UNIVERSIDADE FEDERAL FLUMINENSE

HIGOR DOS SANTOS PINTO

ALINHAMENTO DE CATEGORIAS EM PORTAIS DE DADOS ABERTOS COM BASE EM UM SUBCONJUNTO ABRANGENTE

NITERÓI

2018

UNIVERSIDADE FEDERAL FLUMINENSE

HIGOR DOS SANTOS PINTO

ALINHAMENTO DE CATEGORIAS EM PORTAIS DE DADOS ABERTOS COM BASE EM UM SUBCONJUNTO ABRANGENTE

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: ENGENHARIA DE SISTEMAS E INFORMAÇÃO

Orientadora:
FLAVIA CRISTINA BERNARDINI

Co-orientador:
JOSÉ VITERBO FILHO

NITERÓI

2018

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

P659a Pinto, Higor dos Santos
Alinhamento De Categorias Em Portais De Dados Abertos Com Base Em Um Subconjunto Abrangente / Higor dos Santos Pinto ; Flavia Cristina Bernardini, orientadora ; José Viterbo Filho, coorientador. Niterói, 2018.
202 f. : il.

Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2018.

DOI: http://dx.doi.org/10.22409/PGC.2018.m.10628946775

1. Ciência da computação. 2. Engenharia de sistemas. 3. Transparência na administração pública. 4. Produção intelectual. I. Bernardini, Flavia Cristina, orientadora. II. Viterbo Filho, José, coorientador. III. Universidade Federal Fluminense. Escola de Engenharia. IV. Título.

CDD -

Bibliotecária responsável: Fabiana Menezes Santos da Silva - CRB7/5274

HIGOR DOS SANTOS PINTO

ALINHAMENTO DE CATEGORIAS EM PORTAIS DE DADOS ABERTOS COM BASE EM UM SUBCONJUNTO ABRANGENTE

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: ENGENHARIA DE SISTEMAS E INFORMAÇÃO

Aprovada em NOVEMBRO de 2018.

BANCA EXAMINADORA

Profª. FLÁVIA CRISTINA BERNARDINI - Orientadora, UFF

Prof. JOSÉ VITERBO FILHO - Coorientador, UFF

Profª. LUCIANA CARDOSO DE CASTRO SALGADO, UFF

Prof. DANIEL CARDOSO MORAES DE OLIVEIRA, UFF

Profª. MILENE SELBACH SILVEIRA, PUCRS

Profª. CLAUDIA CAPPELLI, UFRJ

Niterói

2018

Dedico esta dissertação a todos nós, cidadãos, que asseguram a educação e o desenvolvimento da ciência. Que nos esforcemos cada vez mais para levar educação de qualidade a todos e garantir que os frutos da pesquisa científica possam alimentar melhorias significativas no espaço físico e social onde vivemos.

"Vivemos num mundo confuso e confusamente percebido. De um lado, é abusivamente mencionado o extraordinário progresso das ciências e das técnicas, das quais um dos frutos são os novos materiais artificiais que autorizam a precisão e a intencionalidade. De outro lado, há, também, referência obrigatória à aceleração contemporânea e todas as vertigens que cria, a começar pela própria velocidade. Todos esses, porém, são dados de um mundo físico fabricado pelo homem, cuja utilização, aliás, permite que o mundo se torne esse mundo confuso e confusamente percebido."

(Por uma outra globalização, Milton Santos, 2000)

Resumo

A transparência é atualmente um fator chave para manter a confiança das sociedades democráticas em seus governos. A abordagem de dados abertos pode ser utilizada para alavancar a transparência nas gestões públicas. Dessa forma, milhares de conjuntos de dados abertos de governos, organizações e empresas públicas, estão disponíveis em portais na internet. O crescimento no número de portais e na quantidade de informação disponível nos últimos anos, aumentou a dificuldade dos usuários de encontrarem dados úteis para determinadas análises ou pesquisas. Na grande maioria dos portais, esses conjuntos de dados estão distribuídos em tópicos ou categorias. Cada portal utiliza seu próprio conjunto de categorias, mesmo que, em domínios próximos, os conjuntos de dados sejam semelhantes. Essa característica faz com que os usuários precisem de um certo tempo para entenderem a organização da informação em um portal. Neste trabalho apresentamos dois processos desenvolvidos para o alinhamento das categorias de múltiplos portais de dados abertos. No primeiro processo, é obtido um subconjunto reduzido de categorias abrangentes, a partir do grupo de portais em estudo. Na segundo processo, as categorias específicas de cada um desses portais são alinhadas, considerando o subconjunto de categorias abrangentes. Realizamos também uma pesquisa exploratória em 100 portais de cidades americanas densamente populosas, que serviu como estudo de caso para a aplicação dos processos. Ao propor um alinhamento entre as categorias de múltiplos portais, temos como objetivo facilitar a integração de dados nesses portais, oferecendo um único conjunto de categorias que descreve os dados catalogados.

Palavras-chave: dados abertos; portais de dados abertos; categorias; categorização; alinhamento de categorias;

Abstract

Transparency is currently a key factor for maintaining the reliability of democratic societies in their governments. The open data approach can be used to increase the transparency of public administrations. Then, thousands of open datasets from governments, organizations and companies, are available through portals on internet. The growth in the number of portals and in the amount of information available in recent years has increased the difficulty of users in finding useful information for determined analyzes or surveys. In a majority of portals, the datasets are spread across selected topics or categories. But each portal chooses its own set of topics or categories, even if the data sets are very similar on several occasions. This feature takes more time from users to understand the information organization in a portal. In this work we present two processes developed for the categories alignment of multiple open data portals. In the first process, a reduced subset of embracing categories is obtained from the group of portals under study. In the second process, the specific categories of each of these portals are aligned, considering the subset of embracing categories. We also conducted an exploratory research on 100 portals of densely populated American cities, which served as a case study for the application of the processes. In proposing an alignment between the categories of multiple portals, we aim to facilitate the integration of data in these portals, offering a single set of categories that describes the cataloged data.

Keywords: open data; open data portals; categories; categorization; category alignment;

Lista de Figuras

3.1 Categorias de três portais de cidades americanas densamente populosas. . . 28

3.2 Subconjunto Abrangente de Categorias genérico para os portais das três cidades americanas. 28

3.3 Alinhamento de Categorias genérico produzido para as categorias do portal da cidade de Chicago. 29

3.4 Modelo de atividades para Obtenção do Subconjunto Abrangente. 30

3.5 Modelo de atividades para o Alinhamento de Categorias. 38

3.6 Fluxograma genérico que representa o cálculo de similaridade entre duas categorias 40

3.7 Fluxograma genérico que representa o cálculo de similaridade entre duas palavras. 41

4.1 Distribuição da quantidade de categorias por número de portais. 50

4.2 Nuvem de palavras coletadas nas categorias dos diversos portais de dados abertos das cidades americanas. 52

4.3 Distribuição de ocorrência e abrangência das 50 palavras mais frequentes. . 53

4.4 Valores absolutos de similaridade entre as categorias. 59

4.5 Valor de concordância entre os métodos utilizados para o cálculo de similaridade semântica. 61

4.6 Valor de concordância entre os participantes da avaliação no alinhamento das categorias. 64

4.7 Valor de concordância entre os participantes da avaliação no alinhamento das categorias, apresentadas por grupos. 65

A.1 Modelo de atividades para Obtenção do Subconjunto Abrangente. 77

A.2 Categorias dos portais das três cidades americanas mais populosas. 82

A.3 Modelo de atividades para o Alinhamento de Categorias. 96

A.4 Fluxograma genérico que representa o cálculo de similaridade entre duas categorias. 99

A.5 Fluxograma genérico que representa o cálculo de similaridade entre duas palavras. 100

Lista de Tabelas

3.1 Palavras encontradas nas categorias dos portais das três cidades americanas mais populosas e suas frequências. 31

3.2 Abrangência nos portais para cada conjunto C_n de palavras das categorias dos portais das três cidades americanas mais populosas. 33

3.3 Conjunto de Palavras Abrangentes para as palavras das categorias das três cidades americanas mais populosas. 35

3.4 Categorias onde ocorrem as Palavras Mais Abrangentes dos portais das três cidades americanas mais populosas. 36

3.5 Subconjunto Abrangente de Categorias obtidas dos portais das três cidades americanas mais populosas. 36

3.6 Subconjunto Abrangente de Categorias obtidas dos portais das três cidades americanas mais populosas após avaliação de usuário. 37

3.7 Valores de similaridades calculados as entre palavras das categorias *Public Works & Engineering* e *Land Use*. 41

3.8 Modelo de tabela para alinhamento de categorias realizado por pessoas. . . 44

4.1 Amostragem dos 100 portais de cidades americanas densamente populosas. 48

4.2 Abrangência nos portais da quantidade de palavras por grupos, em tamanhos múltiplos de 50. 52

4.3 24 Palavras Mais Abrangentes e suas frequências associadas. 54

4.4 Lista das categorias e frequências associadas a palavra mais frequente *safety*. 55

4.5 Lista das categorias e frequências associadas a palavra mais frequente *land*. 55

4.6 Lista das categorias e frequências associadas a palavra mais frequente *community*. 56

4.7 Categorias mais frequentemente associadas as Palavras Mais Abrangentes obtidas na Pesquisa Exploratória. 57

4.8 Categorias Abrangentes obtidas no processo para os portais das 100 cidades americanas mais populosas. 58

4.9 Resultados de alinhamento no cálculo da similaridade entre palavras e o Valor de Concordância entre os métodos. 60

4.10 Grupos de categorias para o alinhamento dos participantes da avaliação. . 63

4.11 Alinhamento de categorias realizado pelos participantes da avaliação para as categorias do grupo I. 66

4.12 Alinhamento de categorias realizado pelos método de alinhamento para as categorias do grupo I. 67

A.1 Contagem de frequência das palavras do exemplo utilizado nesse guia. . . . 82

A.2 Categorias Abrangentes obtidas na etapa Obter Categorias Mais Abrangentes. 93

A.3 Categorias Abrangentes obtidas após avaliação de usuário. 94

Lista de Abreviaturas e Siglas

API	:	Application Programming Interface;
IC	:	Information Content;
IDE	:	Integrated Development Environment;
JSON	:	JavaScript Object Notation;
NLTK	:	Natural Language Toolkit;
PLN	:	Processamento de Linguagem Natural;
POS	:	Part of Speech;
RDF	:	Resource Description Framework;
SCM	:	Semantic Category Matching;
SGBD	:	Sistemas Gerenciadores de Bancos de Dados;
URL	:	Uniform Resource Locator;
VSM	:	Vector Space Model;

Sumário

1 Introdução **1**

1.1 Contextualização 2

1.1.1 A Sociedade dos Dados 2

1.1.2 Novos Caminhos na *Web* 3

1.1.3 Informação Útil a partir dos Dados 4

1.2 Definição do Problema 5

1.3 Abordagem da Solução 5

1.3.1 Objetivo 6

1.3.2 Definição do Usuário 6

1.3.3 Metodologia 7

1.3.3.1 Pesquisa Bibliográfica 7

1.3.3.2 Pesquisa Exploratória 8

1.3.3.3 Estudo de Caso 8

1.4 Organização 8

2 Fundamentação Teórica e Trabalhos Relacionados **10**

2.1 Dados Abertos 10

2.2 Portais de Dados Abertos 12

2.2.1 Portais de Dados Urbanos 12

2.2.2 Domínio de Conjuntos de Dados Abertos 13

2.2.3 Categorização em Portais 13

2.2.4 Integração de Dados de Diferentes Portais 14

2.3 Processamento de Linguagem Natural 14
2.3.1 Similaridade Semântica 15
2.3.2 Similaridade entre Sentenças 15
2.3.3 Similaridade Palavra-Palavra 16
2.3.3.1 Métodos de Contagem de Arestas 17
2.3.3.2 Métodos Baseados em Conteúdo 18
2.3.4 Tokenização e Remoção das *Stopwords* 20
2.3.5 WordNet e NLTK 21
2.4 Trabalhos Relacionados 22
2.4.1 Catálogos de Portais de Dados Abertos e Pesquisas Exploratórias . 23
2.4.2 Categorização e Alinhamento 24
2.5 Considerações Finais 26

3 Alinhamento de Categorias Baseado em um Subconjunto Abrangente 27

3.1 Obtenção do Subconjunto Abrangente de Categorias 28
3.1.1 Coletar dados dos portais 30
3.1.2 Ler Dados dos Portais 30
3.1.3 Contar Frequência de Palavras 31
3.1.4 Analisar Abrangência das Palavras 32
3.1.5 Definir Parâmetro de Abrangência 34
3.1.6 Obter Conjunto de Palavras Mais Abrangentes 34
3.1.7 Contar Frequência de Categorias 35
3.1.8 Obter Categorias Abrangentes 35
3.1.9 Escrever Categorias Abrangentes 37
3.1.10 Avaliar Categorias Abrangentes 37
3.2 Alinhamento de Categorias dos Portais com o Subconjunto Abrangente . . 38
3.2.1 Coletar Dados dos Portais 39

3.2.2 Ler Dados dos Portais 39
3.2.3 Ler Categorias Abrangentes 39
3.2.4 Calcular Similaridade Semântica entre as Categorias 39
3.2.5 Escrever Resultado do Alinhamento 43
3.2.6 Avaliação do Alinhamento 43
3.3 Avaliação do Processo de Alinhamento de Categorias 44
3.4 Considerações Finais 45

4 Pesquisa Exploratória e Estudo de Caso 46

4.1 Pesquisa Exploratória 46
4.1.1 Análise 46
4.1.1.1 Amostragem de Portais 47
4.1.1.2 Coleta de Informação 48
4.1.2 Resultados 49
4.2 Estudo de Caso 50
4.2.1 Obtenção do Subconjunto Abrangente de Categorias 50
4.2.1.1 Coletar Dados dos Portais 50
4.2.1.2 Ler Dados dos Portais 51
4.2.1.3 Contar Frequência de Palavras 51
4.2.1.4 Analisar a Abrangência das Palavras 52
4.2.1.5 Definir Parâmetro de Abrangência 53
4.2.1.6 Obter Conjunto de Palavras Mais Abrangentes 54
4.2.1.7 Contar Frequência de Categorias 54
4.2.1.8 Obter Categorias Abrangentes 56
4.2.1.9 Escrever Categorias Abrangentes 56
4.2.1.10 Avaliar Categorias Abrangentes 56
4.2.2 Alinhamento de Categorias 58

4.2.2.1 Coletar Dados dos Portais 58
4.2.2.2 Ler Dados dos Portais 58
4.2.2.3 Ler Categorias Abrangentes 59
4.2.2.4 Calcular Similaridade Semântica entre as Categorias . . . 59
4.2.2.5 Escrever Resultado do Alinhamento 61
4.2.2.6 Avaliação do Alinhamento 62
4.2.3 Avaliação do Processo de Alinhamento 62
4.2.3.1 Pesquisa para Avaliação 62
4.2.3.2 Resultados 64
4.3 Considerações Finais 67

5 Conclusão 68
5.1 Contribuições 68
5.2 Limitações 69
5.3 Trabalhos Futuros 69

Referências 71

Apêndice A – GUIA PARA IMPLEMENTAÇÃO DOS PROCESSOS 76
A.1 Processo 1: Obtenção do Subconjunto Abrangente de Categorias 77
A.1.1 Atividade 1: Coletar dados dos portais 78
A.1.2 Atividade 2: Ler Dados dos Portais 80
A.1.3 Atividade 3: Contar a Frequência de Palavras 81
A.1.4 Atividade 4: Analisar a Abrangência das Palavras 85
A.1.5 Atividade 5: Definir Parâmetro de Abrangência 88
A.1.6 Atividade 6: Obter Conjunto de Palavras Mais Abrangentes 88
A.1.7 Atividade 7: Contar Frequência de Categorias 89
A.1.8 Atividade 8: Obter Categorias Abrangentes 91

A.1.9 Atividade 9: Escrever Categorias Abrangentes 92

A.1.10 Atividade 10: Avaliar Categorias Abrangentes 93

A.2 Processo 2: Alinhamento das Categorias dos Portais 96

A.2.1 Atividade 1: Coletar dados dos portais 97

A.2.2 Atividade 2: Ler Dados dos Portais 97

A.2.3 Atividade 3: Ler Categorias Abrangentes 97

A.2.4 Atividade 4: Calcular Similaridade Semântica entre as categorias. . 98

A.2.5 Atividade 5: Escrever resultado do alinhamento 111

A.2.6 Atividade 6: Avaliação do Alinhamento 111

Apêndice B – ABRANGÊNCIA DAS PALAVRAS NOS PORTAIS UTILIZADOS NO EXEMPLO DO GUIA 113

Apêndice C – DADOS DOS PORTAIS COLETADOS NA PESQUISA EXPLORATÓRIA 117

Apêndice D – CHAMADA DAS FUNÇÕES PARA OBTENÇÃO DO SUBCONJUNTO ABRANGENTE 152

Apêndice E – CHAMADA DAS FUNÇÕES PARA O ALINHAMENTO DE CATEGORIAS 154

Apêndice F – RESULTADO DO ALINHAMENTO DE CATEGORIAS PARA O ESTUDO DE CASO 155

Apêndice G – PLANILHA DE AVALIAÇÃO DO PROCESSO DE ALINHAMENTO DE CATEGORIAS 183

Capítulo 1

Introdução

A crescente demanda da sociedade moderna por transparência e o maior engajamento do cidadão nas questões e problemas dos governos democráticos têm produzido diversas soluções baseadas em sistemas computacionais. Na última década, várias cidades ao redor do mundo vêm disponibilizando dados em formatos abertos através de portais disponíveis na internet. Particularmente nos USA, houve um grande esforço e fomento do governo federal para o desenvolvimento de portais de dados abertos de cidades a partir de 2009. As cidades criaram novos cargos administrativos para desenvolver políticas e organizar seus dados [1].

Dados de uma cidade podem ser produzidos por sistemas computacionais, empregados nas diversas organizações, ou podem ter sido coletados por sensores ou equipamentos eletrônicos espalhados pelo território urbano. Diversos setores, como a indústria, academia, imprensa e o próprio poder público, têm utilizado esses dados para produzir análises que aumentam a compreensão de fenômenos urbanos nas grandes cidades ao redor do mundo [2].

Os portais de dados abertos são páginas de internet onde os governos podem disponibilizar e catalogar seus dados. Por meio desses portais, a sociedade pode interagir com os dados disponíveis e produzir informações úteis [3]. Podem ainda desenvolver soluções baseadas em aplicativos para as mais diversas áreas. Muitas vezes, os dados nesses portais estão catalogados por categorias ou grupos. Essas categorias distribuem os conjuntos de dados de um portal em temas diversos. Ao buscar uma mesma informação em portais diferentes, o usuário pode encontrar dificuldade para a navegação entre as diferentes categorias disponíveis nos diversos portais. Neste trabalho, propomos dois processos para o alinhamento de categorias entre portais de dados abertos baseado em um subconjunto abrangente de categorias. Abrangente neste contexto, significa que o conjunto de catego-

rias descreve grande parte dos dados de todos esses portais.

1.1 Contextualização

O acesso à informação de organizações públicas, e também privadas, tem sido discutido como um direito importante nas democracias modernas. A demanda por informação como um requisito para transparência tem alavancado uma sociedade aberta na qual o objetivo é o estabelecimento de cidadãos engajados capazes de entender e acessar as informações disponíveis [4].

A busca por transparência nas gestões públicas das sociedades democráticas vêm proporcionando o surgimento de um grande número de portais de dados abertos. No entanto, apesar do esforço de diversos desenvolvedores, cientistas e pesquisadores, esses portais demandam de uma maior padronização, principalmente no tocante ao vocabulário utilizado pelos editores para descrever os conjuntos de dados [1]. A grande oferta de dados disponíveis nos vários portais, junto a falta de padronização no vocabulário das estruturas de categorias desses portais levam a uma maior dificuldade dos usuários na integração dos conjuntos de dados necessários para produzirem informação útil para uma determinada análise ou pesquisa.

1.1.1 A Sociedade dos Dados

O advento dos computadores nos forneceu a capacidade de ler, escrever e manipular dados e informações de uma forma muito rápida, eficaz e precisa. Os primeiros computadores processavam e armazenavam dados em cartões perfurados que eram ordenados e armazenados em caixas, em locais adequados, algo bem parecido a uma seção comum de arquivos de documentos de uma empresa ou organização. Conforme o uso crescente dos computadores em todas as áreas, a evolução tecnológica produziu inúmeras soluções físicas e lógicas para o armazenamento de informação pelos computadores. Os dispositivos de armazenamento passaram a gravar dados em arquivos, que são lidos de forma muito mais rápida e eficiente. Essa evolução trouxe a tona sistemas de armazenamento denominados sistemas gerenciadores de bancos de dados ou *SGBDs*. Esses sistemas são responsáveis por persistir informações e também gerenciar o acesso de leitura e escrita ao dado [5].

O uso cada vez mais intenso dos sistemas de computadores na solução de problemas em diversas áreas de atuação humana, elevou exponencialmente a quantidade de bases de informações armazenadas nos sistemas de banco de dados. No entanto, assim como os

antigos sistemas de cartão perfurados, essas bases de dados estão isolados umas das outras e, na grande maioria das vezes, mantém dados redundantes. Isso deu início a uma nova discussão na atividade de armazenar e processar informações e dados. Novos paradigmas são propostos e desenvolvidos para permitir uma maior integração das informações disponíveis entre os milhares de sistemas utilizados atualmente no cotidiano da sociedade contemporânea [6].

Novas ferramentas foram emergindo nesse cenário, de grandes quantidades de dados espalhados nos servidores, para dar conta dos problemas que se apresentavam. O *Big Data* é um termo muito utilizado atualmente, e se refere a utilização e integração de grandes bases de dados distribuídas em servidores, muitas vezes distantes [7]. A massificação do uso da internet e a grande disponibilidade de informação trouxe também para a *Web* a necessidade de trabalhar com uma grande quantidade de dados e informações distribuídas, armazenadas em diferentes nós na rede [8].

1.1.2 Novos Caminhos na *Web*

A *World Wide Web* foi projetada inicialmente como um ambiente interativo de informações compartilhadas onde as pessoas podiam se comunicar umas com as outras e com as máquinas. É um espaço abstrato onde podem interagir entre si e com o próprio espaço. Ela cresceu, inicialmente, como um meio para transmissão de material de servidores corporativos, altamente carregados de arquivos contendo diversos tipos de informações, para a massa de consumidores da internet. A conveniência e o potencial impacto social e econômico despertaram interesse de comunidades diversas, muito além do uso que os computadores tinha anteriormente [9].

O volume de informação que era armazenado em bancos de dados de servidores foi crescendo exponencialmente. Surgiram então problemas relacionados à busca e recuperação de informações. Os mecanismos de busca tradicionais provaram ser úteis, já que grandes índices podem ser pesquisados muito rapidamente, retornando diversos documentos. Ao mesmo tempo em que provaram ser ineficientes, na medida em que suas buscas geralmente levam em conta apenas o vocabulário dos documentos, e têm pouco ou nenhum conceito, produzindo assim muitos resultados inúteis [8]. A grande quantidade de informação torna mais difícil a busca, descoberta e a integração das informações obtidas nas pesquisas.

A capacidade humana de encontrar informações na *Web* é limitada aos resultados de indexação textual oferecidos pelos mecanismos de busca. Dessa forma, muito tem se

discutido sobre a melhoria da compreensão da informação contida na *Web* pelas máquinas. Além de ser um espaço navegável por seres humanos, é esperado que a *Web* contenha dados de forma compreensível por máquinas, permitindo que elas tomem parte mais importante na análise da *Web* e resolvendo problemas para nós [9].

A Web Semântica estuda diversos conceitos e técnicas para a estruturação do conteúdo significativo das páginas da *Web*. O objetivo principal é criar um ambiente no qual os agentes de software, em trânsito de uma página para outra, possam realizar prontamente tarefas sofisticadas para os usuários, como buscas de conceitos relacionados [8]. A estruturação do conteúdo contido em formato legível para as máquinas pode proporcionar diversos avanços nas ferramentas de busca e descoberta de informação na *Web* [8].

Existem diversas iniciativas para publicar dados na *Web* de forma que possam ser encontradas por buscas mais refinadas e complexas de sistemas computacionais. Uma iniciativa, muito presente, principalmente no campo governamental, são as políticas de dados abertos. As políticas de dados abertos consistem na publicação e disseminação de dados e informações públicas, seguindo alguns critérios, que possibilitam sua reutilização e o desenvolvimento de aplicativos por toda a sociedade. A maior parte dos dados e informações geradas ou mantidas pelo governo, são públicas. Os benefícios da oferta de dados podem trazer melhorias na eficiência da gestão e na qualidade das políticas públicas. Fomentar a inovação mediante a utilização de dados abertos no desenvolvimento de aplicações e serviços inovadores é, dessa forma, promover o crescimento econômico[10].

1.1.3 Informação Útil a partir dos Dados

A grande disponibilidade de dados em servidores particulares, e na internet, possibilitou o advento de novas atividades profissionais. Uma atividade muito comum atualmente é a análise de dados, que inclui, a seleção de bases de dados, avaliação de sua qualidade, integração e a produção de informação útil para um determinado problema ou fim. As análises de dados se utilizam de várias técnicas e conhecimentos diversos. Habilidades em programação, métodos estatísticos e confecção de gráficos são, muitas das vezes, necessárias para a produção de informação a partir de bases de dados.

Muitas empresas têm se beneficiado das vantagens que as análises e a interpretação de seus dados tem trazido para os negócios. Demandas estratégicas podem ser atendidas facilmente utilizando o potencial dos dados, que, muitas das vezes, já estão armazenados nos sistemas de bancos de dados das empresas. Os governos também podem ser mais eficientes utilizando análises de dados para tomada de decisão e fomento de políticas

públicas. Além disso, ao oferecer seus dados,, através das políticas de dados abertos, os governos propõem para a sociedade o papel de produzir informação útil e acessível com o objetivo de oferecer ao cidadão maior entendimento dos problemas e também maior participação na solução e minimização dos problemas de suas cidades [2].

1.2 Definição do Problema

Um usuário pode querer realizar uma pesquisa sobre a gestão das escolas de sua cidade, sobre o trânsito de seu bairro, ou ainda, sobre os gastos com o máquina pública de seu estado. A definição dos conjuntos de dados ideais para a solução de um determinado problema pode demandar muitas horas de pesquisa nos diferentes portais. Pode-se gastar muito mais tempo ainda integrando os diferentes conjuntos de dados, ou seja, produzindo relações entre os dados [11].

Os administradores publicam os conjuntos de dados agrupando-os em categorias ou tópicos. Os conjuntos de dados disponíveis nesses portais, abrangem diversos tópicos ou categorias diferentes [12]. Os portais utilizam categorias diversas na publicação dos conjuntos de dados. Podemos encontrar ainda várias categorias nos diferentes portais que possuem significado semântico equivalente. Assim, um usuário que queira produzir um estudo entre as diversas cidades de seu estado, ou país, por exemplo, encontra dificuldades para navegar entre as categorias dos diferentes portais.

As categorias de um portal representam o domínio de assuntos coberto pelos conjuntos de dados do portal. Como domínio, definimos, as diferentes áreas onde podem ser produzidos dados abertos. Por exemplo, para portais de cidades, temas como transporte, saúde, educação e segurança pública são extremamente comuns. É importante ressaltar também que a literatura apresenta uma divergência de nomenclatura ao se referir ao domínio de assuntos dos conjuntos de dados nos portais. Oliveira et al. [13] se refere aos assuntos como *domínios* e *tópicos*. Já Barbosa et al. [12] se refere a *tópicos* e *categorias*. Neste trabalho, utilizamos o termo *categoria* para denominar os assuntos referentes aos conjuntos de dados de um portal.

1.3 Abordagem da Solução

O trabalho de Iyengar e Lepper[14] apresenta resultados experimentais sobre a motivação de clientes na escolha de produtos. Ao serem apresentados para diversas marcas de um

mesmo produto, o cliente tem menos motivação para comprar o produto. Dessa forma, oferecer ao usuário poucas opções de escolha é um paradigma que tem sido adotado cada vez mais nos trabalhos gráficos e de organização da informação. Assim, ao integrar dados ou catálogos de diferentes portais, pretendemos oferecer um único conjunto de categorias para organização e pesquisa de dados nos diversos portais.

Neste trabalho apresentamos dois processos para alinhamento entre as categorias de diversos portais de dados abertos. Ao propor uma categorização única entre os diversos portais, pretendemos facilitar a integração de dados nesses diferentes portais, através das categorias. Também apresentamos, neste trabalho, uma pesquisa exploratória realizada em vários portais de dados abertos de cidades americanas densamente populosas.

1.3.1 Objetivo

Dadas as oportunidades e os desafios apontados na categorização e na integração de conjuntos de dados abertos nos portais, descritos anteriormente, propomos dois processos para alinhamento de categorias entre diversos portais, para que, ao se integrar os conjuntos de dados desses diferentes portais, possa ser oferecido ao usuário um único conjunto abrangente de categorias para pesquisa.

1.3.2 Definição do Usuário

Existem diversos portais de dados abertos que integram catálogos de dados de diferentes portais. O portal Data.gov [15], mantido pelo governo federal americano, é um exemplo de portal que integra catálogos de outros portais. No Brasil, temos o exemplo do portal Dados.gov [16], mantido pelo governo federal brasileiro e que integra diversos catálogos de portais de dados abertos de instituições federais.

Os processos descritos neste trabalho podem ser utilizados por editores de portais desse tipo para integração dos catálogos dos vários portais de onde se coletam os dados. Dessa forma, o perfil de usuário para utilização dos processos apresentados neste trabalho inclui: conhecimentos em portais de dados abertos, integração de dados e linguagem de programação *Python*, a qual utilizamos para codificação das atividades algorítmicas dos processos desenvolvidos. Na próxima seção vamos discutir a metodologia utilizada no desenvolvimento desse trabalho.

1.3.3 Metodologia

Para garantir que uma pesquisa cientifica seja reconhecidamente sólida e relevante, tanto pelo campo acadêmico, quanto pela sociedade em geral, ela deve demonstrar que é passível de debate e verificação. O uso de metodologias que dão suporte à pesquisa científica é cada vez mais necessário e exigido. Nesse sentido, ao descrever sua pesquisa através de uma metodologia, o pesquisador fornece ao leitor a capacidade de validação dos passos e soluções empregadas no processo de pesquisa [17].

A pesquisa científica pode ser classificada de acordo com diferentes critérios. Podemos caracterizar tipos de pesquisa de acordo com sua natureza, objetivos ou procedimentos técnicos. Um trabalho de pesquisa pode não limitar-se a um único tipo. Além disso, alguns tipos de pesquisa podem servir de base para outros [18].

Neste trabalho, utilizamos alguns tipos de pesquisa científica para propor e comunicar a solução para o problema abordado. São eles: a pesquisa bibliográfica, a pesquisa exploratória e o estudo de caso.

1.3.3.1 Pesquisa Bibliográfica

A pesquisa bibliográfica sugere o estudo de artigos, teses, livros e outras publicações usualmente indexadas e disponibilizadas por editoras. A pesquisa bibliográfica é um passo fundamental e prévio para qualquer trabalho científico, mas não produz qualquer conhecimento novo. Ela caracteriza uma pesquisa quanto aos seus procedimentos técnicos [18].

Neste trabalho foi realizado uma extensa pesquisa bibliográfica para, primeiro, definir o tema e o problema abordado. Segundo, para produzir uma fundamentação teórica consistente que servisse de base para a proposição da solução adequada ao problema. E terceiro, para encontrar trabalhos relacionados a este.

Neste capítulo realizamos uma contextualização histórica e do panorama atual do estado da arte com relação aos portais de dados abertos e a integração de dados. No Capítulo 2 apresentamos a fundamentação teórica e os trabalhos relacionados obtidos com a pesquisa bibliográfica.

1.3.3.2 Pesquisa Exploratória

Uma pesquisa exploratória é um tipo de pesquisa onde o autor não necessariamente tem uma hipótese ou objetivo definido. A pesquisa exploratória é uma classificação da pesquisa científica quanto ao seu objetivo. Ela pode ser considerada, muitas vezes, como o primeiro estágio de um processo de pesquisa mais longo. Na pesquisa exploratória o autor vai examinar um fenômeno buscando características que sejam pouco conhecidas ou demonstradas. Assim, formando uma base de conhecimento para uma pesquisa mais elaborada [18].

Uma pesquisa exploratória foi definida e realizada neste trabalho para se verificar, na prática, o problema abordado. Através dela produzimos dados que demonstram a grande variedade de categorias utilizadas para descrever os conjuntos de dados em diversos portais de dados abertos. Essa pesquisa exploratória é apresentada na Seção 4.1 do Capítulo 4.

1.3.3.3 Estudo de Caso

Um estudo de caso envolve o estudo profundo de um ou poucos objetos de maneira que se permita o seu amplo e detalhado conhecimento [19]. Os resultados obtidos com o estudo de caso não devem ser generalizadores. Ou seja, não podem ser usados para representar todos os indivíduos, mas sim apenas aqueles que foram diretamente investigados [18].

Neste trabalho o estudo de caso foi utilizado para aplicar os processos desenvolvidos, como solução do problema, nos dados obtidos na pesquisa exploratória. O estudo de caso realizado neste trabalho é apresentado na Seção 4.2 do Capítulo 4.

1.4 Organização

Neste capítulo apesentamos o cenário atual de uso e desenvolvimento de soluções e ferramentas ligadas a abordagem de dados abertos. Contextualizamos o problema de se encontrar conjuntos de dados adequados para a construção de pesquisas e análises nos portais de dados abertos, já que as ferramentas de buscas nos portais oferecem poucas oportunidades de integração. Ainda neste capítulo, descrevemos o objetivo principal deste trabalho, o desenvolvimento de processos para alinhamento de categorias para facilitar a integração de dados nos portais de dados abertos pelo uso de categorias.

No Capítulo 2 apresentamos os principais conceitos necessários para um melhor entendimento desse trabalho. São abordados temas de dados abertos e de portais de dados

abertos. Também descrevemos a estrutura de categorização encontrada nesses portais. Ainda no Capítulo 2, são apresentados conceitos relacionados a tarefas frequentemente realizadas na área de Processamento de Linguagem Natural (*PLN*). Neste trabalho, foram utilizadas várias técnicas desenvolvidas na área de *PLN*.

No Capítulo 3 apresentamos os dois processos desenvolvidos neste trabalho. O primeiro realiza a Obtenção do Subconjunto Abrangente de Categorias, que produz um conjunto de categorias que descreve os dados de todos os portais na entrada do processo. O segundo processo realiza o Alinhamento de Categorias dos portais com as categorias do Subconjunto Abrangente. Descrevemos todas as etapas desenvolvidas para a realização do processo.

No Capítulo 4 apresentamos a Pesquisa Exploratória que mostra a categorização de 100 portais de dados abertos de cidades americanas densamento populosas. A pesquisa exploratória ainda foi utilizada para produzir dados para o Estudo de Caso onde aplicamos os processos desenvolvidos e apresentados no Capítulo 3. Descrevemos também, ainda no Capítulo 4, uma avaliação realizada para comparação dos resultados obtidos no alinhamento de categorias, realizado pelo processo definido neste trabalho, com o alinhamento produzido por pessoas.

No Capítulo 5 discutimos as principais contribuições e limitações desse trabalho, e ainda apresentamos os trabalhos que podem ser realizados utilizando os resultados e contribuições descritos neste trabalho. No Apêndice A é apresentado um Guia de Implementação dos Processos, onde descrevemos as atividades dos processos definidos neste trabalho e disponibilizamos os códigos necessários para a realização das atividades algorítmicas.

Capítulo 2

Fundamentação Teórica e Trabalhos Relacionados

Neste capítulo abordamos alguns conhecimentos básicos, de diferentes áreas, que utilizamos no desenvolvimento dos processos propostos neste trabalho. Definimos e conceituamos dados abertos e discutimos a organização e distribuição em categorias dos conjuntos de dados nos portais. Também apresentamos algumas técnicas de Processamento de Linguagem Natural *(PLN)* utilizadas nos processos para alinhamento de categorias, descrito no Capítulo 3, e no tratamento dos dados obtidos na Pesquisa Exploratória, descrita no Capítulo 4.

2.1 Dados Abertos

A Open Definition[20], definida pela Open Knowledge International [21], estabelece os principais princípios que definem os conceitos de dados abertos, em relação ao dado em si e em relação ao conteúdo. De maneira geral, define dados abertos como: "dados e conteúdos que podem ser usados, modificados e compartilhados por qualquer pessoa para qualquer propósito"[20]. Segundo a Open Knowledge International [21], as principais características atribuídas aos dados abertos são [22]:

- Disponibilidade: os dados devem estar disponíveis a um custo de reprodução razoável, de preferência através de download pela internet. Os dados também devem estar disponíveis de forma conveniente e modificável.
- Reutilização e redistribuição: os dados devem ser fornecidos em termos que permitam a reutilização e redistribuição, incluindo o intercâmbio com outros conjuntos de dados. Os dados devem ser legíveis por máquina.

- Participação universal: todos devem poder usar, reutilizar e redistribuir, não deve haver discriminação contra os campos de trabalho ou contra pessoas ou grupos. [22]

Uma outra grande contribuição da Open Definition[20], são os 8 princípios dos dados públicos abertos, os quais abrangem conceitos relativos aos campos de completibilidade, primariedade, temporalidade, acessibilidade, possibilidade de processamento por máquina, não ser discriminatório, não proprietário e de livre licença. [23]

As políticas e princípios de dados abertos são atividades inovadoras e inspiradoras, e podem ser observadas em diversas áreas. Permitem ainda, o surgimento de novos negócios. Equipam formas tradicionais e novas de jornalismo com conjuntos de dados abertos e reutilizáveis. O governo aberto está impulsionando mudanças e crescimento na economia digital no Reino Unido [24].

Governos das diferentes instâncias da administração pública possuem diversos dados armazenados em servidores e sistemas. Através da utilização das políticas de dados abertos, os governos podem oferecer esses dados para sociedade e contribuir para a produção de novas ferramentas e, consequentemente, novos negócios. Assim, podem impulsionar o processo de controle social dos gastos e das políticas públicas ao disponibilizar informações de suas gestões, contribuindo para melhorias no sistemas democráticos e engajamento dos cidadãos nas decisões políticas.

A transparência nos órgãos públicos é fundamental para manter a confiança da sociedade em seus governos [25]. Além de alavancar a transparência das gestões públicas, as políticas de dados abertos podem contribuir para o surgimento de diversas oportunidades econômicas e de negócios. Também podem dar suporte a sistemas democráticos mais eficientes e a cidadãos mais engajados politicamente [2].

Os dados abertos de cidades dão suporte para o desenvolvimento de novos aplicativos e ferramentas. Atualmente, podemos encontrar diversos aplicativos que consomem dados abertos. O *CityMapper*[26] é um aplicativo que utiliza dados em tempo real de cidades para calcular melhores rotas de transportes urbanos. Integrando diversas modalidades de transporte e combinando com dados de horário, localização e estimativas de chegada das linhas, a ferramenta oferece o tempo mais curto de deslocamento entre dois pontos. O *SafeEats*[27] exibe dados de localização e de resultados de inspeções sanitárias nos restaurantes da cidade de Nova York no *smartphone* do usuário. Ainda no contexto do uso de dados abertos, podemos dizer que as visualizações de dados, em formatos de fácil compreensão pelo usuário, podem ajudar a sociedade civil a acompanhar situações críticas

em ambientes urbanos, como infestações de doenças graves [28].

2.2 Portais de Dados Abertos

Com o objetivo de tornar seus dados públicos, vários governos de cidades ao redor do mundo vêm construindo portais de dados abertos. Esses portais disponibilizam os conjuntos de dados para visualização e download pelos usuários. Quando falamos de dados públicos abertos, quatro conceitos são fundamentais: descoberta de dados, consumo de dados, interoperabilidade e engajamento da comunidade. Dessa forma, ferramentas, plataformas e governos, precisam estar em sintonia com esses conceitos para oferecerem maior aderência as políticas de dados abertos. [3].

2.2.1 Portais de Dados Urbanos

Podemos definir portais de dados urbanos como páginas especializadas onde conjuntos de dados descritos por metadados de alta qualidade podem ser publicados, visualizados e baixados [3]. Com o aumento do número de portais, organizações e empresas têm desenvolvido soluções em tecnologias de dados abertos para governos e gestões privadas. Dessa forma, essas empresas e comunidades oferecem plataformas para o desenvolvimento de portais de dados abertos. O CKAN [29] e o Socrata [30] são duas das plataformas mais conhecidas atualmente.

Toda a informação disponível sobre um conjunto de dados está contida em seus metadados. Informações como autor, data de publicação, formato, licença, palavras-chave, entre outras, são comumente encontradas nos modelos de metadados de diversas plataformas de dados abertos. As categorias relacionadas aos conjuntos de dados também podem ser encontradas, muitas das vezes, no conjunto de metadados. Metadados são informações que criamos, armazenamos e compartilhamos para descrever objetos, nos permitem interagir com esses objetos para obter o conhecimento que precisamos. A definição clássica e literal, com base na etimologia da própria palavra, é "dados sobre dados". Os metadados são essenciais em diversas áreas que utilizam sistemas digitais e aplicações. Aplicativos de mídia digital, páginas de busca, transações bancárias, tudo isso utiliza metadados para armazenar informações e históricos [31].

2.2.2 Domínio de Conjuntos de Dados Abertos

A estrutura de categorias disponível em um portal representa o domínio de assuntos cobertos pelos conjuntos de dados desse portal. As áreas de atuação do governo de uma cidade definem o domínio dos conjuntos de dados de seu portal. Como transporte, saúde, educação e segurança pública.

A Open Knowledge International[21] cita aspectos da abrangência do domínio na utilização de dados abertos com vários usos potenciais em aplicações. Os domínios citados incluem: Cultura, Ciência, Finanças, Estatísticas, Clima e Ambiente [22]. A abrangência de domínio dos conjuntos de dados em um portal de dados abertos pode ser grande e heterogênea. Como demonstrado nos trabalhos de Barbosa et al. [12] e Oliveira et al. [13], ao analisar diferentes portais de dados abertos pode-se encontrar diferentes tópicos ou categorias para classificar os conjuntos de dados. Também pode-se encontrar diversas palavras para se referir a um único sentido ou conceito de categoria. Neste trabalho vamos explorar o domínio dos conjuntos de dados das diferentes cidades americanas densamente populosas através das palavras utilizadas para definir as categorias nos diferentes portais de dados abertos.

2.2.3 Categorização em Portais

Um portal da *Web* é um site que fornece conteúdo de informações sobre um tópico comum ou um interesse específico. Permite que indivíduos interessados em algum tópico possam receber notícias, encontrar e conversar com pessoas, criar uma comunidade e encontrar links para recursos da rede de interesse comum. Normalmente, os portais da *Web* podem definir uma ontologia para o comunidade. Esta ontologia define terminologias para descrever conteúdo e serve como índice para recuperação de conteúdo [32].

No contexto de sistemas computacionais, uma ontologia é um documento ou arquivo que define formalmente as relações entre os termos que descrevem um conteúdo. O tipo mais comum de ontologia para a *Web* possui uma taxonomia e um conjunto de regras de inferência associados. A taxonomia define classes de objetos e relações entre eles. Classes, subclasses e relações entre entidades são uma ferramenta muito poderosa para uso na *Web*. Pode expressar um grande número de relações entre entidades, atribuindo propriedades a classes e permitindo que subclasses herdem essas propriedades [8]. Outras maneiras de categorizar a informação podem estar baseadas em modelos de domínio ou simples hierarquias de categorias.

A recuperação da informação por outro lado, tem explorado modelos de representação de textos para levar a descoberta de padrões. No caso de portais de dados abertos, em geral, a categorização se dá por especialistas. Assim, a representação da informação pode auxiliar a descoberta de padrões. No entanto, adquirir conhecimento com especialistas para evoluir na representação do conhecimento é uma tarefa árdua.

2.2.4 Integração de Dados de Diferentes Portais

Uma atividade normalmente executada nas análises de dados é a integração de conjuntos de dados de diferentes portais. Por exemplo, podemos querer integrar conjuntos de dados que informam a disponibilidade, qualidade e os gastos das escolas públicas em diferentes cidades. Dessa forma, precisamos visitar os portais das várias cidades, baixar os conjuntos de dados e produzir a integração desses conjuntos. No entanto, essa atividade pode demandar processos de formatação e conversão de tipos de dados e arquivos. Assim, muitas ferramentas tem sido desenvolvidas para facilitar o processo de integração de dados de diferentes portais.

As plataformas disponíveis para construção de portais oferecem algumas ferramentas que podem auxiliar a integração de dados. Destacamos aqui uma extensão desenvolvida para o uso no CKAN[29], o CKAN Harvester [33]. Essa ferramenta facilita muito a importação de conjuntos de dados de uma instância remota do CKAN, por exemplo, um portal municipal, para uma outra instância do CKAN, por exemplo, um portal federal, onde poderão ser catalogados todos os dados das instâncias municipais. Essa extensão é altamente personalizável, permitindo definir tags, grupos, usuários e permissões padrão para os conjuntos de dados importados [33]. O método de alinhamento de categorias proposto neste trabalho pode ser utilizado na integração de dados de diferentes portais, já que produz um conjunto de categorias, ou grupos, abrangente e alinha as categorias das instâncias remotas com o conjunto abrangente.

2.3 Processamento de Linguagem Natural

O objetivo do Processamento de Linguagem Natural *(PLN)* é fornecer às máquinas a capacidade de processar e compor textos. Dentre as tarefas que um computador pode fazer ao processar um texto estão incluídas: reconhecer o contexto, fazer análise sintática, semântica, léxica e morfológica, criar resumos, extrair informação, interpretar os sentidos, analisar sentimentos e até aprender conceitos com os textos processados [34]. As diversas

aplicações possíveis de *PLN* incluem, por exemplo, a tradução automática, a categorização de textos, o processamento de discurso e a web semântica.

Neste trabalho é proposto um alinhamento semântico das categorias de diversos portais de dados abertos. Para isso, utilizamos algumas técnicas comumente empregadas nas soluções de problemas de *PLN*. A determinação da similaridade semântica entre dois segmentos de textos é uma tarefa muito utilizada em diversas aplicações, como na categorização de textos e análise de discurso.

2.3.1 Similaridade Semântica

O problema de calcular a similaridade semântica entre dois conceitos, palavras ou frases é um antigo problema na área de *PLN*. Em geral, para se calcular a semelhança semântica entre duas palavras é medida a distância conceitual entre dois objetos em uma determinada ontologia de conceitos semânticos [35]. Esses objetos representam as palavras ou conceitos que se quer comparar. A determinação da similaridade semântica é utilizada em uma ampla gama de aplicações, como em motores de busca na internet, determinação de similaridade genética em bases de códigos genéticos, recuperação de informação, classificação e categorização de textos na web.

Na literatura, podemos encontrar várias técnicas para a medida da similaridade semântica entre dois conceitos. Um conjunto de métodos bastante utilizados são os que utilizam uma base de dados léxica, ou ontologia, que contém conceitos e ligações entre eles. Nas técnicas que utilizam bases de dados léxicas, a similaridade é calculada utilizando a estrutura hierárquica das palavras e significados distribuídas em forma de árvore. Esses métodos, basicamente, utilizam a distância entre os conceitos na árvore [35]. Uma das bases de dados léxicas mais extensas e utilizadas na língua inglesa é o *WordNet* [36]. O *WordNet* possui conceitos relacionados de nomes, verbos, adjetivos e advérbios, que são agrupados em conjuntos de sinônimos.

2.3.2 Similaridade entre Sentenças

Além dos métodos que calculam similaridade semântica entre duas palavras, podemos encontrar diversas técnicas na literatura para calcular a similaridade semântica entre textos mais longos, porém poucos trabalhos abordam o cálculo de similaridade entre pequenas sentenças ou frases [37]. Diferentemente de se calcular a similaridade entre textos longos, que pode ser obtida através da análise de frequência das palavras que ocorrem nos

dois textos, na similaridade de pequenas sentenças as palavras quase, ou nunca, se repetem [37]. Dessa forma, outras abordagens são necessárias para a solução do problema de se comparar dois segmentos de texto curtos. Na solução proposta neste trabalho, a medida de similaridade entre curtos segmentos de texto é fundamental, dado a necessidade do alinhamento semântico de categorias, que são formadas por curtos segmentos de texto.

Uma abordagem frequentemente explorada é calcular a similaridade entre as palavras de cada segmento de texto, e então propor uma relação entre a similaridade semântica dos termos e as palavras de cada termo. Dados dois curtos segmentos de texto, as categorias, queremos obter medidas que indicam sua similaridade em nível semântico. Dessa forma, propomos uma adaptação do método apresentado por Mihalcea et al. [38]. Assim, modelamos a similaridade semântica dos segmentos de texto como uma função da similaridade semântica de suas palavras. E então, determinamos as similaridades das palavras entre as duas sentenças. De modo que, dadas duas sentenças T_1 e T_2, para cada palavra na sentença T_1 calcula-se a maior similaridade dessa palavra com todas as outras palavras da sentença T_2. O mesmo processo é feito para T_2.

$$sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w \in T_1} maxSim(w, T_2)}{|T_1|} + \frac{\sum_{w \in T_2} maxSim(w, T_1)}{|T_2|}\right) \tag{2.1}$$

Na Eq. 2.1 é apresentada a equação que calcula a similaridade sim entre duas sentenças T_1 e T_2, onde $maxSim(w, T_i)$ é a similaridade máxima de uma palavra da sentença T_j com todas as palavras da sentença T_i.

É importante observar que o método originalmente proposto por Mihalcea et al. [38] utiliza a frequência inversa de documento para balancear os termos da soma. No entanto, nós balanceamos a equação apenas com o valor do tamanho do conjunto de palavras-chave de cada sentença. Para o domínio aplicado e dado o curto tamanho das sentenças avaliadas, não aplicamos informação de conteúdo no cálculo de similaridade semântica entre as sentenças. Na seção a seguir, apresentamos como pode ser realizado esse cálculo de similaridade entre palavras.

2.3.3 Similaridade Palavra-Palavra

Para avaliar a similaridade semântica de pequenos segmentos de textos, precisamos calcular a similaridade entre as palavras de cada segmento de texto, de acordo com a Eq. 2.1. Neste trabalho, obtemos as palavras das categorias do portais utilizando as técnicas descritas na Seção 2.3.4. Nesta seção, vamos apresentar alguns métodos disponíveis para

o cálculo de similaridade semântica entre palavras.

Podemos dividir os métodos para determinar a similaridade semântica entre palavras basicamente em duas categorias principais:

- Métodos de Contagem de Arestas: Essas técnicas medem a similaridade entre dois conceitos C_1, C_2 determinando o caminho que conecta os termos na taxonomia.
- Métodos baseados em Conteúdo: Nesta categoria, as medidas de similaridade são tomadas com base no conteúdo da informação de cada conceito.

Existem outra maneiras de se classificar os métodos de similaridade palavra-palavra, no entanto, para melhor entendimento das técnicas utilizadas neste trabalho, vamos adotar a classificação apresentada.

2.3.3.1 Métodos de Contagem de Arestas

1. **Menor caminho (*path similarity - path*)** [39]

   A distância conceitual, sim_{ps}, entre dois nós é geralmente proporcional ao número de arestas que separam os dois nós na hierarquia. É uma medida do quão perto dois conceitos estão em uma determinada ontologia [40], é apresentada na Eq.2.2 .

$$sim_{ps}(c_1, c_2) = 2MAX - L \tag{2.2}$$

   onde MAX é o comprimento máximo do caminho entre dois conceitos na taxonomia, ou seja, o maior caminho possível na hierarquia. L é o número mínimo de arestas entre os conceitos c_1 e c_2. Essa equação é uma adaptação proposta por Varelas[41] do método do menor caminho apresentado no trabalho de Rada[39]. Existem muitas questões em aberto sobre o realismo cognitivo da medida de caminho mínimo, no entanto, é uma medida simples, de prática fácil e aceitável em redes semânticas hierárquicas [41].

2. **Wu and Palmer (*wup similarity - wup*)** [42]

   Essa medida de similaridade, $sim_{wp}(c_1, c_2)$, considera a posição dos conceitos c_1 e c_2 na taxonomia em relação à posição do conceito de maior especificidade comum C, apresentada na Eq. 2.3. Onde N_1 e N_2 é o número de links hiperônimos de

c_1 e c_2, respectivamente, para o conceito mais comum C, e H é o número de links hiperônimos de C para a raiz da taxonomia.

$$sim_{wp}(c_1, c_2) = \frac{2H}{N_1 + N_2 + 2H} \tag{2.3}$$

O conceito de maior especifidade comum C é o pai comum entre os conceitos c_1 e c_2 com o número mínimo de links hiperônimos. Como pode haver vários pais para cada conceito, dois conceitos podem compartilhar pais por vários caminhos [41].

3. **Leacock and Chodorow (*lch similarity - lch*)** [43]

   A medida de parentesco proposta por Leacock e Chodorow[43] é dada pela Eq. 2.4, onde $length$ é o comprimento do caminho mais curto entre os dois conceitos c_1 e c_2, usando a contagem de nós. D é a profundidade máxima da taxonomia.

$$sim_{lch}(c_1, c_2) = -\log \frac{length}{2 * D} \tag{2.4}$$

2.3.3.2 Métodos Baseados em Conteúdo

A noção de conteúdo informacional do conceito está diretamente relacionada à frequência do termo em uma determinada coleção de documentos. As frequências de termos na taxonomia são estimadas usando frequências nominais em algumas grandes coleção de textos. A ideia por trás das técnicas que utilizam conteúdo de informação é que a semelhança de dois conceitos está relacionada a informações que eles compartilham em comum, como indicado por um conceito específico que inclui os dois termos [41].

Associando probabilidades a conceitos na taxonomia, para cada conceito c, $p(c)$ é a probabilidade de encontrar o conceito c na taxonomia. A probabilidade de um conceito é definida como $p(c) = freq(c)/N$, onde N é o número total de termos na taxonomia, $freq(c) = \sum_{n \in words(c)} n$ e $words(c)$ é o conjunto de termos incluídos por c [44].

Dadas estas probabilidades, várias medidas de similaridade semântica foram definidas. Todas estas medidas usam o conteúdo informativo do compartilhamento dos pais dos dois termos c_1 e c_2, onde $S(c_1, c_2)$ é o conjunto de conceitos que incluem $c1$ e $c2$. Dessa forma, segundo a Eq. 2.5, p_{mis} é definida como a menor probabilidade dos pais compartilhados

de c_1 e c_2 [41]:

$$p_{mis}(c_1, c_2) = min_{c \in S(c1,c2)} p(c) \tag{2.5}$$

4. **Resnik (*res similiraty - res*)** [44]

   Essa medida se baseia em que, quanto mais informações dois termos compartilham em comum, mais semelhantes eles são, e as informações compartilhadas por dois termos são indicadas pelo conteúdo informativo do termo que os inclui na taxonomia [41].

$$sim_{res}(c_1, c_2) = -\ln p_{mis} \tag{2.6}$$

5. **Lin (*lin similarity - lin*)** [45]

   Essa medida, sim_{Lin}, utiliza tanto o conteúdo informativo dos pais compartilhados entre os dois termos quanto o conteúdo informativo dos termos comparados, conforme a Eq. 2.7.

$$sim_{Lin}(c_1, c_2) = \frac{2 \ln p_{mis}(c_1, c_2)}{\ln p(c_1) + \ln p(c_2)} \tag{2.7}$$

   A medida de Resnik[44] depende unicamente do conteúdo informativo dos pais compartilhados, e há tantas valores possíveis para similaridade quanto termos de ontologia. Usando o conteúdo informativo dos termos comparados e do pai compartilhado, o número de valores possíveis é quadrático no número de termos que aparecem na ontologia, aumentando assim a probabilidade de ter valores diferentes para pares diferentes de termos. Consequentemente, essa medida oferece uma medida mais sensível de similaridade do que a medida Resnik [41].

6. **Jiang et al. (*jcn similarity - jcn*)** [46]

Ao contrário das medidas de similaridade apresentadas acima, essa é uma medida de *distância semântica*, $dist_{jcn}(c_1, c_2)$, dada pela Eq. 2.8. Dessa forma, a similaridade entre dois conceitos c_1 e c_2, é dada pela Eq. 2.9

$$dist_{jcn}(c_1, c_2) = -2 \ln p_{mis}(c_1, c_2) - (\ln p(c_1) + \ln p(c_2)) \tag{2.8}$$

$$sim_{jcn}(c1, c2) = 1 - dist_{jcn}(c1, c2) \tag{2.9}$$

Essa medida pode resultar em valores arbitrariamente grandes, como a medida de Resnik[44], embora na prática tenha um valor máximo de $2\ln(N)$, onde N é o tamanho do corpus. Além disso, combina conteúdo informativo do pai compartilhado e dos conceitos comparados, como a medida *Lin*[45] faz. Assim, esta medida parece combinar as propriedades das diversas medidas de similaridade apresentadas [41].

Todas as medidas de similaridade apresentadas utilizam uma ontologia para identificar e relacionar os conceitos semânticos entre as palavras. Na Seção 2.3.5 vamos descrever a ontologia e as implementações dos métodos de cálculo de similaridade palavra-palavra utilizadas neste trabalho. Na próxima seção vamos descrever o processo de obtenção de palavras-chave a partir de um segmento de texto.

2.3.4 Tokenização e Remoção das *Stopwords*

Para possibilitar que a máquina processe informação textual, são necessários processamentos que abstraem e estruturam a língua, deixando apenas o que é informação relevante. Esse pré-processamento reduz o vocabulário e torna os dados menos esparsos, característica conveniente para o processamento computacional. O processo de tokenização tem como objetivo separar palavras ou sentenças em unidades.

Uma outra tarefa muito utilizada no pré-processamento de textos é a remoção de *stopwords*. Esse método consiste em remover palavras muito frequentes, tais como “a”, “de”, “o”, “da”, “que”, “e”, “do” entre outras, pois na maioria das vezes não são informações relevantes para o domínio aplicado. Diversas listas de *stopwords* podem ser encontradas na internet para aplicações distintas.

O método de similaridade entre sentenças descrito na Seção 2.3.2 calcula o valor de similaridade para as palavras de cada segmento de texto. Dessa forma, devemos gerar as palavras para cada sentença, categoria, a qual queremos calcular a similaridade.

Obtemos as palavras de cada categorias aplicando os processos de tokenização e remoção das *stopwords*.

2.3.5 WordNet e NLTK

O *WordNet*[36] é um grande banco de dados léxico em língua inglesa. Substantivos, verbos, adjetivos e advérbios são agrupados em conjuntos de sinônimos cognitivos, denominados *synsets*, cada um expressando um conceito distinto. Os *synsets* são interligados por meio de relações conceituais, semânticas e lexicais. A estrutura do *WordNet* faz dele uma ferramenta útil para linguística computacional e processamento de linguagem natural.

O *WordNet* se assemelha superficialmente a um dicionário de sinônimos, pois agrupa as palavras com base em seus significados. No entanto, existem algumas distinções importantes. Primeiro, o *WordNet* interliga não apenas formas de palavras, mas sentidos específicos de palavras. Como resultado, as palavras que estão próximas umas das outras na rede não possuem ambiguidade semântica. Segundo, o *WordNet* rotula as relações semânticas entre palavras, enquanto o agrupamento de palavras em um dicionário de sinônimos não segue nenhum padrão explícito [36].

A principal relação entre as palavras no *WordNet* é a sinonímia. Sinônimos são palavras que denotam o mesmo conceito e são intercambiáveis em muitos contextos. No *WordNet* são agrupados em conjuntos não ordenados, os *synsets*. Cada um dos 117.000 *synsets* do *WordNet* é vinculado a outros *synsets* por meio de um pequeno número de relações conceituais. Além disso, um *synset* contém uma definição breve e, na maioria dos casos, uma ou mais sentenças curtas ilustrando o uso dos membros do *synset*. As palavras com vários significados distintos são representadas em tantos *synsets* quantos forem os significados distintos. Assim, cada significado no *WordNet* é único [36].

A relação mais frequentemente codificada entre os *synsets* é a relação super-subordinada (também chamada de hiperonímia, hiponímia ou relação *is-a*). Ela vincula *synsets* mais gerais aos mais específicos. Assim, o *WordNet* afirma que a categoria *mobiliário* inclui *cama*, que por sua vez inclui *beliche*. Por outro lado, conceitos como *cama* e *beliche* compõem a categoria *mobiliário* [36].

Outras relações também são codificadas no *WordNet*, são elas: meronímia, que é a relação entre as partes e o todo, por exemplo, cadeira possui encosto e pernas. E a antonímia, que expressa a relação de oposição entre as palavras. A maioria das relações

conecta palavras da mesma parte do discurso *(POS - Part of Speech)*. Assim, o *WordNet* realmente consiste em quatro sub-redes, substantivos, verbos, adjetivos e advérbios, com poucos ponteiros entre pontos [36].

Ao comparar duas frases, temos muitos pares de palavras que possuem múltiplos *synsets*. Portanto, não considerar o sinônimo adequado ao contexto da sentença, pode introduzir erros na fase inicial do cálculo de similaridade. Assim, o sentido da palavra afeta significativamente a medida geral de semelhança. Identificar o sentido da palavra faz parte da área de pesquisa desambiguação do sentido da palavra [35]. Desambiguação do sentido da palavra é o processo de atribuir um significado a um determinado palavra com base no contexto em que ocorre. Muitas vezes o conjunto de possíveis significados para uma palavra é conhecido antes do tempo, e é determinado pelo sentido inventário de um dicionário legível por máquina ou banco de dados léxico.

No domínio dessa pesquisa, não possuímos o contexto das palavras pré-definidos. Assim, utilizamos um algoritmo para desambiguação de palavras conhecido como similaridade máxima. Esse algoritmo propõe utilizar a maior similaridade entre os *synsets* de um par de palavras [47]. A Equação 2.10 define a similaridade máxima $argmax_{synset(a)}$ entre todos os *synsets* de um par de palavras. Dessa forma são calculados a similaridade entre todos os *synsets* das duas palavras e a maior similaridade é o resultado final.

$$argmax_{synset(a)} \sum_{i}^{n} max_{synset(i)}(sim(i, a)) \tag{2.10}$$

O *NLTK*, *Natural Language Toolkit* [48] é uma plataforma de desenvolvimento para programas em Python[49] que implementa diversos métodos para as atividades de *PLN*. O framework fornece interfaces fáceis para mais de 50 recursos léxicos, como o *WordNet*, juntamente com um conjunto de bibliotecas de processamento de texto para classificação, tokenização, *stemming*, *tagging*, análise e raciocínio semântico . Neste trabalho utilizamos a implementação do *WordNet* fornecido no *NLTK*[48]. Todos os processos de tokenização e remoção das *stop words* também foram implementados utilizando os métodos disponíveis no *NLTK*[48].

2.4 Trabalhos Relacionados

Nesta seção vamos apresentar alguns trabalhos que apresentam o estado da arte em relação a catálogos de portais de dados abertos e a interoperabilidade entre esse catálogos.

Mostramos que pesquisas exploratórias têm sido realizadas em portais de dados abertos. Também apresentamos trabalhos relacionados a conceitos de alinhamento semântico, muito utilizados na área de processamento de linguagem natural e alinhamento de ontologias.

2.4.1 Catálogos de Portais de Dados Abertos e Pesquisas Exploratórias

Encontramos na literatura alguns trabalhos que produziram pesquisas exploratórias em portais de dados abertos. Esses trabalhos mostram aspectos relevantes e lacunas ainda não resolvidas na área de dados abertos.

Barbosa et al. [12] coletaram mais de 9.000 conjuntos de dados de 20 portais de cidades da América do Norte e analisaram diferentes aspectos desses dados, incluindo seu conteúdo e tamanho, quão recentes e dinâmicos eles são, formatos usados, qualidade dos dados e oportunidades para integração.

Outro trabalho bastante motivador é apresentado por Oliveira et al. [13], onde são analisados 13 portais de dados abertos brasileiros. Nesse trabalho foram coletados dados manualmente e também utilizando extração de metadados por interfaces de aplicação das plataformas de publicação. Também apresentaram características dos conjuntos de dados como tamanho, formato, domínio, atualização, entre outros.

Uma atividade complexa e que tem demandado bastante atenção dos cientistas da área é a integração de diferentes catálogos de portais. Catálogos de dados governamentais são encontrados em portais de dados abertos fornecendo descrições dos conjuntos de dados disponíveis nesses portais. Esses catálogos, muitas vezes, permanecem isolados. No entanto, diversas iniciativas para federação desses catálogos, seja geograficamente sobrepostos ou tematicamente complementares, tem sido realizadas. Maali et al. [50] propuseram um vocabulário *RDF* como um formato de intercâmbio entre catálogos de dados e como uma maneira de trazê-los para a web semântica, onde eles podem desfrutar de interoperabilidade entre si e com outros conjuntos de dados implantados. O design do vocabulário foi definido por meio de uma pesquisa exploratória com sete catálogos de dados de cinco países diferentes, e foi aplicado na unificação de quatro catálogos de dados para permitir consultas cruzadas e navegação entre esses catálogos.

2.4.2 Categorização e Alinhamento

Uma das lacunas atuais na avaliação de portais de dados abertos se refere a qualidade das estruturas de categorização nesses portais. Yang et al. [51] produziram avaliações sobre a qualidade das estruturas de categorização de portais de dados abertos investigando, automaticamente, a coerência dos conjuntos de dados na mesma categoria. A estrutura de categorização de diversos portais de dados abertos de Taiwan foram avaliadas comparando-se a estrutura do portal a um estrutura padrão, definida pelos autores. A qualidade da categorização pôde ser medida também pela investigação da similaridade de dados dentro da mesma categoria. A similaridade entre essas estruturas de categorias é calculada usando o modelo vetorial[52], utilizado em sistemas de recuperação da informação. O modelo vetorial, conhecido mais comumente como modelo de espaço vetorial, *VSM* (*Vector Space Model*), representa documentos e consultas como vetores de termos. Termos são ocorrências únicas nos documentos. Os documentos devolvidos como resultado para uma consulta são representados por um vetor montado através de um cálculo de similaridade. Aos termos das consultas e documentos são atribuídos pesos que especificam o tamanho e a direção do vetor que o representa. Ao ângulo formado por estes vetores dá-se o nome de q. O termo $cos(q)$ determina a proximidade da ocorrência. O cálculo da similaridade é baseado neste ângulo entre os vetores que representam o documento e a consulta [52].

Hoshiai et. al. [53] encontraram pares de categorias semanticamente correspondentes entre dois sistemas de categorização diferentes aplicando tecnologias, desenvolvidas por eles, de correspondência semântica de categorias, *SCM (Semantic Category Matching)* nos problemas de alinhamento de ontologias. A correspondência semântica de categorias baseia-se em uma abordagem estatística que pega amostras de documentos de cada categoria, dados de descrição da estrutura hierárquica e produz saídas para todos os pares de categorias que correspondem semanticamente aos dois sistemas de classificação. Essa abordagem foi utilizada em vários problemas conhecidos de alinhamento de ontologias. Para a promover a correspondência semântica entre as categorias, eles também utilizaram modelos de espaço vetorial.

O alinhamento de ontologias é um processo complexo que ajuda a reduzir a lacuna semântica entre diferentes representações sobrepostas de um mesmo domínio. A existência de tais representações diferentes obedece ao instinto humano natural de ter diferentes perspectivas e, portanto, de modelar problemas de maneiras diferentes [54]. Diversas técnicas e abordagens diferentes para o alinhamento de ontologia tem sido produzidas.

Algumas técnicas são relevantes para esse trabalho. Dentre elas podemos destacar as técnicas baseadas em strings [54]. Estas técnicas são baseadas na similaridade das strings que representam os nomes e descrições das entidades nas ontologias. Existem várias métricas de distância de strings que podem ser utilizadas nestes métodos Levenshtein, Jaccard, Jaro-Winkler, Euclidiana, TFIDF, etc. [55]. Tais técnicas estão presentes, por exemplo no trabalho de Akbari et al. [56]. Outros tipos de técnicas comumente utilizadas no alinhamento de ontologias são as técnicas baseadas em linguagem natural [54]. Essas técnicas dependem do processamento de linguagem natural, pois elas não consideram nomes simplesmente como strings, mas palavras em alguma linguagem natural. Técnicas nesta categoria usam, por exemplo, tokenização, lemmatização ou remoção de *stopwords*. Algumas dessas técnicas são aplicadas por Shah e Syeda-Mahmood [57]. Esses tipos de técnicas de alinhamento de ontologias também aproveitam as vantagens de recursos externos para encontrar semelhanças entre os termos, usando, por exemplo, dicionários léxicos. Encontramos esse tipo de abordagem no trabalho de He et. al. [58], que utilizaram o banco de dados *WordNet*[36] como recurso externo.

A unidade de descrição para alinhamento de ontologias é a classe (ou instância), e a unidade de descrição para um *SCM* é a categoria (domínio do objeto) dos tópicos do documento. Assim a granularidade do alinhamento de ontologias é menor que a do *SCM*. Além disso, no alinhamento de ontologias, propriedades para os atributos de ambos os objetos e relações entre esses objetos podem ser descritos. Por outro lado, não podemos descrever nenhum relacionamento lógico predefinido entre qualquer uma das partes de um documento no SCM. Então, as informações descritas no alinhamento de ontologias são mais detalhadas que as informações descritas no *SCM* [53].

Neste trabalho descrevemos processos para o alinhamento semântico de categorias. Dessa forma, aplicamos uma abordagem de *SCM* para portais de dados abertos, utilizando técnicas de *PLN*. Diferentemente do trabalho de Hoshiai et. al. [53] não avaliamos os documentos relacionados as categorias, nesse caso os conjuntos de dados, e sim produzimos um alinhamento entre as próprias categorias, utilizando similaridade semântica entre pequenos segmentos de textos ou sentenças, atividade comum na área de *PLN*. Assim, os documentos podem ser entendidos como os dados ou catálogos dos portais, e as categorias são atributos desses documentos.

2.5 Considerações Finais

Neste capítulo, apresentamos os principais conceitos relacionados a fundamentação teórica deste trabalho. Discutimos a abordagem de dados abertos e o cenário atual dos portais. Também apresentamos as técnicas de *PLN* que utilizamos no tratamento dos dados da pesquisa exploratória, descrita no Capítulo 4, e no desenvolvimento dos processos de alinhamento de categorias, descritos no Capítulo 3. A similaridade semântica entre segmentos de textos é a principal ferramenta, descrita neste capítulo, utilizada na construção do alinhamento de categorias. Ferramentas de tokenização e remoção de *stop words* também foram apresentadas neste capítulo. Ainda, discutimos trabalhos relacionados que mostram o estado da arte em relação aos portais de dados abertos e à abordagem da solução empregada neste trabalho.

Capítulo 3

Alinhamento de Categorias Baseado em um Subconjunto Abrangente

Neste capítulo apresentamos os processos desenvolvidos para o Alinhamento de Categorias nos portais de dados abertos, baseado em um Subconjunto Abrangente de Categorias. Apresentamos dois processos para executar o Alinhamento de Categorias. No primeiro processo é extraído o Subconjunto Abrangente de Categorias, que contém categorias que ocorrem frequentemente nos portais estudados. No segundo processo, são alinhadas as categorias de vários portais com esse Subconjunto Abrangente.

Apresentamos abaixo um exemplo genérico para o alinhamento de categorias. Todos os processos, de obtenção das categorias abrangentes e de alinhamento, desse exemplo, foram realizados manualmente. Na Figura 3.1 são apresentadas as categorias de portais de três cidades americanas densamente populosas, Nova York, Los Angeles e Chicago. Essas categorias foram obtidas nos sites dos portais. Podemos produzir um Subconjunto Abrangente de Categorias genérico que descreve as categorias mais frequentes nesses portais. Um Subconjunto Abrangente de Categorias genérico é apresentado na Figura 3.2. Esse conjunto genérico de categorias abrangentes foi obtido manualmente de acordo com as categorias dos portais, mostradas na Figura 3.1. Podemos produzir também um Alinhamento de Categorias genérico para os portais. Na Figura 3.3 é apresentado um alinhamento genérico para as categorias da cidade de Chicago. As categorias do portal foram alinhadas manualmente com as categorias do Subconjunto Abrangente da Figura 3.2.

O objetivo das atividades descritas nesse capítulo é a realização automática desses processos, com algumas intervenções pontuais do usuário. As atividades dos processos estão divididas entre atividades do usuário e atividades de algoritmo. Todas as atividades estão descritas completamente no Guia de Implementação dos Processos, disponível no

Figura 3.1: Categorias de três portais de cidades americanas densamente populosas.

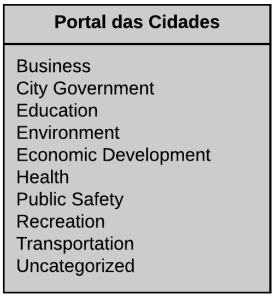

Figura 3.2: Subconjunto Abrangente de Categorias genérico para os portais das três cidades americanas.

Apêndice A junto com os códigos, em linguagem Python[49], necessários para as atividades de algoritmo.

Na próxima seção, apresentamos o primeiro processo para o alinhamento de categorias. Nessa primeira parte é obtido o Subconjunto Abrangente de Categorias que descreve o conteúdo dos diversos portais em estudo, por meio das Categorias Abrangentes.

3.1 Obtenção do Subconjunto Abrangente de Categorias

Nesta seção, vamos apresentar as etapas desenvolvidas para a Obtenção do Subconjunto Abrangente de Categorias, as quais vamos denominar de Categorias Abrangentes. Essas categorias são obtidas através da análise de frequência das categorias entre os diversos

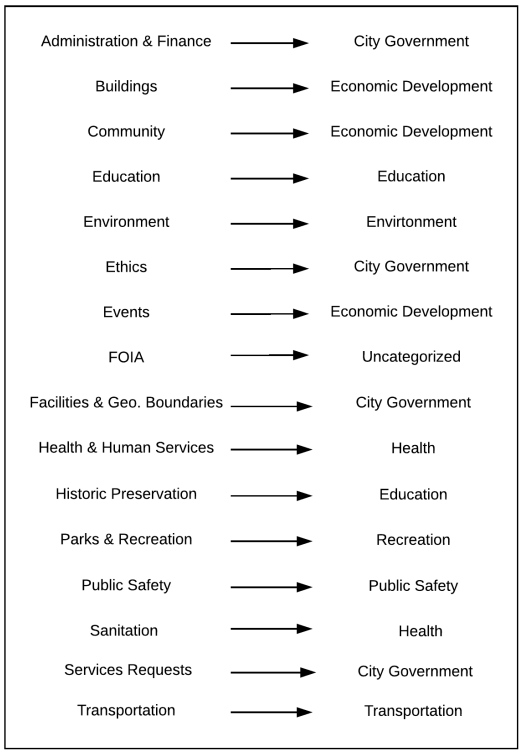

Figura 3.3: Alinhamento de Categorias genérico produzido para as categorias do portal da cidade de Chicago.

portais de dados abertos estudados. Dado um número de portais de dados abertos, o processo extrai um conjunto comum de categorias mais abrangentes entre os diversos portais. Por meio da contagem de frequência das palavras encontradas nas categorias, podemos propor um conjunto de categorias mais abrangentes.

Na Figura 3.4 é apresentado o modelo do processo e suas atividades. As atividades estão descritas em sequência uma das outras. Para melhor entendimento do processo, não são apresentadas no modelo as entradas e saídas das atividades.

No Apêndice A, descrevemos o Guia de Implementação dos Processos, onde são apresentadas detalhadamente todas as atividades, suas entradas e saídas. Existem atividades realizadas pelo usuário, onde não necessárias funções computacionais, e atividades realizadas por algoritmos, onde apresentamos as funções computacionais utilizadas para realização da atividade.

Nas atividades de algoritmo o usuário precisa obter os códigos necessários para realização dessas atividades, e executá-los, de acordo com a ordem das atividades do modelo. Para melhor visualização dos códigos, aconselhamos o usuário a utilizar alguma *IDE* para

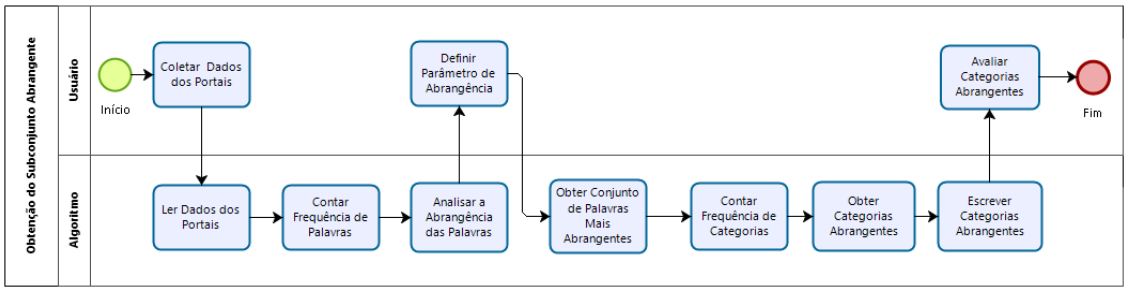

Figura 3.4: Modelo de atividades para Obtenção do Subconjunto Abrangente.

Python.

3.1.1 Coletar dados dos portais

O processo se inicia com a coleta de dados dos portais pelo usuário. A coleta de dados dos portais pode ser realizada manualmente, automaticamente, ou ainda, pode ter sido obtida de uma base de dados disponível em sites ou ferramentas online, como o Data Portals [59].

Coletar as categorias manualmente significa visitar cada portal que se queira e então anotar todas as categorias de cada portal. Muitos dos portais de dados abertos disponibilizam seu catálogo para acesso através de *API*. Dessa forma, obtendo os catálogos por aplicação, a coleta de dados dos portais pode ser automatizada. As *URL's* dos portais devem ser conhecidas e também as chamadas das funções das *API's*. Ainda, como uma terceira possibilidade, pode-se encontrar bases de dados de informações de portais. O Data Portals [59] oferece um catálogo de vários portais de dados abertos ao redor do mundo.

Definimos como entrada das atividades de algoritmo um arquivo *JSON* contendo os dados dos portais necessários para a execução das atividades. No Apêndice A apresentamos um exemplo do arquivo *JSON* utilizado na entrada da atividade Ler Dados dos Portais.

3.1.2 Ler Dados dos Portais

A leitura dos dados dos portais é realizada através de um arquivo *JSON*. Os dados dos portais incluem o nome do portal, as categorias e outras características. O nome do portal e suas categorias são dados necessários para o funcionamento do processo, os demais dados

são opcionais.

Definimos uma classe de objetos *Portal* para entrada das atividades de algoritmo. Os dados dos portais são lidos no arquivo *JSON* e então são criados objetos do tipo *Portal* para armazenar esses dados em memória. A definição da classe, o código de leitura e o exemplo do arquivo *JSON* estão disponíveis e descritos no Guia de Implementação dos Processos, no Apêndice A.

3.1.3 Contar Frequência de Palavras

As categorias dos portais são a entrada da atividade Contar Frequência de Palavras. As categorias são decompostas em palavras, ou *tags*, conforme descrito na seção 2.3.4, e então são contadas as frequências de ocorrência de todas as palavras nos diversos portais. Como resultado, essa atividade produz um dicionário contendo todas as palavras dos portais e suas frequências de ocorrência, ordenado por frequência.

Utilizamos o código escrito para essa atividade, apresentado no Apêndice A, Seção A.1.3, para contar a frequência das palavras nas categorias dos portais da Figura 3.1. Na Tabela 3.1 são apresentadas as palavras encontradas nas categorias desses portais e suas frequências.

Tabela 3.1: Palavras encontradas nas categorias dos portais das três cidades americanas mais populosas e suas frequências.

education	2	social	1	ethics	1
environment	2	livable	1	events	1
health	2	sustainable	1	facilities	1
safety	2	prosperous	1	boundaries	1
recreation	2	safe	1	human	1
services	2	well	1	historic	1
transportation	2	run	1	preservation	1
business	1	administration	1	parks	1
government	1	finance	1	sanitation	1
housing	1	buildings	1	service	1
development	1	community	1	requests	1

A ordem de leitura dos portais e, consequentemente, das categorias é um fator importante no algoritmo. Quando as palavras possuem a mesma frequência, a ordenação entre as palavras de mesma frequência na lista é realizada de acordo com a leitura dos dados nos portais. Assim a ordenação dos portais no arquivo de entrada é de extrema importância e deve ser levada em conta pelo usuário. No exemplo utilizado, ordenamos os portais de

acordo com a densidade populacional de cada cidade, a mais populosa aparece primeiro. Na Tabela 3.1 as palavras são apresentadas em ordem de frequência (entre as palavras de mesma frequência a ordem é pela entrada, palavra lida primeiro fica na frente) de cima para baixo e da esquerda para direita.

3.1.4 Analisar Abrangência das Palavras

O objetivo dessa atividade é avaliar o conjunto de palavras encontradas nos portais, a partir das mais frequentes para as menos frequentes, em termos da ocorrência em uma quantidade de portais. A análise da Abrangência das Palavras nos portais será utilizada para obter um subconjunto de palavras mais frequentes, que por sua vez, será utilizado como entrada na atividade Contar Frequência de Categorias. Lembramos que o objetivo final desse processo é eleger um conjunto de Categorias Abrangentes, as quais serão alinhadas, no processo Alinhamento de Categorias, com as categorias dos portais de interesse.

Definimos a Abrangência como a quantidade de portais onde um grupo de palavras ocorre. Assim, de acordo com um Valor de Abrangência definido pelo usuário, podemos eleger um conjunto de palavras mais abrangentes. A Tabela 3.1 mostra a lista de palavras, e suas frequências, encontradas nos portais das três cidades americanas mais populosas, utilizadas como exemplo. Seja C_1 o conjunto formado pela primeira palavra da lista, C_2 o conjunto formado pela primeira e segunda palavras da lista, C_3, o conjunto formado pela primeira, segunda e terceira palavras da lista, e assim sucessivamente. C_n é o conjunto formado por n palavras da lista. Para cada um desses conjuntos devemos calcular a quantidade de portais onde ocorrem simultaneamente todas as palavras do conjunto. A quantidade de portais onde ocorrem as palavras do conjunto é o Valor de Abrangência desse conjunto. Definimos os conjuntos C_n abaixo:

Definição 3.1: *Seja L uma lista com todas as palavras das categorias dos portais em ordem de frequência, e entrada, definimos os conjuntos C_n como*:

$$\begin{aligned} C_1 &= L(1) \\ C_2 &= L(1) \cup L(2) \\ C_3 &= L(1) \cup L(2) \cup L(3) \\ &\ldots \\ C_n &= L(1) \cup L(2) \cup L(3) \cup \ldots \cup L(n) \end{aligned} \tag{3.1}$$

Definição 3.2: *Seja T_n o conjunto de portais onde ocorrem todas as palavras do conjunto C_n, definimos o Valor de Abrangência $A(C_n)$ dos conjuntos C_n como*:

$$\begin{aligned} A(C_1) &= |T(1)| \\ A(C_2) &= |T(1) \cap T(2)| \\ A(C_3) &= |T(1) \cap T(2) \cap T(3)| \\ &... \\ A(C_n) &= |T(1) \cap T(2) \cap T(3) \cap ... \cap T(n)| \end{aligned} \tag{3.2}$$

Na Tabela 3.2 são mostrados os valores de abrangência calculados, de acordo com as Definições 3.1 e 3.2 para todos os conjuntos C_n com as palavras das categorias dos portais das cidades americanas mais populosas, apresentadas na Figura 3.1. Os conjuntos C_n são representados pela única palavra diferente em relação ao conjunto anterior, $C_n - 1$. O conjunto C_1 é formado pela palavra *(education)*. O conjunto C_2 é formado pelas palavras *(education, environment)*, e é representado pelo palavra *environment*. O conjunto C_3 é formado pelas palavras *(education, environment, health)*, e é representado pelo palavra *health*. E assim por diante.

Tabela 3.2: Abrangência nos portais para cada conjunto C_n de palavras das categorias dos portais das três cidades americanas mais populosas.

C_1	education	66.7%	C_{12}	social	66.7%	C_{23}	ethics	100%
C_2	environment	66.7%	C_{13}	livable	100%	C_{24}	events	100%
C_3	health	66.7%	C_{14}	sustainable	100%	C_{25}	facilities	100%
C_4	safety	66.7%	C_{15}	prosperous	100%	C_{26}	boundaries	100%
C_5	recreation	66.7%	C_{16}	safe	100%	C_{27}	human	100%
C_6	services	66.7%	C_{17}	well	100%	C_{28}	historic	100%
C_7	transportation	66.7%	C_{18}	run	100%	C_{29}	preservation	100%
C_8	business	66.7%	C_{19}	administration	100%	C_{30}	parks	100%
C_9	government	66.7%	C_{20}	finance	100%	C_{31}	sanitation	100%
C_{10}	housing	66.7%	C_{21}	buildings	100%	C_{32}	service	100%
C_{11}	development	66.7%	C_{22}	community	100%	C_{33}	requests	100%

A análise da abrangência é necessária para escolher o conjunto C_n que será utilizado para definir as palavras mais frequentes nos portais. O tamanho desse conjunto será o tamanho do conjunto de Categorias Abrangentes. Essa análise é fundamental para a escolha do Parâmetro de Abrangência, descrito na Seção 3.1.5.

3.1.5 Definir Parâmetro de Abrangência

O usuário deve definir, logo após a atividade Analisar a Abrangência das Palavras, seção 3.1.4, o valor de corte para a abrangência das palavras nos portais. Isso irá definir o tamanho do conjunto de palavras mais frequentes, e consequentemente, o tamanho do Subconjunto Abrangente de Categorias. A atividade Analisar Abrangência das Palavras produz informações necessárias para a definição do Parâmetro de Abrangência.

A definição do Parâmetro de Abrangência é uma atividade importante e deve ser realizada com cuidado, pois, dependendo da frequência das palavras, um valor alto para o parâmetro de abrangência produzirá um conjunto maior de Categorias Abrangentes, um valor mais baixo produzirá um conjunto menor de Categorias Abrangentes. O usuário deve analisar a abrangência das palavras, utilizando o algoritmo proposto na atividade Analisar Abrangência das Palavras, seção 3.1.4, e definir o valor adequado do parâmetro.

Na Tabela 3.2 são apresentados todos os conjuntos C_n e suas respectivas Abrangências para as categorias dos portais das três cidades americanas mais populosas. Observamos que existem apenas dois valores possíveis para a Abrangência de todos os conjuntos, $66,7\%$ e 100%. Se utilizarmos o Parâmetro de Abrangência igual a $66,7\%$ o primeiro conjunto a satisfazer a condição é o conjunto C_1, que contém apenas uma palavra. Essa não é uma boa escolha, pois o conjunto de Palavras Abrangentes terá apenas uma palavra, e consequentemente, o conjunto de Categorias Abrangentes também terá apenas uma categoria. Para o Parâmetro de Abrangência igual a 100%, o primeiro conjunto a satisfazer a condição é o conjunto C_{13}. Dessa forma existem 13 palavras no conjunto para serem eleitas como Palavras Mais Abrangentes. Para o caso do exemplo das três cidades americanas mais populosas, o Parâmetro de Abrangência no valor de 100% se mostra adequado. Isso nos diz que essas 13 palavras ocorrem conjuntamente em todos os portais estudados no exemplo.

3.1.6 Obter Conjunto de Palavras Mais Abrangentes

O objetivo dessa etapa é obter as Palavras Mais Abrangentes nos diversos portais após Definir o Parâmetro de Abrangência, seção 3.1.5, utilizando a análise obtida na atividade Analisar Abrangência das Palavras, seção 3.1.4. O primeiro conjunto C_n que, entre todos os conjuntos obtidos pela Definição 3.1, tiver o valor de Abrangência, obtido pela Definição 3.2, igual ao do Parâmetro de Abrangência é eleito como Conjunto de Palavras Mais Abrangentes.

Utilizando as categorias dos portais das três cidades americanas mais populosas, descritas na Figura 3.1, na atividade Definir Parâmetro de Abrangência, seção 3.1.5, concluímos que o valor de 100% produz um tamanho para conjunto de Palavras Abrangentes adequado. Da Tabela3.2 obtemos o conjunto C_{13} como primeiro conjunto a satisfazer esse valor. Assim o conjunto C_{13} é eleito como conjunto de Palavras Mais Abrangente.

Tabela 3.3: Conjunto de Palavras Abrangentes para as palavras das categorias das três cidades americanas mais populosas.

education	business
environment	government
health	housing
safety	development
recreation	social
services	livable
transportation	

3.1.7 Contar Frequência de Categorias

Nessa atividade, para cada palavra do conjunto C_n de Palavras Mais Abrangentes são contadas as frequências das categorias dos portais onde ocorrem cada palavra. Assim, podemos associar, na próxima etapa, para cada Palavra Mais Abrangente, a categoria mais frequente, nos diversos portais, onde ocorre a palavra.

Utilizamos o algoritmo proposto no Guia de Implementação dos Processos, disponível na Seção A.1.7, do Apêndice A para realizar a contagem de frequência das categorias dos portais das três cidades americanas mais populosas. Na Tabela 3.4 são apresentadas as categorias e suas frequência de ocorrência para o conjunto C_{13} de Palavras Mais Abrangentes, obtido na etapa anterior.

3.1.8 Obter Categorias Abrangentes

Nesse etapa são escolhidas as categorias mais frequentes para serem associadas a cada Palavra Mais Abrangente. Para cada Palavra Mais Abrangente escolhe-se a categoria mais frequente onde ocorre a palavra. Caso haja um empate entre as frequências das categorias as categorias que apresentam a mesma frequência são escolhidas para o conjunto de Categorias Abrangentes.

Na Tabela 3.4 são apresentadas as categorias e suas frequência de ocorrência para o conjunto C_{13} de Palavras Mais Abrangentes. Coincidentemente todas as categorias

Tabela 3.4: Categorias onde ocorrem as Palavras Mais Abrangentes dos portais das três cidades americanas mais populosas.

Palavra	Categoria	Freq
education	Education	2
environment	Environment	2
health	Health	1
	Health & Human Services	1
safety	Public Safety	2
recreation	Recreation	1
	Parks & Recreation	1
services	Social Services	1
	Health & Human Services	1
transportation	Transportation	2
business	Business	1
government	City Government	1
housing	Housing & Development	1
development	Housing & Development	1
social	Social Services	1
livable	A Livable and Sustainable City	1

mostradas nessa tabela são associadas às Palavras Mais Abrangentes, formando assim o conjunto de Categorias Abrangentes, ou o Subconjunto Abrangente de Categorias. Na Tabela 3.5 são apresentadas todas as categorias do Subconjunto Abrangente obtido utilizando o código descrito para essa etapa no Apêndice A. Ressaltamos que essas categorias ainda precisam ser avaliadas pelo usuário, na próxima etapa, para remover categorias iguais e com sentidos semânticos diferentes.

Tabela 3.5: Subconjunto Abrangente de Categorias obtidas dos portais das três cidades americanas mais populosas.

Education	Health & Human Services
Environment	Transportation
Health	Business
Health & Human Services	City Government
Public Safety	Housing & Development
Recreation	Housing & Development
Parks & Recreation	Social Services
Social Services	A Livable and Sustainable City

3.1.9 Escrever Categorias Abrangentes

Essa etapa escreve as Categorias Abrangentes em um arquivo *JSON*. Assim, essa é a etapa de saída das funções algorítmicas desse processo. No entanto essa saída ainda precisa passar pela última etapa do processo que é a avaliação do usuário. O arquivo *JSON* obtido na saída dessa atividade para o exemplo dos portais das três cidades americanas pode ser encontrado na Seção A.1.9 do Apêndice A.

3.1.10 Avaliar Categorias Abrangentes

A última atividade realizada pelo usuário é a avaliação do resultado final, ou seja, a avaliação do conjunto de Categorias Abrangentes, ou Subconjunto Abrangente de Categorias. Na saída das atividades de algoritmo, o conjunto de Categorias Abrangentes pode conter categorias iguais ou que possuem conteúdo semântico equivalente. Podem existir ainda, categorias que não descrevem dados com a generalização necessária para uma integração entre portais ou algumas categorias dos portais podem não ter sido alinhadas com quaisquer das Categorias Abrangentes. Assim, é de responsabilidade do usuário a avaliação final e, por consequência, a definição final do conjunto de Categorias Abrangentes. O usuário pode ainda inserir categorias nesse conjunto. O usuário deve, ao final da avaliação, editar o arquivo JSON de saída para que ele contenha somente as Categorias Abrangentes que considerar necessárias.

Na Tabela 3.6 são apresentadas as categorias do Subconjunto Abrangente após nossa avaliação de usuário. Essas categorias serão utilizadas na entrada do próximo processo para alinhamento de todas as categorias dos portais das três cidades americanas mais populosas.

Tabela 3.6: Subconjunto Abrangente de Categorias obtidas dos portais das três cidades americanas mais populosas após avaliação de usuário.

Education	Transportation
Environment	Business
Health & Human Services	City Government
Public Safety	Housing & Development
Parks & Recreation	Social Services

3.2 Alinhamento de Categorias dos Portais com o Subconjunto Abrangente

Nesta seção apresentamos o processo desenvolvido para Alinhamento de Categorias de Portais com as Categorias do Subconjunto Abrangente, obtido no processo anterior, descrito na Seção 3.1. Dessa forma podemos produzir um alinhamento de diferentes categorias de diversos portais com um único conjunto de categorias, afim de oferecer a possibilidade de implementação de uma busca integrada entre diversos portais com um único conjunto de categorias.

Esse processo pode ser executado independente do processo anterior, a Obtenção do Subconjunto Abrangente de Categorias, descrito na seção 3.1, desde que o usuário forneça, na entrada do processo, o conjunto de categorias que quer utilizar como Subconjunto Abrangente.

A Figura 3.5 apresenta o modelo do processo e suas atividades. As atividades estão descritas em sequência uma das outras. Não são apresentadas no modelo as entradas e saídas das atividades. As entradas e saídas das atividades são descritas no Guia de Implementação dos Processos, descrito no Apêndice A. Algumas atividades são realizadas pelo usuário e as demais são realizadas por algoritmo. Nesta seção vamos descrever todas as atividades do processo. Os códigos necessários para implementação das atividades algorítmicas também estão descritos no Guia do Apêndice A.

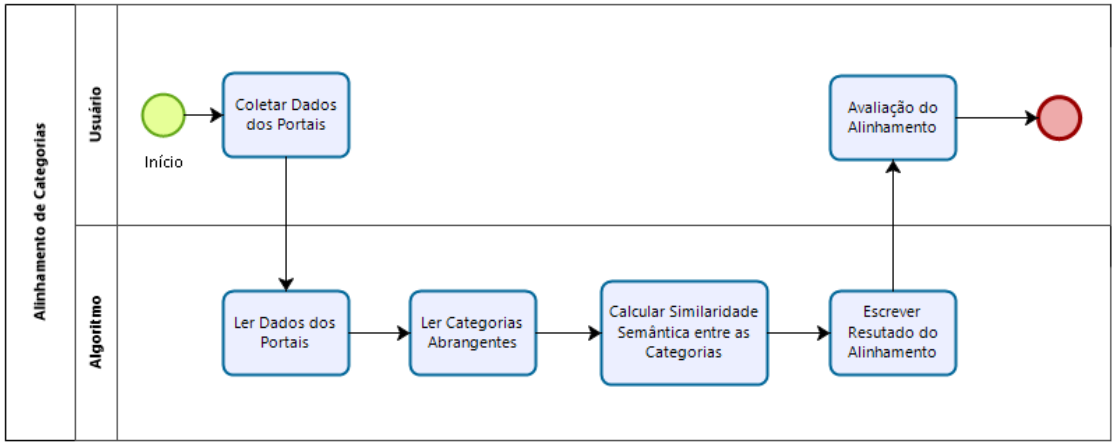

Figura 3.5: Modelo de atividades para o Alinhamento de Categorias.

3.2.1 Coletar Dados dos Portais

Essa atividade é exatamente igual a atividade descrita na seção 3.1.1. Dessa forma, essa atividade já foi descrita na referida seção. No entanto, devemos lembrar que o objetivo dessa seção é produzir um arquivo *JSON* com os dados dos portais de interesse do usuário.

3.2.2 Ler Dados dos Portais

Novamente, essa atividade é idêntica à atividade de mesmo nome no processo anterior, e foi descrita, na Seção 3.1.2. Lembramos que o objetivo dessa etapa é produzir uma lista de objetos *Portal* contendo os dados dos portais lidos do arquivo *JSON*.

3.2.3 Ler Categorias Abrangentes

Essa atividade realiza a leitura das categorias abrangentes, obtidas no processo Obtenção do Subconjunto Abrangente de Categorias, descrito na Seção 3.1. O código utilizado para realização dessa atividade está disponível e descrito na Seção A.2.3, no Guia de Implementação dos Processos, disponível no Apêndice A.

3.2.4 Calcular Similaridade Semântica entre as Categorias

O objetivo dessa atividade é o cálculo da Similaridade Semântica entre as categorias de um portal e as Categorias Mais Abrangentes, conforme descrito na Seção 2.3.1. Para isso são calculadas as similaridades entre todas as palavras que formam as categorias, conforme descrito na Seção 2.3.3. Utilizamos todos os métodos apresentados na Seção 2.3.3 para calcular a similaridade semântica entre as palavras das categorias.

Na Figura 3.6 é apresentado um fluxograma genérico que representa o cálculo da similaridade semântica entre duas categorias.

Uma categoria pode ser formada por diversas palavras e conectivos gramaticais. Para calcular a similaridade semântica entre os segmentos de textos que formam as categorias, calculamos a similaridade entre as palavras que compõem a categoria, de acordo com a formulação descrita na seção 2.3.2.

Na Equação 2.1 é apresentado o cálculo da similaridade sim entre duas sentenças T_1 e T_2, onde $maxSim(w, T_i)$ é a similaridade máxima de uma palavra da sentença T_j com todas as palavras da sentença T_i. Assim deve-se calcular a similaridade semântica entre

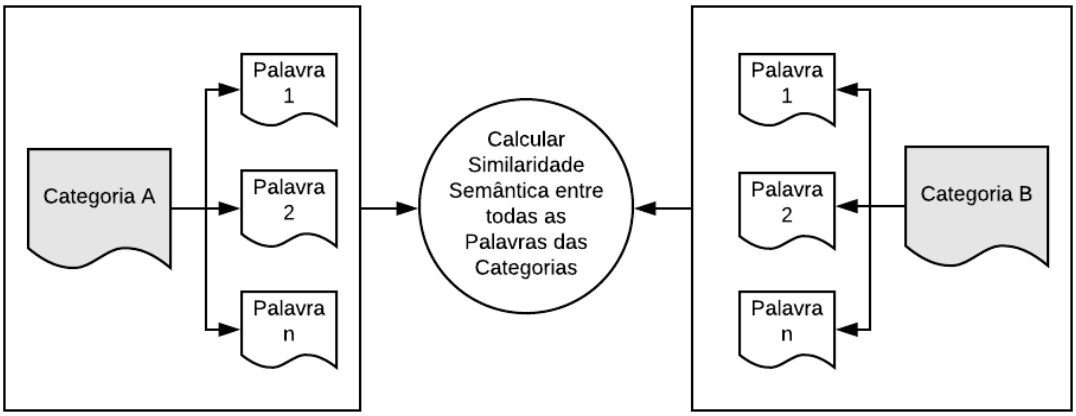

Figura 3.6: Fluxograma genérico que representa o cálculo de similaridade entre duas categorias

todas as palavras das duas categorias.

Dadas duas categorias, devemos calcular a similaridade semântica entre elas através do cálculo da similaridade semântica entre todas as palavras que formam as categorias, de acordo com a Equação 2.1.

Na Figura 3.7 é apresentado um fluxograma genérico que representa o cálculo da similaridade entre duas palavras, utilizando os seis métodos diferentes.

Como um exemplo, calculamos a similaridade semântica entre as categorias *Public Works & Engineering* e *Land Use*. Para isso, precisamos calcular a similaridade entre as palavras *public* e *land*, *public* e *use*, *works* e *land*, *works* e *use*, *engineering* e *land*, e *engineering* e *use*, conforme descrito na Seção 2.3.3.

Na Tabela 3.2.4 são apresentados os valores de similaridade para cada par de palavras das duas categorias do exemplo proposto. Nas duas primeiras colunas são mostradas as palavras as quais calculamos as similaridades. Na terceira e na quarta coluna são mostrados os *synsets* de cada palavra que apresentaram o maior valor de similaridade. E na quinta e última coluna é mostrado o valor obtido no cálculo da similaridade. Todos os valores de similaridade foram obtidos usando o método do Menor Caminho (*path similarity - path*), descrito na Seção 2.3.3.1. Cabe ressaltar que para o método utilizado nesse exemplo os valores de similaridade variam de 0 a 1. Para facilitar o entendimento do exemplo, apresentamos o resultados de apenas um método. O uso de outros métodos trará resultados diferentes nos valores das similaridades. Realizamos também um pequeno

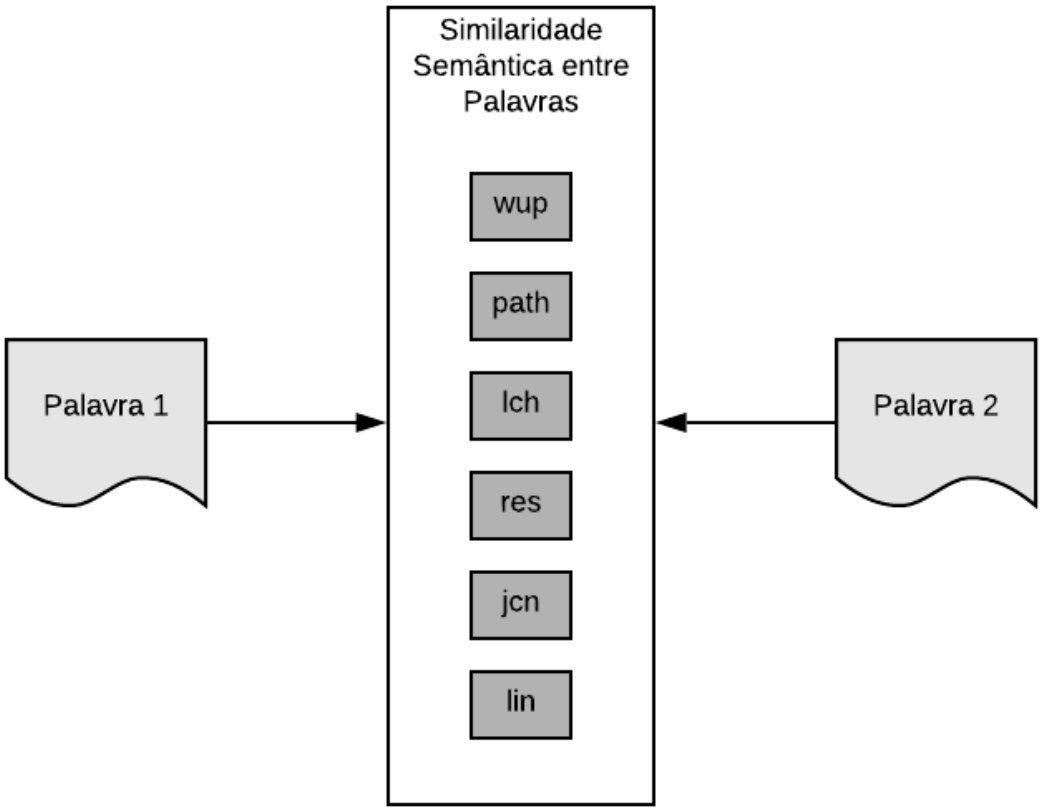

Figura 3.7: Fluxograma genérico que representa o cálculo de similaridade entre duas palavras.

processo de desambiguação dos conceitos, descrito na Seção 2.3.5, dado que, em princípio não há informação do significado da palavra na categoria. Dessa forma, calculamos a similaridade entre todos os sinônimos, ou *synsets* no *WordNet*, de cada palavra, e escolhendo a maior similaridade, de acordo com a Eq. 2.10.

Palavra 1	Palavra 2	Synset 1	Synset 2	Similaridade
Public	Land	'populace.n.01'	'nation.n.02'	0.33
Public	Use	'populace.n.01'	'function.n.02'	0.125
Works	Land	'employment.n.02'	'farming.n.02'	0.33
Works	Use	'work.v.12'	'use.v.01'	0.5
Engineering	Land	'technology.n.01'	'farming.n.02'	0.25
Engineering	Use	'technology.n.01'	'use.n.01'	0.33

Tabela 3.7: Valores de similaridades calculados as entre palavras das categorias *Public Works & Engineering* e *Land Use*.

Um *synset* é identificado com um nome de 3 partes: *word.pos.nn*, que correspondem ao sinônimo, a parte do discurso da qual o sinônimo faz parte e ordem na lista de *synsets* [36]. Podemos verificar na Tabela 3.2.4 que a maior similaridade entre os *synsets* das palavras *Public* e *Land* é 0.33. Os *synsets* que pontuam essa similaridade são: *populace.n*.01 e *nation.n*.02. Nesse caso, os *synsets* mais similares são: *populace*, primeiro substantivo na lista de *synsets* da palavra *Public*, e *nation*, segundo substantivo na lista de *synsets* da palavra *Land*. A maior similaridade entre os *synsets* das palavras *Public* e *Use* é 0.125. Os *synsets* mais similares são *populace.n*.01 e *function.n*.02. Entre as palavras *Works* e *Land*, encontramos uma maior similaridade entre os *synsets employment.n*.02 e *farming.n*.02, com um valor de 0.33. A maior similaridade calculada entre as palavras *Works* e *Use* apresenta o valor de 0.5, entre os *synsets work.v*.12 e *use.v*.01. A similaridade máxima entre os *synsets* das palavras *Engineering* e *Land* apresenta o valor de 0.25, calculada entre *technology.n*.01 e *farming.n*.02. Os *synsets* mais similares entre as palavras *Engineering* e *Use* são *technology.n*.01 e *use.n*.01, sendo a máxima similaridade de 0.33.

Após calcular as similaridades de todos os pares de palavras das categorias do exemplo, calculamos a similaridade semântica entre as categorias, de acordo com a Equação 2.1. As listas de palavras chaves das categorias *Public Works & Engineering* e *Land Use* são T_1 e T_2, respectivamente. Dessa forma, temos que:

$T1 = (Public, Works, Enginnering)$, $T2 = (Land, Use)$

$$sim(T_1, T_2) = \frac{1}{2}(\frac{0.33 + 0.5 + 0.33}{3} + \frac{0.33 + 0.5}{2}) = 0.4$$

Para a primeira palavra de T_1, *Public*, escolhemos a maior similaridade entre as palavras de T_2, *Land* ou *Use*. Podemos verificar que a maior similaridade de *Public* é com *Land*, no valor de 0.33. Dessa mesma forma procedemos para as outras palavras de T_1, e então escrevemos o primeiro termo da equação.

O segundo termo da equação é escrito tomando as palavras de T_2 e comparando as similaridades com T_1. Para a primeira palavra de T_2, *Land*, verificamos que a maior similaridade é com *Works* ou *Public*, que possuem o mesmo valor de 0.33. A segunda palavra, *Use*, possui similaridade máxima com *Works*, no valor de 0.5. Na próxima seção descrevemos o processo de alinhamento de categorias utilizando o cálculo de similaridade das palavras-chave.

A implementação do *WordNet*[36] no *NLTK*[48] fornece diversos métodos diferentes

para o cálculo de similaridade entre palavras. Cada técnica possui suas particularidades, vantagens e desvantagens, potenciais e limitações. Ao invés de avaliar o desempenho e eficiência de cada uma das técnicas, decidimos combiná-las para tomada de decisão da similaridade. Utilizamos seis métodos diferentes para o cálculo de similaridade palavra-palavra fornecidos pelo *WordNet*, que podem ser aplicados no domínio explorado neste trabalho. As técnicas para cálculo de similaridade entre palavras são: $path$, wup, lch, res, lin e jcn. Todos essas técnicas foram descritas resumidamente na Seção 2.3.3. Ao utilizar os seis métodos para compor o resultado, evitamos a análise da formulação mais adequada ou eficiente para o alinhamento proposto nesse trabalho, já que essas análises não foram alvo desse estudo e podem ser realizadas em trabalhos posteriores. O algoritmo proposto conta a quantidade de resultados iguais entre as técnicas. O resultado do alinhamento que tiver sido eleito por mais métodos é apontado como resultado final do alinhamento. Caso haja empate no número de resultados iguais entre os métodos, o algoritmo escolhe aleatoriamente o resultado final para o alinhamento.

Os cálculos de similaridade entre palavras baseados em conteúdo da informação, res, lin e jcn, precisam de um corpus para produzir a medida estatística necessária para o cálculo de similaridade. A maneira mais convencional de medir o conteúdo de informação (IC) dos sentidos da palavra é combinar o conhecimento da estrutura hierárquica de uma ontologia, como *WordNet*[36], com estatísticas sobre seu uso real no texto derivado de um grande corpus. O *Brown Corpus*[60] foi o primeiro corpus eletrônico de milhões de palavras do inglês, criado em 1961 na Brown University. Esse corpus contém texto de 500 fontes e as fontes foram categorizadas por gênero, como notícias, editorial e assim por diante [48]. Utilizamos o Brown Corpus nos métodos de cálculo de similaridade baseados em conteúdo.

3.2.5 Escrever Resultado do Alinhamento

Essa atividade escreve o alinhamento produzido na atividade anterior em um arquivo *JSON*, para então ser avaliado pelo usuário na próxima atividade.

3.2.6 Avaliação do Alinhamento

Nessa etapa, o usuário pode avaliar o alinhamento produzido pelo algoritmo, verificar sua consistência e corrigir algum alinhamento que considerar inadequado. Algumas categorias podem não ser alinhadas com nenhuma categoria abrangente, assim o usuário pode

manualmente alinhar para uma das Categorias Abrangentes.

3.3 Avaliação do Processo de Alinhamento de Categorias

Apresentamos nesta seção uma avaliação para o processo de Alinhamento de Categorias. Essa avaliação tem como objetivo comparar os resultados do alinhamento obtido pelo processo descrito nesse trabalho e o alinhamento produzido por pessoas.

Neste processo de avaliação, solicitamos que algumas pessoas realizem o alinhamento de algumas categorias dos portais com as categorias do Subconjunto Abrangente. Comparamos então o resultado do alinhamento executado por pessoas com os resultados obtidos com o processo de alinhamento. Essa avaliação pode demonstrar resultados importantes sobre o processo de alinhamento de categorias. No entanto, essa pesquisa não faz parte de nenhum processo descrito neste trabalho. No Capítulo 4 apresentamos um estudo de caso onde utilizamos essa avaliação para obter informações sobre o resultado do alinhamento.

Na Tabela 3.8 é apresentado um modelo de tabela para alinhamento de categorias por pessoas. Na primeira linha são mostradas todas as categorias do Subconjunto Abrangente. Na primeira coluna são mostradas as categorias dos portais as quais se queira alinhar e comparar com alinhamento produzido pelo processo de alinhamento. Deve-se preencher, para cada categoria dos portais, na primeira coluna, um marcador (como um X, por exemplo) para a coluna correspondente a Categoria Abrangente que o preenchedor considerar mais adequado ao alinhamento.

Tabela 3.8: Modelo de tabela para alinhamento de categorias realizado por pessoas.

	Cat. Abrang. 1	Cat. Abrang. 2	Cat. Abrang. 3	...	Cat. Abrang. n
Categoria 1					
Categoria 2					
Categoria 3					
...					
Categoria n					

Esse instrumento de avaliação foi construído de acordo com os princípios básicos que orientam uma ferramenta de Carding Sorting [61]. O Carding Sorting é uma abordagem muito utilizada que ajuda a entender o modelo mental de como as pessoas agrupam conteúdo e funcionalidades ou como interpretam o significado desses grupos de forma

que faça sentido para elas e assim aumentar a capacidade do usuário conseguir localizar informações de forma rápida dentro de um sistema [61].

3.4 Considerações Finais

Neste Capítulo descrevemos os processos para Obtenção do Subconjunto Abrangente de Categorias, seção 3.1, e para o Alinhamento de Categorias baseado em um Subconjunto Abrangente, 3.1. Também apresentamos uma avaliação do processo de alinhamento, que pode fornecer informações sobre a comparação do alinhamento produzido pelo processo e o alinhamento produzido por pessoas.

O alinhamento de categorias proposto neste capítulo pode facilitar a implementação de uma busca integrada entre diversos portais de dados abertos em um portal que coleta dados, ou catálogos, de diversos portais de dados abertos. Essa solução pode ser adicionada à ferramentas de integração de dados de diversos portais, como o CKAN Harvester [33], descrito na Seção 2.2.4.

Capítulo 4

Pesquisa Exploratória e Estudo de Caso

Neste capítulo apresentamos dois pontos fundamentais da metodologia proposta neste trabalho, a Pesquisa Exploratória, onde verificamos a problematização abordada, e o Estudo de Caso, onde aplicamos a solução proposta neste trabalho.

4.1 Pesquisa Exploratória

Nesta seção descrevemos uma Pesquisa Exploratória realizada nos portais de dados abertos de cidades americanas densamente populosas. Para realizar essa pesquisa, visitamos portais de dados abertos das cidades americanas, e coletamos informação sobre a categorização dos conjuntos de dados encontrados nesses portais. A condução dessa pesquisa foi fundamental para a verificação prática do problema o qual este trabalho apresenta uma tecnologia para solução. Os dados coletados permitiram desenvolver o Estudo de Caso onde realizamos a aplicação da solução. Vamos descrever detalhadamente a Pesquisa Exploratória nesta seção.

4.1.1 Análise

Existem algumas iniciativas para catalogação de portais de dados abertos ao redor do mundo. Uma delas é o Data Portals [59], ele contém informação de 551 portais de dados abertos em todos os continentes. O catálogo do site é disponibilizado em formatos abertos. Plataformas como o CKAN [29], também disponibilizam os catálogos dos portais por interfaces públicas. Apesar dos catálogos e ferramentas disponíveis, realizamos a Pesquisa Exploratória manualmente, pois assim, poderíamos ter mais entendimento do cenário atual, já que a amostragem de portais utilizada não dependeria de disponibilidade de

ferramentas ou catálogos acessíveis por máquina.

Visitamos os 100 portais de grandes cidades americanas, em número de habitantes, e coletamos informações sobre a utilização de categorias para distribuição dos conjuntos de dados nesses diferentes portais. Coletamos os termos, expressões e segmentos de texto utilizados para apresentar as categorias dos conjuntos de dados nos diversos portais.

Nos últimos anos, nos EUA, houve um grande esforço para o desenvolvimento de portais de dados abertos de cidades [1]. Assim, muitas cidades americanas divulgam seus dados por meio de portais na internet. Por isso, as cidades americanas, juntas, oferecem um grande número de portais que podem ser estudados em nossa pesquisa.

Os métodos de similaridade semântica necessários para a implementação da solução proposta nesse trabalho, descrita no Capítulo 3, calculam a similaridade entre palavras em língua inglesa. Assim, as categorias que deveríamos alinhar precisavam ser em língua inglesa. Dessa forma, decidimos procurar por portais de cidades americanas densamente populosas.

Na próxima seção, apresentamos o processo de amostragem desses portais.

4.1.1.1 Amostragem de Portais

Anualmente, o *United States Census* Bureau [62] através do *Census Bureau's Population Estimates Program* [63] produz estimativas populacionais para os EUA, seus estados, condados, cidades e centros urbanos. Utilizando os resultados do censo americano, e criando uma lista de cidades ordenadas por população, podemos estudar os portais das cidades americanas mais populosas. Através do *American Fact Finder* [64], que disponibiliza informações relevantes sobre o censo, listamos as maiores cidades americanas em número populacional, utilizando dados do censo de 2016. De posse dessa lista de cidades, ordenadas por tamanho em população, foi possível então iniciar o processo de pesquisa dos portais.

Para cada cidade na lista de mais populosas, procuramos seu portal através de uma busca no *Google*[65], utilizando a sequência de busca contendo o nome da cidade seguido de "open data portal". Esse processo de busca é descrito pelo Socrata[30] em seus manuais de *API*[66]. De posse dos resultados da pesquisa, utilizando a sequência descrita acima, examinamos apenas a primeira página de resultados do *Google* para verificar a existência do portal. Muitas vezes o link para o portal foi encontrado na página de internet oficial da cidade.

Na Tabela 4.1 é apresentada a lista com os nomes das 100 cidades americanas, densamente populosas, com portais de dados abertos encontrados. Na lista são apresentadas as cidades em ordem decrescente de população da esquerda para direita, onde realizamos a pesquisa sobre categorias utilizadas pelos diversos portais dessas cidades. Alguns desses portais disponibilizam dados do condado onde a cidade está geograficamente localizada.

Tabela 4.1: Amostragem dos 100 portais de cidades americanas densamente populosas.

NYC	Los Angeles	Chicago	Houston
Phoenix	Philadelphia	San Antonio	San Diego
Dallas	San Jose	Austin	Jacksonville
San Francisco	Columbus	Indianapolis	Fort Worth
Charlotte	Seattle	Denver	Washington
Boston	Detroit	Nashville	Portland
Oklahoma City	Las Vegas	Louisville	Baltimore
Tucson	Sacramento	Mesa	Kansas City
Atlanta	Long Beach	Colorado Springs	Raleigh
Miami	Virginia Beach	Omaha	Oakland
Minneapolis	Tulsa	New Orleans	Wichita
Tampa	Aurora	Honolulu	Anaheim
Santa Ana	Riverside	Lexington	St. Louis
Pittsburgh	Saint Paul	Cincinnati	Anchorage
Henderson	Greensboro	Plano	Newark
Orlando	Chula Vista	Jersey City	Durham
Laredo	Madison	Scottsdale	Glendale
Reno	Norfolk	Chesapeake	Fremont
Baton Rouge	Richmond	Boise	San Bernardino
Spokane	Birmingham	Tacoma	Oxnard
Fayetteville	Montgomery	Little Rock	Akron (County of Summit)
Grand Rapids	Salt Lake City	Huntsville	Mobile
Tallahassee	Knoxville	Worcester	Tempe
Santa Clarita	Cape Coral	Providence	Chattanooga
Santa Rosa	Sioux Falls	McKinney	ElkGrove

4.1.1.2 Coleta de Informação

De posse da lista dos 100 portais, visitamos cada portal, e então, coletávamos manualmente todas as informações necessárias para a pesquisa. Verificamos a plataforma utilizada para construir o portal da cidade. Verificamos, também, a existência de categorias para distribuir os conjuntos de dados. Coletamos todos os termos e expressões utilizadas nas categorias pelo portal. Muitas vezes os portais apresentavam as categorias em níveis de hierarquia. Assim, coletamos os termos e expressões tanto na hierarquia mais alta quanto nas hierarquias mais baixas.

É importante observar que nesse processo, apenas coletamos os dados e não realizamos nenhuma interpretação qualitativa sobre o portal pesquisado, devido ao objetivo deste trabalho de obter informação dos diferentes portais, e produzir estatísticas relacionadas a categorização dos conjuntos de dados disponíveis.

4.1.2 Resultados

Foram coletadas 976 categorias nos 100 portais das cidades americanas. Foram processadas um total de 1530 palavras utilizadas nas categorias dos conjuntos de dados nos diferentes portais. Desse total, 507 palavras são diferentes. As outras ocorrências são repetições das mesmas palavras. Não fizemos nenhuma interpretação semântica das palavras, apenas contamos ocorrências. Por isso, em nossos resultados é comum encontrarmos palavras com sentidos semânticos semelhantes. Os dados coletados dos portais nesta pesquisa estão dispostos em formato de arquivo *JSON* no Apêndice C.

Os resultados da pesquisa nos mostraram que 88% dos portais utilizam alguma forma de distribuição dos conjuntos de dados em categorias. Ou seja, apresentam alguma forma de categorização textual ao usuário. Esse percentual mostra um grande interesse de classificar os conjuntos de dados em categorias no portais de dados abertos. Essa distribuição facilita a busca e a descoberta dos conjuntos de dados por um usuário. Alguns portais utilizam termos diferentes para se referir a categorias, neles encontramos: *tópicos*, *camadas* e *dados disponíveis*.

Uma outra diferença marcante entre os portais é a quantidade de categorias utilizadas para distribuir os conjuntos de dados. Há uma grande variação de quantidade de categorias entre os diversos portais. Alguns utilizam poucas categorias. O valor mínimo encontrado foi de 3 categorias. Outros portais utilizam grandes quantidades de tópicos ou categorias, principalmente os que apresentam as categorias em estruturas hierárquicas. Nesses casos, o maior valor encontrado foi de 70 categorias. Na Figura 4.1 é apresentada a distribuição da quantidade de categorias por portal das cidades. A quantidade de categorias foi agrupada em grupos de 5. E então foram contados os números de portais que continham a quantidade de categorias referente a cada grupo.

Mais resultados desta pesquisa, como as plataformas de desenvolvimento utilizadas nos 100 portais, foram publicados em Pinto et al. [67].

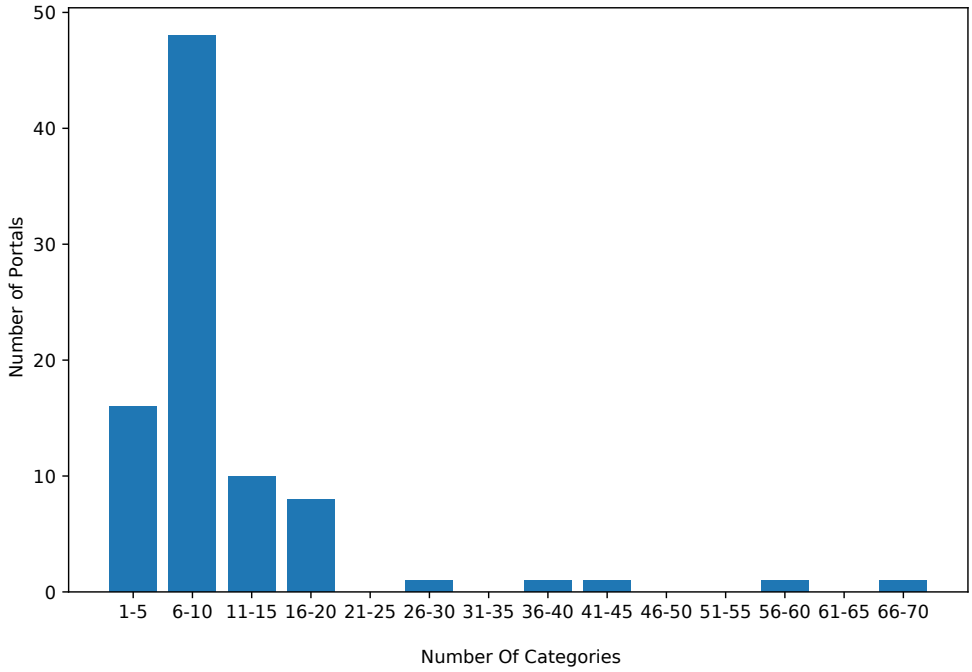

Figura 4.1: Distribuição da quantidade de categorias por número de portais.

4.2 Estudo de Caso

Nesta seção vamos descrever a aplicação dos processos propostos, como solução para o problema abordado neste trabalho, nos dados dos portais coletados na Pesquisa Exploratória.

4.2.1 Obtenção do Subconjunto Abrangente de Categorias

Nesta seção executamos o processo para Obtenção do Subconjunto Abrangente de Categorias nos dados dos portais obtidos na Pesquisa Exploratória. O objetivo desse processo é obter um conjunto de categorias que podem descrever satisfatoriamente os dados disponíveis nos diversos portais. Seguimos a descrição do processo descrito na Seção 3.1.

4.2.1.1 Coletar Dados dos Portais

Os dados dos portais utilizados nesse estudo de caso foram obtidos na Pesquisa Exploratória, descrita na seção 4.1. No Apêndice C são apresentados os dados dos 100 portais

em formato de arquivo *JSON*. Esse arquivo foi utilizado para entrada das atividades de algorítimo.

4.2.1.2 Ler Dados dos Portais

Utilizamos os código disponível na Seção A.1.2 do Guia de Implementação de Processos, descrito no Apêndice A, para realizar a leitura dos dados dos portais contidos no arquivo *JSON* do Apêndice C. Assim, foi criada uma lista de objetos *Portal* para entrada das atividades algorítmicas.

4.2.1.3 Contar Frequência de Palavras

Após o processo de leitura das categorias, para minimizar os equívocos com a semântica dos termos e expressões utilizadas, realizamos um processo de tratamento das palavras coletadas. Primeiramente, separamos os termos e expressões em palavras, com um processo de tokenização, e então removemos as *stopwords*, conforme descrito na Seção 2.3.4. Em todas as tarefas utilizamos as ferramentas disponíveis no *NLTK*[48] para Python[49]. Ainda removemos algumas palavras coletadas durante a pequisa que pudessem adicionar ruídos as estatísticas de frequência. São elas: '&', 'gis', '/' , 'kc', 'fy', 'foia', 'geo', 'city', 'data', 'go', '-' , 'houston', 'use', 'public', 'department'. São palavras frequentes e muito genéricas do domínio de portais governamentais. Denominamos essas palavras como termos auxiliares no contexto de categorias de portais. Foram coletadas 976 categorias nos 100 portais das cidades americanas densamente populosas. Após as técnicas de tokenização e remoção das *stop words*, chegamos a um total de 507 palavras diferentes.

Listamos as palavras que ocorrem nos diversos portais, calculamos a frequência de ocorrência de cada palavra, e então as listamos por ordem de frequência. Na Figura 4.2 são apresentadas as 80 palavras mais frequentes encontradas em nossa Pesquisa Exploratória. A nuvem de palavras ilustra a grande diversidade no domínio de assuntos dos conjuntos de dados encontrados nos portais. Também podemos verificar algumas palavras com sentido semântico equivalente, como por exemplo: *economy* e *economic*. Nessa Pesquisa Exploratória, não realizamos nenhuma análise semântica entre as palavras coletadas nas categorias dos portais. Dessa forma, é comum encontrarmos palavras semelhantes entre as mais frequentes.

Figura 4.2: Nuvem de palavras coletadas nas categorias dos diversos portais de dados abertos das cidades americanas.

4.2.1.4 Analisar a Abrangência das Palavras

Realizamos a Análise de Abrangência utilizando o código disponível no Guia de Implementação dos Processo, no Apêndice A, para essa atividade. No entanto, para melhor visualização dos resultados obtidos, agrupamos as 508 palavras coletadas em grupos, inicialmente de 50 palavras e, posteriormente, aumentando em 50 o número de palavras do grupo. Assim, visualizamos a abrangência das 50 palavras mais frequentes, depois das 100 palavras mais frequentes, depois das 150 palavras mais frequentes, e assim por diante. Na Tabela 4.2 é apresentada a abrangência nos diversos portais das palavras coletadas agrupadas em grupos de tamanho múltiplos de 50.

Podemos verificar na Tabela 4.2 que chegamos a 100% de cobertura dos portais utilizando 200 palavras mais frequentes. Podemos verificar ainda, que entre os conjuntos de 50, 100 e 150, a abrangência nos portais não muda, sendo fixada em 98%. Dessa forma, vamos analisar o conjunto das 50 primeiras palavras através da frequência de ocorrência desse conjunto nos diversos portais.

Tabela 4.2: Abrangência nos portais da quantidade de palavras por grupos, em tamanhos múltiplos de 50.

50	100	150	200	250	300	350	400	450	500
98%	98%	98%	100%	100%	100%	100%	100%	100%	100%

Na Figura 4.3 é apresentado o gráfico com a distribuição do percentual de frequência e abrangência nos diversos portais das 50 palavras mais frequentes. Na barra mais escura, vemos o percentual de frequência de ocorrência nos portais de cada palavra. Na barra mais clara, podemos ver o valor com que cada palavra, junto com todas as anteriores,

conforme descrito na Seção 3.1.4, contribui para a abrangência nos portais.

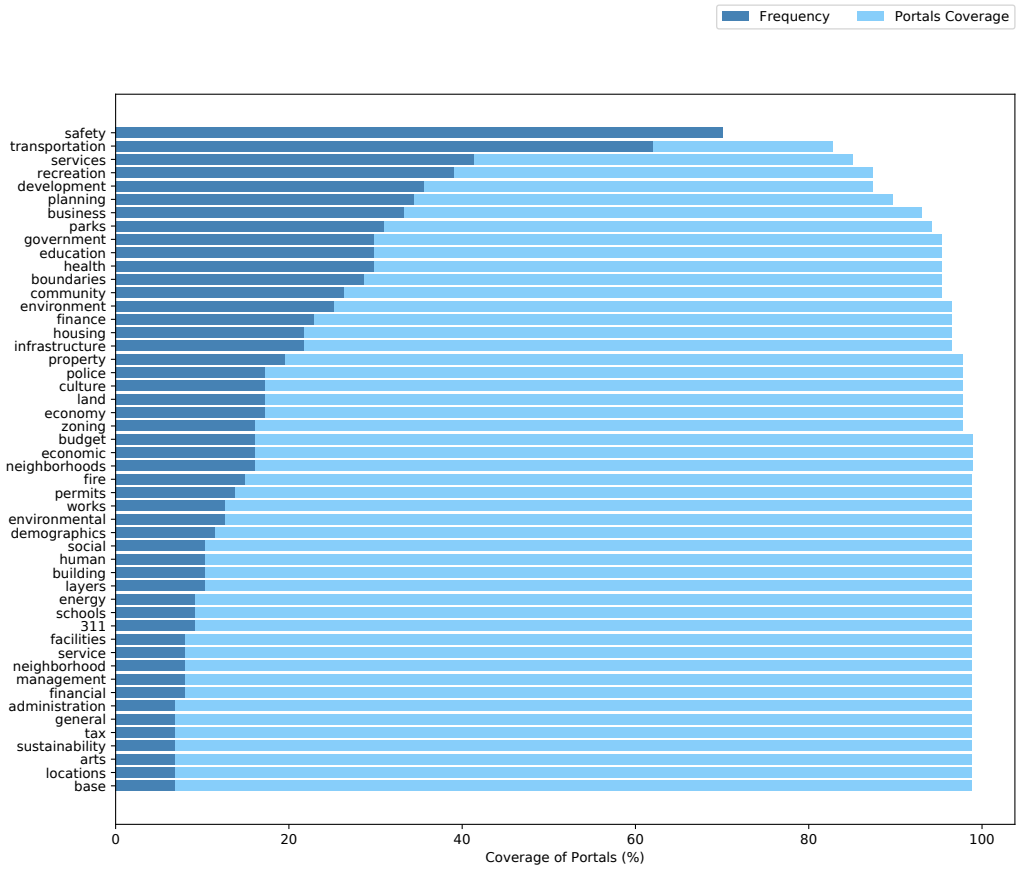

Figura 4.3: Distribuição de ocorrência e abrangência das 50 palavras mais frequentes.

No Capítulo 3, Seção 3.1.4, são apresentados todas as definições necessárias para realização da Análise de Abrangência. Na próxima etapa vamos avaliar o Parâmetro de Abrangência que foi utilizado para obtenção das Palavras Mais Abrangentes.

4.2.1.5 Definir Parâmetro de Abrangência

Podemos verificar no gráfico da Figura 4.3 que após a vigésima quarta palavra da lista, não há mais variação na quantidade acumulada da cobertura de portais. Dessa forma, podemos utilizar um percentual de abrangência de 98%. Assim, podemos compor a lista de Palavras Mais Abrangentes, nos diversos portais de cidades americanas, com o conjunto C_{24}, que contém as primeiras 24 palavras mais frequentes nos portais.

4.2.1.6 Obter Conjunto de Palavras Mais Abrangentes

Da etapa anterior, definimos o Parâmetro de Abrangência com valor de 98%, dessa forma o primeiro conjunto a satisfazer esse valor de Abrangência é o conjunto C_{24}. As palavras contidas nesse conjunto e suas frequências de ocorrência nos portais são apresentadas na Tabela 4.3.

Tabela 4.3: 24 Palavras Mais Abrangentes e suas frequências associadas.

safety	61	transportation	54
services	36	recreation	34
development	31	planning	30
business	29	parks	27
government	26	education	26
health	26	boundaries	25
community	23	environment	22
finance	20	housing	19
infrastructure	19	property	17
police	15	culture	15
land	15	economy	15
zoning	14	budget	14

4.2.1.7 Contar Frequência de Categorias

De posse da lista de Palavras Mais Abrangentes, o conjunto C_{24}, obtidas na etapa anterior, contamos a frequência de ocorrência de cada categoria onde aparecem cada uma das Palavras Mais Abrangentes.

Por exemplo, para a palavra $safety$, a palavra mais frequente encontrada nos diversos portais, as categorias que aparecem relacionadas à palavra estão dispostas ao lado de sua frequência na Tabela 4.4. Podemos verificar que a categoria mais frequente é *Public Safety*, que ocorre 45 vezes das 61 vezes que $safety$ aparece nas categorias, ou seja, *Public Safety* aparece em 73.8% das vezes em que $safety$ aparece. Assim podemos afirmar que *Public Safety* é uma categoria frequente e abrangente nos diversos portais. Dessa forma, obtemos as categorias e as frequências associadas a cada palavra da lista de Palavras Mais Abrangentes. Assim, podemos apresentar a Categoria Abrangente relacionada a cada Palavra Mais Abrangente.

Um outro exemplo que destacamos aqui são as categorias relacionadas à palavra *land*. Podemos ver na Tabela 4.5 que existem três categorias relacionadas à palavra *land* que possuem a mesma frequência, *Land Base*, *Land Use* e *Land Records*. Em caso de empate

Tabela 4.4: Lista das categorias e frequências associadas a palavra mais frequente *safety*.

Palavra	**Freq**	**Categorias**	**Freq**	**%**
safety	61	**Public Safety**	**45**	**73.8**
		Safety	5	8.2
		Community Safety	3	4.9
		Public Health & Safety	1	1.6
		Neighborhood & Safety	1	1.6
		Building and Safety	1	1.6
		Public Safety and Preparedness	1	1.6
		Public Safety and Emergency Management	1	1.6
		Public Safety & Justice	1	1.6
		Justice, Safety, Police, Crime	1	1.6
		Community Safety and Well-Being	1	1.6

Tabela 4.5: Lista das categorias e frequências associadas a palavra mais frequente *land*.

Palavra	**Freq**	**Categorias**	**Freq**	**%**
land	15	**Land Base**	**2**	**13.3**
		Land Use	**2**	**13.3**
		Land Records	**2**	**13.3**
		Real State / Land Records	1	6.7
		Land Development	1	6.7
		Housing, Land Use, and Blight	1	6.7
		Real Estate, Land Records	1	6.7
		Property and Land	1	6.7
		Land Use and Environment	1	6.7
		Future Land Use	1	6.7
		Land Use & Permits	1	6.7
		Land Layers	1	6.7

nas frequências das categorias, todas as categorias mais frequentes serão utilizadas para compor o conjunto de Categorias Abrangentes. Cabe ao usuário, na atividade final de Avaliar Categorias Abrangentes, decidir quais categorias serão utilizadas no Subconjunto Abrangente.

Como último exemplo de categorias mais frequentes, vamos apresentar as categorias relacionadas à palavra mais frequente *community*. Podemos verificar na Tabela 4.6 as categorias relacionadas a palavra *community*. As categorias mais frequentes são: *Community* e *Community Services*. Novamente, todas as categorias mais frequentes, com mesmo valor de frequência, são atribuídas ao conjunto de Categorias Abrangentes, cabendo ao usuário avaliar quais delas serão utilizadas.

Tabela 4.6: Lista das categorias e frequências associadas a palavra mais frequente *community*.

Palavra	Freq	Categorias	Freq	%
community	23	**Community**	**4**	**17.4**
		Community Services	**4**	**17.4**
		Economy and Community	3	13.0
		Community Safety	3	13.0
		Economy & Community	2	8.7
		Community Development	2	8.7
		City of Houston House & Community Development	1	4.3
		Community Risk Reduction	1	4.3
		Community Safety and Well-Being	1	4.3
		Community and Economic Development	1	4.3
		Office of Community Engagement	1	4.3

4.2.1.8 Obter Categorias Abrangentes

Nesta seção, propomos um conjunto mais frequente de categorias nos diversos portais baseado nas Palavras Mais Abrangentes. Na Tabela 4.7 é apresentada a lista de categorias mais frequentes e suas frequências. Todas essas categorias formam o Subconjunto Abrangente de Categorias.

4.2.1.9 Escrever Categorias Abrangentes

Utilizamos o código descrito na Seção A.1.9 do Guia de Implementação dos Processos, descrito no Apêndice A, para escrever as Categorias Abrangentes em um arquivo *JSON*, para que possam ser avaliadas pelo usuário na próxima etapa.

4.2.1.10 Avaliar Categorias Abrangentes

Podemos verificar na Tabela 4.7 que existem categorias semelhantes, ou repetidas, entre as categorias do conjunto obtido. É o caso das categorias *Recreation*, *Parks & Recreation*, *Culture and Recreation*, e também o caso de *Business* e *Businees & Budget*. Como são categorias com sentidos semânticos muito semelhantes, podemos escolher uma delas para representar o conceito. Também possuem sentido semântico equivalente, ou igual, as categorias das palavras *services*, *community*, *land* e *zoning*. Assim, podemos escolher apenas uma categoria entre as mais frequentes, destas palavras, para compor o Subconjunto Abrangente de Categorias.

Tabela 4.7: Categorias mais frequentemente associadas as Palavras Mais Abrangentes obtidas na Pesquisa Exploratória.

Palavra Abrangente	**Categoria Abrangente**	**Freq**
safety	Public Safety	45
transportation	Transportation	46
services	City Services	4
	Community Services	4
recreation	Recreation	8
development	Economic Development	6
planning	Planning	10
business	Business	15
parks	Parks & Recreation	6
government	Government	12
education	Education	19
health	Health	13
boundaries	Boundaries	14
community	Community	4
	Community Services	4
environment	Environment	12
finance	Finance	9
housing	Housing	19
infrastructure	Infrastructure	19
property	Property	7
police	Police	7
culture	Culture and Recreation	4
land	Land Base	2
	Land Use	2
	Land Records	2
economy	Economy	4
zoning	Planning & Zoning	4
	Zoning	4
budget	Business & Budget	3

Propomos assim, o conjunto de Categorias Abrangentes entre todas as categorias coletadas nos 100 portais visitados na Pesquisa Exploratória. Na Tabela 4.8 é apresentado o conjunto de Categorias Abrangentes após nossa análise de usuário.

As Categorias Abrangentes formam o Subconjunto Abrangente de Categorias que são entrada para o processo de Alinhamento de Categorias. Esse conjunto de categorias representa uma forma de categorização genérica para organização dos conjuntos de dados nos portais de dados abertos pesquisados.

Apresentamos no Apêndice D, o código fonte escrito em linguagem Python para as

Tabela 4.8: Categorias Abrangentes obtidas no processo para os portais das 100 cidades americanas mais populosas.

Public Safety	Community
Transportation	Environment
City Services	Finance
Parks & Recreation	Housing
Economic Development	Infrastructure
Planning	Property
Business	Police
Government	Land Use
Education	Economy
Health	Zoning
Boundaries	

chamadas das funções para realização de todas as atividades algorítmicas desse processo, disponíveis no Guia de Implementação dos Processos, no Apêndice A.

Na próxima seção, descrevemos os passos necessários para o Alinhamento de Categorias dos portais com o Subconjunto Abrangente de Categorias obtido nesta seção. Para o desenvolvimento de todas as etapas do processo utilizamos os dados obtidos na Pesquisa Exploratória com os portais de dados abertos das cidades americanas densamente populosas, descrito na Seção 4.1.

4.2.2 Alinhamento de Categorias

Nesta seção, vamos descrever o processo de Alinhamento de Categorias com os dados dos portais obtidos na Pesquisa Exploratória, descrita na Seção 4.1.

4.2.2.1 Coletar Dados dos Portais

Os dados dos portais foram coletados na Pesquisa Exploratória, com 100 cidades americanas densamente populosas. Produzimos um arquivo *JSON* contendo todos esses dados. Este arquivo está escrito no Apêndice C.

4.2.2.2 Ler Dados dos Portais

Utilizamos o código disponível na Seção A.2.2 do Guia de Implementação dos Processos, descrito no Apêndice A, para realizar a leitura dos dados do arquivo *JSON* que contém todos os dados dos portais necessários para realização desse processo.

4.2.2.3 Ler Categorias Abrangentes

Nessa atividade foi realizada a leitura das Categorias Abrangentes, obtidas no processo anterior, Obtenção do Subconjunto Abrangente de Categorias. O código utilizado para leitura do arquivo *JSON* contendo as Categorias Abrangentes está disponível na Seção A.2.3, do Guia de Implementação de Processos, no Apêndice A.

4.2.2.4 Calcular Similaridade Semântica entre as Categorias

Alinhamos todas as categorias de cada portal com as categorias do Subconjunto Abrangente, descrito na Seção 4.2.1. Esta atividade foi descrita completamente na Seção 3.2.4.

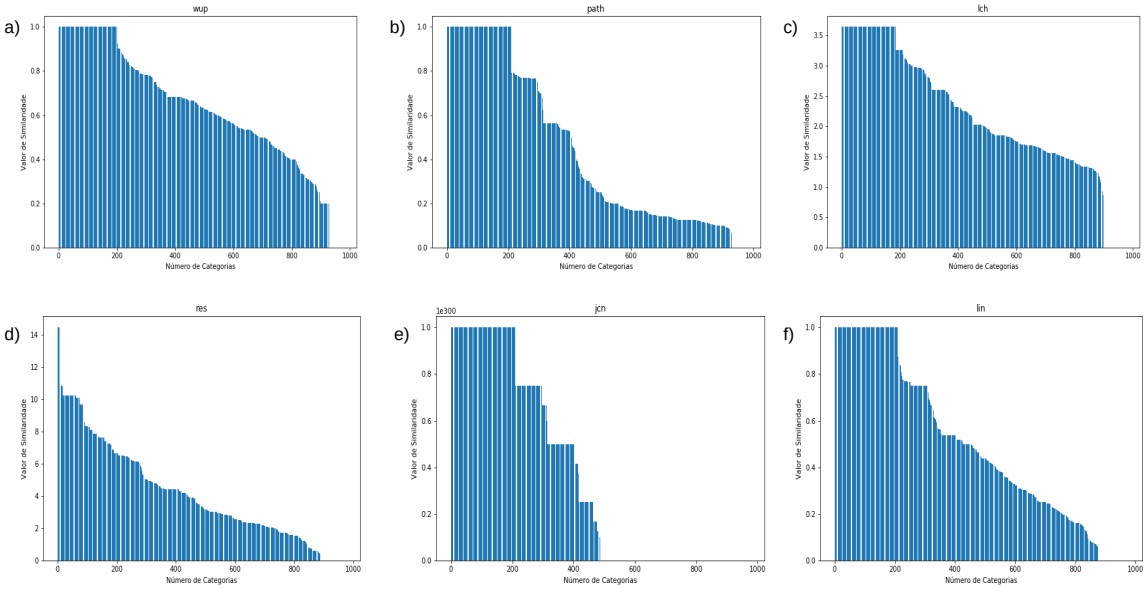

Figura 4.4: Valores absolutos de similaridade entre as categorias.

Na Figura 4.4 são apresentados os valores dos cálculos de similaridade semântica, para todas as categorias coletadas na Pesquisa Exploratória, de acordo com cada técnica para o cálculo de similaridade entre palavras, *wup*, *path*, *lch*, *res*, *jcn* e *lin*, em a), b), c), d) e e), respectivamente. Cada valor apresentado é o maior valor de similaridade obtido no alinhamento de uma determinada categoria. Ou seja, cada barra representa o maior valor de similaridade do alinhamento de uma categoria de um portal com a categoria mais similar do Subconjunto Abrangente, para todas as categorias de todos os portais. Existem algumas técnicas que não conseguem alinhar todas as categorias, já que, devido a características próprias das formulações, algumas técnicas não conseguem avaliar

a similaridade entre palavras que não fazem parte da mesma parte do discurso *POS* [41].

O resultado do alinhamento produzido por cada método foi obtido e comparado para calcular o Valor de Concordância. Definimos o Valor de Concordância como sendo a quantidade de técnicas que exibiram o mesmo resultado para um alinhamento. Para um dado alinhamento, se os seis métodos apresentarem o mesmo resultado, o Valor de Concordância é 6. Já para o caso de um alinhamento onde os seis métodos apresentam resultados diferentes, o Valor de Concordância é 1. A categoria escolhida, em um dado alinhamento, pelo maior número de métodos é apresentada como a categoria mais similar. Ou seja, a categoria escolhida para o alinhamento é a que apresenta o maior Valor de Concordância entre as formulações.

Na Tabela 4.9 são apresentados alguns resultados do alinhamento obtido pelos diferentes métodos utilizados no alinhamento das categorias obtidas na Pesquisa Exploratória. O Valor de Concordância para cada alinhamento de categoria também é mostrado. Para o exemplo *Public Works*, $(Education, 4), (CityServices, 2)$ significa que a categoria *Public Works* foi alinhada com *Education* em quatro dos seis métodos e alinhada com *City Services* nas outras duas formulações. Assim o Valor de Concordância nesse exemplo é 4, e *Education* é o resultado final do alinhamento, com concordância de quatro métodos.

Tabela 4.9: Resultados de alinhamento no cálculo da similaridade entre palavras e o Valor de Concordância entre os métodos.

Categoria	Alinhamento das Formulações	Valor de Concordância
Public Works	(Education,4),(City Services,2)	4
Culture & Arts	(Community,2), (Economy,1), (Economic Development,1), (Business,1),(Education,1)	2
Government Boundaries	(Boundaries,6)	6

Na Figura 4.5 são apresentados os resultados para o Valor de Concordância no alinhamento de todas as categorias coletadas na Pesquisa Exploratória com o Subconjunto Abrangente de Categorias obtido na Seção 3.1. O valor igual a 6 indica que todas as seis formulações utilizadas para produzir a similaridade apresentaram o mesmo resultado. O valor igual a 5 indica que 5 delas apresentaram o mesmo resultado, e assim sucessivamente. O valor igual a 0 indica que não houve alinhamento possível por nenhuma das técnicas utilizadas. No eixo vertical, são apresentados os números de categorias que foram alinhadas e apresentaram os mesmos valores para o Valor de Concordância.

Podemos verificar, ainda na Figura 4.5, que, para a maioria das categorias dos portais

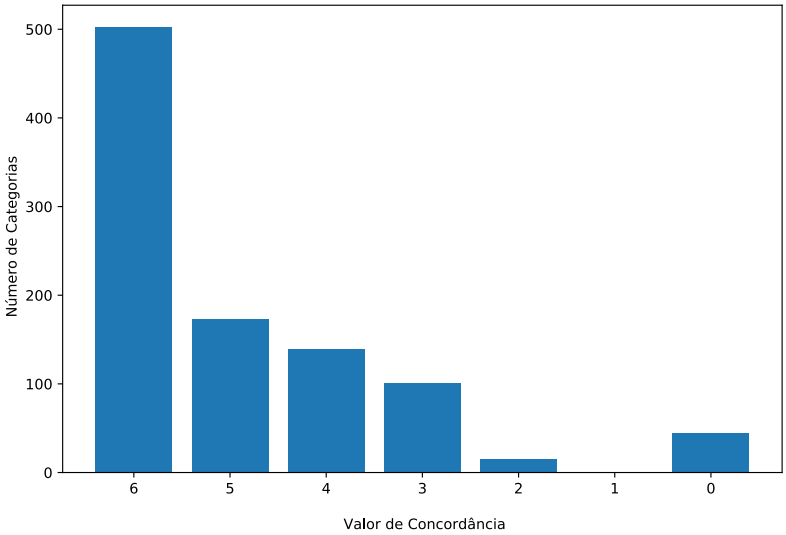

Figura 4.5: Valor de concordância entre os métodos utilizados para o cálculo de similaridade semântica.

analisados, os métodos propuseram uma mesma categoria para alinhamento, ou seja, para a maioria das categorias coletadas nos portais, os seis métodos concordaram no resultado do alinhamento. Uma outra observação é que, em todos os casos em que foi possível alinhar as categorias, pelo menos dois métodos propuseram o mesmo resultado de alinhamento. Ou seja, não houve nenhum caso onde as formulações propuseram resultados completamente diferentes umas das outras.

4.2.2.5 Escrever Resultado do Alinhamento

Utilizamos o código disponível na Seção A.2.5 do Guia de Implementação dos Processos, no Apêndice A para escrever o arquivo *JSON* que mostra a saída do resultado do Alinhamento produzido nesse processo.

No Apêndice E estão todas as chamadas das funções das atividades do processo de Alinhamento de Categorias descrito nessa seção. No Apêndice F estão escritos todos os resultados do alinhamento produzido para as categorias dos 100 portais das cidades americanas densamente populosas.

4.2.2.6 Avaliação do Alinhamento

Existem alguns casos em que os métodos não conseguem atribuir nenhuma categoria dentre o Subconjunto Abrangente como sendo a mais similar. Essas categorias podem estar descritas em contextos específicos ou utilizarem siglas na descrição, o que dificulta a análise semântica do segmento de texto utilizado. Podemos visualizar na Figura 4.5 a quantidade de categorias não alinhadas na barra correspondente ao Valor de Concordância igual a 0, onde nenhum dos métodos utilizados propuseram resultados de alinhamento. Nesses casos, a categorização pode ser feita manualmente, ou ainda, pode-se criar uma categoria para agrupar todas as categorias não alinhadas.

Na próxima seção apresentamos uma avaliação para o alinhamento produzido nesse estudo de caso. Nessa avaliação pedimos algumas pessoas para alinharem algumas das mesmas categorias que foram alinhadas por processos, obtidas na Pesquisa Exploratória.

4.2.3 Avaliação do Processo de Alinhamento

Essa avaliação, realizada com estudantes de graduação e pós-graduação nas áreas de computação, tem como objetivo comparar os resultados de alinhamento obtidos com o método apresentado nesse trabalho, com os resultados de alinhamento de categorias feito por pessoas.

Nessa avaliação, pedimos as pessoas para realizar o alinhamento de algumas categorias dos portais, obtidas na Pesquisa Exploratória apresentada neste capítulo, com as categorias do Subconjunto Abrangente, obtido na Seção 3.1. Comparamos então o resultado do alinhamento executado por pessoas com os resultados obtidos com o método de alinhamento, apresentado na Seção 3.2.

4.2.3.1 Pesquisa para Avaliação

Para realizar a avaliação, primeiramente, selecionamos algumas categorias dos portais, obtidas na Pesquisa Exploratória. As categorias foram selecionadas de acordo com o resultado do alinhamento realizado pelo processo de alinhamento. Conforme proposto na Seção 4.2.2.4, o Valor de Concordância representa a quantidade de métodos diferentes que obtiveram o mesmo resultado de alinhamento para uma dada categoria. Ou seja, como são seis métodos utilizados, um Valor de Concordância igual a 6 indica que todas as formulações concordaram. O valor 5 indica que cinco das seis formulações obtiveram o mesmo

resultado, e assim sucessivamente. Dessa forma, utilizamos o Valor de Concordância para selecionar as categorias a serem alinhadas pelo público alvo da avaliação.

Grupo I	Grupo II	Grupo III
City Government	Sanitation	Planning Building and Code Enforcement
Housing & Development	A Livable and Sustainable City	Map Features
Administration & Finance	A Safe City	Culture & Arts
Facilities & Geo. Boundaries	Service Requests	Recreation & Culture
Government Boundaries	Permitting and Licensing	Area Plans
Planning & Development	Flood Hazard	Code Interpretations
City Infrastructure	Staff Salaries	Peer Review
Energy & Environment	Auditor	Scope Statement
Budget and Finance	Public Works	Housing, Land Use, and Blight
Community Development	City Management and Ethics	Recreation and Culture

Tabela 4.10: Grupos de categorias para o alinhamento dos participantes da avaliação.

Propomos três grupos diferentes de categorias. No primeiro grupo, selecionamos, aleatoriamente, as categorias alinhadas, pelo método proposto neste trabalho, que tiveram Valor de Concordância igual a 6. Ou seja, todos os seis métodos utilizados apresentaram o mesmo resultado de alinhamento. No segundo grupo, selecionamos, aleatoriamente, as categorias que foram alinhadas com Valor de Concordância entre 3 e 5. No terceiro grupo, as categorias que foram alinhadas com Valor de Concordância menor ou igual a 2. Na Tabela 4.10 são apresentados os grupos descritos e as categorias, escolhidas aleatoriamente de acordo com cada grupo, utilizados na avaliação proposta. Essas categorias foram alinhadas, pelos participante da avaliação, com o Subconjunto Abrangente de Categorias obtido nesse estudo de caso.

Produzimos uma planilha online e compartilhada onde pedimos os participantes para responderem algumas questões para determinação do público alvo da pesquisa. Nessa mesma planilha, também apresentamos as categorias dos grupos descritos, distribuídas horizontalmente, e as categorias do Subconjunto Abrangente, distribuídas verticalmente. Essa planilha foi disponibilizada online[1] para os participantes, sendo uma planilha de mesmo modelo para cada participante. A planilha é apresentada no Apêndice G.

A avaliação foi realizada com 13 estudantes de graduação e pós-graduação nas áreas de Computação, Engenharia e Sistemas de Informação. As idades dos participantes variam

[1]https://github.com/higorspinto/urban-data-categories/blob/master/planilha_de_avaliacao_matriz.xlsx

de 25 a 40 anos. Todos concordaram em participar da pesquisa e na utilização dos dados fornecidos nos resultados deste trabalho. Também mantemos as identidades dos participantes anônimas. Na próxima seção apresentamos os resultados obtidos na avaliação e as comparações com os resultados do método descrito neste trabalho.

4.2.3.2 Resultados

Para comparar os resultados do alinhamento feito pelas técnicas diferentes utilizadas neste trabalho com o alinhamento feito pelos participantes da avaliação, aplicamos o conceito do Valor de Concordância nos resultados do alinhamento dos participantes.

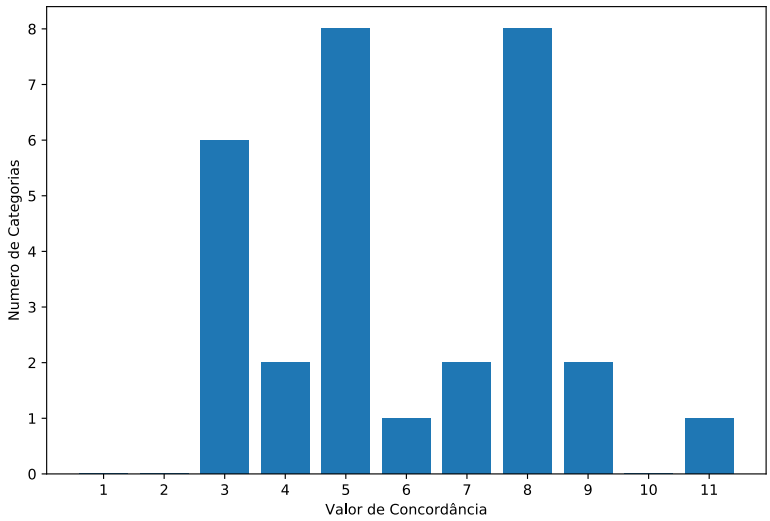

Figura 4.6: Valor de concordância entre os participantes da avaliação no alinhamento das categorias.

Dois participantes desistiram do preenchimento da planilha antes do terminar todo o preenchimento, assim, terminamos com 11 planinhas preenchidas. Como são 11 planilhas de avaliação totalmente preenchidas, o Valor de Concordância mais alto, nesse caso, é 11, onde todos os participantes apontaram o mesmo alinhamento. Valor de Concordância igual a 10 significa que dez dos onze participantes apontaram o mesmo resultado no alinhamento. E assim sucessivamente para os outros valores possíveis. Na Figura 4.6 são apresentados os resultados do Valor de Concordância obtidos com o alinhamento dos participantes da avaliação. Verificamos que em apenas um alinhamento de uma dada categoria os 11 participantes apontaram o mesmo resultado. Podemos verificar também, que o menor valor apresentado para o Valor de Concordância é 3, ou seja, no mínimo 3

participantes concordaram ao alinharem as categorias. A altura da barra representa quantidade de categorias alinhadas com aquele determinado Valor de Concordância apontado no eixo horizontal.

Dito isso, podemos destacar, de uma comparação entre as Figuras 4.5 e 4.6, que o alinhamento manual entre categorias é uma tarefa complexa e que produz resultados diferentes para diferentes pessoas que alinham as categorias. Durante a realização da avaliação, recebemos diversos relatos dos participantes sobre a dificuldade da tarefa proposta. Tivemos também 2 casos de desistência durante o preenchimento da planilha de avaliação. Isso pode mostrar uma desmotivação para realização do alinhamento durante o preenchimento da planilha. Assim, mostramos indícios da dificuldade e complexidade na tarefa de alinhar categorias.

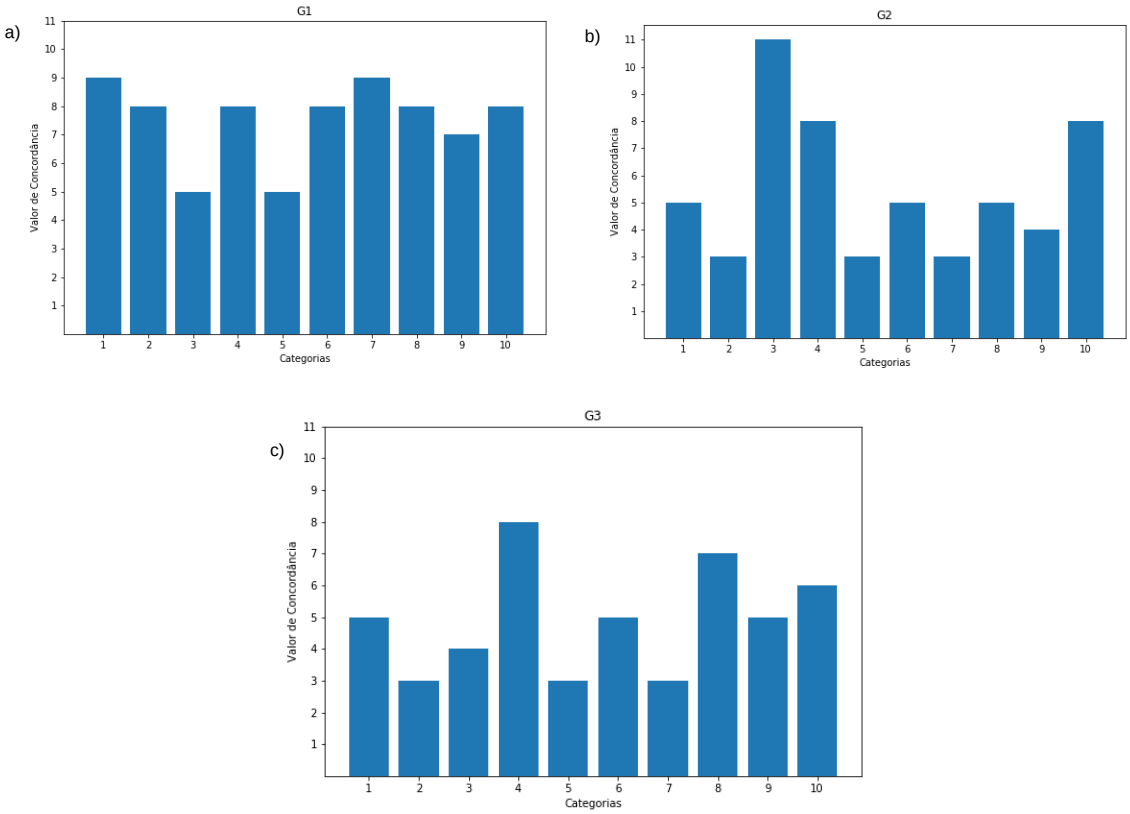

Figura 4.7: Valor de concordância entre os participantes da avaliação no alinhamento das categorias, apresentadas por grupos.

Na Figura 4.7 são apresentados os resultados dos Valores de Concordância para o alinhamento feito pelo participantes da avaliação por grupos de categorias, conforme apresentados na Tabela 4.10. Para cada categoria dos grupos I, II e III, são mostrados os

Valores de Concordância entre os participantes da avaliação, em a), b) e c), respectivamente. Verificamos que, como esperado, os valores são mais altos no grupo I, a), onde todas as técnicas utilizadas no método, descrito neste trabalho, apresentaram resultados iguais para o alinhamento das mesmas categorias. Nos grupos II, b), e III, c), os Valores de Concordância também são semelhantes à concordância obtida nas formulações, nos mesmos grupos.

Na Tabela 4.11 são apresentadas as categorias escolhidas pelos participantes para o alinhamento das categorias do grupo I, e as frequências com que foram escolhidas. Observamos que, mesmo para o grupo I, onde ocorreram as maiores concordâncias no alinhamento, o conjunto de categorias escolhidas pelos participantes para alinhar a uma categoria é diverso, mesmo que apresente categorias com conceitos semelhantes.

Grupo I	**Alinhamento dos Participantes**
City Government	('Government', 9), ('Public Safety', 1), ('City Services', 1)
Housing & Development	('Housing', 8), ('Property', 2), ('Finance', 1)
Administration & Finance	('Finance', 5), ('Economy', 3), ('Business', 2), ('Economic Development', 1)
Facilities & Geo. Boundaries	('Boundaries', 8), ('Zoning', 2), ('City Services', 1)
Government Boundaries	('Boundaries', 5), ('Government', 4), ('Property', 1), ('Zoning', 1)
Planning & Development	('Planning', 8), ('Business', 2), ('Economic Development', 1)
City Infrastructure	('Infrastructure', 9), ('Parks & Recreation', 1), ('City Services', 1)
Energy & Environment	('Environment', 8), ('Infrastructure', 2), ('City Services', 1)
Budget and Finance	('Finance', 7), ('Business', 3), ('Economy', 1)
Community Development	('Community', 8), ('Economic Development', 2), ('Education', 1)

Tabela 4.11: Alinhamento de categorias realizado pelos participantes da avaliação para as categorias do grupo I.

Na Tabela 4.12 são apresentados os resultados do alinhamento realizado pelo processo de Alinhamento de Categorias proposto neste trabalho para as categorias do Grupo I. É possível uma comparação direta dos resultados obtidos do método com os resultados do alinhamento dos participantes da avaliação, apresentados na Tabela 4.11. Para o grupo de categorias apresentado, os participantes e o método discordam em 30% dos alinhamentos realizados.

Grupo I	**Alinhamento do Método**
City Government	Government
Housing & Development	Economic Development
Administration & Finance	Finance
Facilities & Geo. Boundaries	Boundaries
Government Boundaries	Boundaries
Planning & Development	Economic Development
City Infrastructure	Infrastructure
Energy & Environment	Environment
Budget and Finance	Finance
Community Development	Economic Development

Tabela 4.12: Alinhamento de categorias realizado pelos método de alinhamento para as categorias do grupo I.

4.3 Considerações Finais

Apresentamos neste capítulo uma Pesquisa Exploratória, desenvolvida para a fundamentação do problema discutido neste trabalho, sobre a grande diversidade de termos utilizados na categorização dos portais de dados abertos. Com essa pesquisa, pretendemos fundamentar a importância de estudos e do desenvolvimento de técnicas para a integração dos conjuntos de dados entre diversos portais através das categorias. Essa Pesquisa Exploratória também possibilitou um Estudo de Caso para a aplicação dos processos desenvolvidos nesse trabalho. Esse Estudo de Caso foi apresentado e seus resultados foram discutidos neste capítulo.

Verificamos, ainda como resultados da Pesquisa Exploratória, que a grande maioria dos portais das cidades americanas mais populosas distribuem os conjuntos de dados em categorias. O que mostra o grande interesse dos editores em organizar os conjuntos de dados por temas.

Capítulo 5

Conclusão

Com o aumento do número de portais de dados abertos disponíveis na internet, e consequentemente o aumento da quantidade de conjuntos de dados disponíveis, se faz necessário o desenvolvimento de soluções e ferramentas capazes de simplificar as pesquisas e as integrações de dados nesses portais. Estudos sobre como portais de dados abertos divulgam seu dados, podem dar suporte ao desenvolvimento de ferramentas capazes de aprimorar a interação do usuário com as informações disponíveis nos portais.

Neste trabalho apresentamos uma Pesquisa Exploratória realizada com 100 portais das cidades dos EUA densamente populosas. Mostramos os resultados relacionados e categorização dos dados e aplicamos a solução desenvolvida, neste trabalho, aos portais das 100 cidades, formando assim um Estudo de Caso. Foram apresentados dois processos para integração de catálogos de dados por categorias. O primeiro processo realiza a Obtenção do Subconjunto Abrangente de Categorias. O segundo processo realiza o Alinhamento das Categorias do portais com o Subconjunto Abrangente. Apresentamos brevemente, nas próximas seções, as contribuições, limitações e os possíveis trabalhos futuros que podem ser realizados a partir das discussões apresentadas.

5.1 Contribuições

O processo de Alinhamento de Categorias produz resultados práticos na integração de dados de portais e possibilita o desenvolvimento de novas ferramentas, além de adicionar também experiência e expertise na área de integração de dados abertos. Este trabalho contribui com uma tecnologia que auxilia na tarefa de integração de dados entre diversos portais, já que possibilita uma organização entre as categorias dos portais que minimiza a estrutura de categorização e descreve, através de categorias, os dados disponíveis nesse

portais.

A Pesquisa Exploratória coletou dados de diversos portais e mostrou resultados importantes sobre a categorização de dados nesses portais. Ainda, produziu-se um catálogo com informações dos diversos portais de dados pesquisados. Esse catálogo pode ser útil em futuras pesquisas e análises.

5.2 Limitações

Os processos de alinhamento de categorias, descrito no Capítulo 3, foram aplicados apenas ao Estudo de Caso com os dados coletados na Pesquisa Exploratória. Dessa forma, não há garantias, a priori, sobre a aplicação dos processos em outros domínios de assuntos abordados pelos vários portais de dados abertos encontrados atualmente na internet.

O alinhamento foi realizado apenas com as categorias do catálogo formado na Pesquisa Exploratória. Não obtemos nenhuma informação sobre o conteúdo dos conjuntos de dados classificados nas categorias. O conteúdo dos conjuntos de dados também podem oferecer informações importantes para integração dos catálogos por categorias.

5.3 Trabalhos Futuros

Uma questão em aberto, que é resultado direto deste trabalho, é a implementação dos processos para alinhamento de categorias como uma ferramenta, utilizando tecnologias apropriadas. O CKAN[29] oferece a possibilidade de se desenvolver extensões que adicionam funcionalidades a plataforma, já que mantém todo o código da plataforma aberto. O CKAN Harvester [33] é uma extensão do CKAN que integra conjuntos de dados de diferentes portais. Essa extensão pode ser modificada para oferecer o alinhamento de categorias entre os diversos portais, utilizando os processos propostos neste trabalho.

Os conjuntos de dados nos portais estão disponíveis em formatos que facilitam a leitura por máquinas. Muitas vezes, o catálogo completo do portal está disponível também em algum formato que seja de fácil leitura por máquinas. Isso permite o processamento em massa dos conjuntos de dados nos portais [50]. Um formato muito utilizado para a disponibilização do catálogo de portais é o *RDF* [68], que viabiliza a descrição conceitual ou de modelagem de informação, de modo que essa descrição possa ser lida por máquinas. A leitura do catálogo dos portais em *RDF* pode ser incorporado aos processos de alinhamento de categorias descritos nesse trabalho. O alinhamento final produzido tam-

bém pode ser apresentado como um vocabulário em um catálogo *RDF*, que descreve os dados contidos nos diversos portais. Desse modo, a implementação dos processos em ferramentas de integração de dados de portais, como o CKAN Harvester [33], pode atender as demandas de interoperabilidade entre os portais incentivadas na abordagem de dados abertos, e também, oferecer maior facilidade ao usuário no processo de integração dos dados de diferentes portais.

No alinhamento das categorias, utilizamos alguns métodos que calculam a similaridade entre duas palavras ou conceitos. Esses métodos foram resumidamente descritos na Seção 2.3.3. Todas essas técnicas de cálculo de similaridade entre palavras foram utilizadas na composição do resultado final do alinhamento de categorias. Não realizamos, neste trabalho, análises sobre o desempenho ou eficiência das técnicas utilizadas. Vimos na Seção 4.2.2.4 que nem todas os métodos conseguem produzir resultados de alinhamento para todas as categorias analisadas, pois algumas técnicas de cálculo de similaridade não conseguem comparar palavras de classes morfológicas diferentes, por exemplo, um verbo e um substantivo. Dessa forma, realizar análises prévias sobre as classes morfológicas das palavras comparadas durante o cálculo de similaridade pode informar qual das técnicas irá propor um resultado. Análises de desempenho e eficiência podem também ser realizadas para escolha dos métodos mais adequados para o domínio de aplicação abordado. Uma análise completa, realizada entre duas técnicas diferentes para o cálculo de similaridade semântica, é encontrada na trabalho de Pawar [35]. Métricas para comparações com julgamentos humanos sobre similaridades também são explorados nesse trabalho, e também no trabalho de Seco et al. [69].

Outros trabalhos podem ainda ser conduzidos para a aplicação dos processos de alinhamento de categorias em diferentes domínios. As etapas dos processos podem ser seguidas, ou adaptadas se necessário, para aplicação em outros assuntos abordados por portais de dados abertos.

Referências

[1] THORSBY, J.; STOWERS, G. N.; WOLSLEGEL, K.; TUMBUAN, E. Understanding the content and features of open data portals in american cities. *Government Information Quarterly*, v. 34, n. 1, p. 53 – 61, 2017.

[2] DAVIES, T. *Open data, democracy and public sector reform: A look at open government data use from data.gov.uk*. Practical Participation, 2010.

[3] VAN DER WAAL, S.; WĘCEL, K.; ERMILOV, I.; JANEV, V.; MILOŠEVIĆ, U.; WAINWRIGHT, M. Lifting open data portals to the data web. In: *Linked Open Data–Creating Knowledge Out of Interlinked Data*. Springer, 2014. p. 175–195.

[4] ALÓ, C. C.; LEITE, J. D. P. Uma abordagem para transparência em processos organizacionais utilizando aspectos. *Rio de Janeiro*, 2009.

[5] DATE, C. J. *Introdução a sistemas de bancos de dados*. Elsevier Brasil, 2004.

[6] CASANOVA, M. A. *Princípios de sistemas de gerência de bancos de dados distribuídos*. Campus, 1985.

[7] VIEIRA, M. R.; FIGUEIREDO, J. M. D.; LIBERATTI, G.; VIEBRANTZ, A. F. M. Bancos de dados nosql: conceitos, ferramentas, linguagens e estudos de casos no contexto de big data. *Simpósio Brasileiro de Bancos de Dados*, 2012.

[8] BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. et al. The semantic web. *Scientific american*, v. 284, n. 5, p. 28–37, 2001.

[9] BERNERS-LEE, T. Www: Past, present, and future. *Computer*, v. 29, n. 10, p. 69–77, 1996.

[10] DADOS.GOV.BR. Cartilha técnica para publicação de dados abertos no brasil v1.0. Disponível em: <http://dados.gov.br/pagina/cartilha-publicacao-dados-abertos>. Acesso em: 20 de setembro de 2018.

[11] HENDLER, J. Data integration for heterogenous datasets. *Big data*, v. 2, n. 4, p. 205–215, 2014.

[12] BARBOSA, L.; PHAM, K.; SILVA, C.; VIEIRA, M. R.; FREIRE, J. Structured open urban data: understanding the landscape. *Big data*, v. 2, n. 3, p. 144–154, 2014.

[13] OLIVEIRA, M. I. S.; DE OLIVEIRA, H. R.; OLIVEIRA, L. A.; LóSCIO, B. F. Open government data portals analysis: The brazilian case. In: . dg.o '16. New York, NY, USA: ACM, c2016. p. 415–424.

[14] IYENGAR, S. S.; LEPPER, M. R. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology*, v. 79, n. 6, p. 995, 2000.

[15] GOV, D. Data gov. Disponível em: <https://www.data.gov/>. Acessado em: 30 de Outubro de 2018.

[16] DADOS.GOV.BR. Portal brasileiro de dados abertos. Disponível em: <http://dados.gov.br/>. Acesso em: 16 de fevereiro de 2019.

[17] LACERDA, D. P.; DRESCH, A.; PROENÇA, A.; ANTUNES JÚNIOR, J. Design science research: método de pesquisa para a engenharia de produção. *Gestão & produção*, v. 20, n. 4, p. 741–761, 2013.

[18] WAZLAWICK, R. S. Uma reflexão sobre a pesquisa em ciência da computação à luz da classificação das ciências e do método científico. *Revista de Sistemas de Informação da FSMA*, v. 6, p. 3–10, 2010.

[19] DA SILVA, E. L.; MENEZES, E. M. Metodologia da pesquisa e elaboração de dissertação. *UFSC, Florianópolis, 4a. edição*, v. 123, 2005.

[20] KNOWLEDGE INTERNATIONAL, O. The open definition. Disponível em: <http://opendefinition.org/>. Acessado em: 07 de janeiro 2018.

[21] KNOWLEDGE INTERNATIONAL, O. Open knowledge international. Disponível em: <https://okfn.org/>. Acessado em: 07 de janeiro de 2018.

[22] KNOWLEDGE INTERNATIONAL, O. what is open? Disponível em: <https://okfn.org/opendata/>. Acessado em: 07 de janeiro de 2018.

[23] KNOWLEDGE INTERNATIONAL, O. The annotated 8 principles of open government data. Disponível em: <https://opengovdata.org/>. Acessado em: 07 de janeiro de 2018.

[24] TINATI, R.; CARR, L.; HALFORD, S.; POPE, C. Exploring the impact of adopting open data in the uk government. October 2012.

[25] HARRISON, T. M.; SAYOGO, D. S. Transparency, participation, and accountability practices in open government: A comparative study. *Government Information Quarterly*, v. 31, n. 4, p. 513 – 525, 2014.

[26] MAPPER, C. City mapper - our apps for iphone and android. Disponível em: <https://citymapper.com/company>. Acessado em: 21 de janeiro de 2018.

[27] EATS, S. Safe eats - the app. Disponível em: <http://www.safeeats.toughturtle.com/>. Acessado em: 21 de janeiro de 2018.

[28] DE MENDONÇA, P. G. A.; MACIEL, C.; VITERBO, J. Visualizing aedes aegypti infestation in urban areas: A case study on open government data mashups. *Information Polity*, v. 20, n. 2, 3, p. 119–134, 2015.

[29] CKAN. Ckan - the open source da portal software. Disponível em: <https://ckan.org/>. Acessado em: 07 de janeiro de 2018.

[30] SOCRATA. Socrata: Data-drive innovation of government programs. Disponível em: <https://socrata.com/>. Acessado em: 07 de janeiro de 2018.

[31] NISO. *Understanding metadata*. 4733 Bethesda Avenue, Suite 300, Bethesda, MD 20814 USA: NISO, 2004. ISBN: 1880124629.

[32] SU, X. A text categorization perspective for ontology mapping. *Norway: Department of Computer and Information Science, Norwegian University of Science and Technology*, 2002.

[33] CKAN. Ckan data management system documentation. Disponível em: <https://docs.ckan.org/en/ckan-1.7.4/harvesting.html>. Acesso em: 05 de novembro de 2018.

[34] MANNING, C. D.; MANNING, C. D.; SCHÜTZE, H. *Foundations of statistical natural language processing*. MIT press, 1999.

[35] PAWAR, A.; MAGO, V. Calculating the similarity between words and sentences using a lexical database and corpus statistics. *CoRR*, v. abs/1802.05667, 2018.

[36] FELLBAUM, C. *Wordnet: An electronic lexical database*. Cambridge, MA: MIT Press., 1998.

[37] LI, Y.; MCLEAN, D.; BANDAR, Z. A.; O'SHEA, J. D.; CROCKETT, K. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. on Knowl. and Data Eng.*, Piscataway, NJ, USA, v. 18, n. 8, p. 1138–1150, Aug. 2006.

[38] MIHALCEA, R.; CORLEY, C.; STRAPPARAVA, C. Corpus-based and knowledge-based measures of text semantic similarity. In: . AAAI'06. AAAI Press, c2006. p. 775–780.

[39] RADA, R.; MILI, H.; BICKNELL, E.; BLETTNER, M. Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, v. 19, n. 1, p. 17–30, 1989.

[40] BULSKOV, H.; KNAPPE, R.; ANDREASEN, T. On measuring similarity for conceptual querying. In: . c2002. p. 100–111.

[41] VARELAS, G.; VOUTSAKIS, E.; RAFTOPOULOU, P.; PETRAKIS, E. G.; MILIOS, E. E. Semantic similarity methods in wordnet and their application to information retrieval on the web. In: . c2005. p. 10–16.

[42] WU, Z.; PALMER, M. Verbs semantics and lexical selection. In: . c1994. p. 133–138.

[43] LEACOCK, C.; CHODOROW, M. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, v. 49, n. 2, p. 265–283, 1998.

[44] RESNIK, P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *CoRR*, v. abs/1105.5444, 2011.

[45] LIN, D. Principle-based parsing without overgeneration. In: . c1993. p. 112–120.

[46] JIANG, J. J.; CONRATH, D. W. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[47] PEDERSEN, T.; BANERJEE, S.; PATWARDHAN, S. Maximizing semantic relatedness to perform word sense disambiguation. Technical report, Research Report UMSI 2005/25, University of Minnesota Supercomputing Institute, 2005.

[48] LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: . ETMTNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, c2002. p. 63–70.

[49] FOUNDATION, P. S. Welcome to python.org. Disponível em: <https://www.python.org/>. Acessado em: 21 de janeiro de 2018.

[50] MAALI, F.; CYGANIAK, R.; PERISTERAS, V. Enabling interoperability of government data catalogues. In: . c2010. p. 339–350.

[51] YANG, H.-C.; LIN, C. S.; YU, P.-H. Toward automatic assessment of the categorization structure of open data portals. In: . c2015. p. 372–380.

[52] SALTON, G.; MCGILL, M. J. *Introduction to modern information retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[53] HOSHIAI, T.; YAMANE, Y.; NAKAMURA, D.; TSUDA, H. A semantic category matching approach to ontology alignment. In: . c2004.

[54] OTERO-CERDEIRA, L.; RODRÍGUEZ-MARTÍNEZ, F. J.; GÓMEZ-RODRÍGUEZ, A. Ontology matching: A literature review. *Expert Systems with Applications*, v. 42, n. 2, p. 949–971, 2015.

[55] COHEN, W. W.; RAVIKUMAR, P.; FIENBERG, S. E. et al. A comparison of string distance metrics for name-matching tasks. In: . c2003. v. 2003. p. 73–78.

[56] AKBARI, I.; FATHIAN, M.; BADIE, K. An improved mlma+ and its application in ontology matching. In: . c2009. p. 56–60.

[57] SHAH, G.; SYEDA-MAHMOOD, T. Searching databases for sematically-related schemas. In: . c2004. p. 504–505.

[58] HE, W.; YANG, X.; HUANG, D. A hybrid approach for measuring semantic similarity between ontologies based on wordnet. In: . c2011. p. 68–78.

[59] DATAPORTALS.ORG. Data portals - a comprehensive list of open data portals from around the world. Disponível em: <http://dataportals.org/>. Acessado em: 30 de outubro de 2018.

[60] FRANCIS, W. N.; KUCERA, H. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

[61] SPENCER, D.; WARFEL, T. Card sorting: a definitive guide. *Boxes and Arrows*, p. 2, 2004.

[62] BUREAU, U. S. C. Census.gov. Disponível em: <https://www.census.gov/>. Acessado em: 08 de janeiro de 2018.

[63] BUREAU, U. S. C. Population estimates program. Disponível em: <https://factfinder.census.gov/faces/nav/jsf/pages/programs.xhtml?program=pep>. Acessado em: 08 de janeiro de 2018.

[64] BUREAU, U. S. C. American fact finder - annual estimates of the resident population for incorporated places of 50,000 or more, ranked by july 1, 2016. Disponível em: <https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk>. Acessado em: 08 de janeiro de 2018.

[65] GOOGLE. Gogle. Disponível em: <https://www.google.com>. Acessado em: 07 de janeiro de 2018.

[66] SOCRATA. Getting started with the soda consumer api. Disponível em: <https://dev.socrata.com/consumers/getting-started.html>. Acessado em: 07 de janeiro de 2018.

[67] PINTO, H. D. S.; BERNARDINI, F.; VITERBO, J. How cities categorize datasets in their open data portals: An exploratory analysis. In: . dg.o '18. New York, NY, USA: ACM, c2018. p. 25:1–25:9.

[68] BRICKLEY, D.; GUHA, R. V.; MCBRIDE, B. Rdf schema 1.1. *W3C recommendation*, v. 25, p. 2004–2014, 2014.

[69] SECO, N.; VEALE, T.; HAYES, J. An intrinsic information content metric for semantic similarity in wordnet. In: . c2004. v. 16. p. 1089.

APÊNDICE A – GUIA PARA IMPLEMENTAÇÃO DOS PROCESSOS

Neste guia estão descritas todas as etapas e atividades necessárias para a implementação dos processos para o alinhamento de categorias entre portais de dados abertos. Apresentamos os códigos-fonte escritos em linguagem Python [49], versão 3.7, para as atividades que podem ser realizadas por algoritmo.

Este guia é voltado para profissionais que atuam na divulgação de dados nos portais de dados abertos. Principalmente profissionais que trabalham com portais federados, ou seja, portais que coletam dados de outros portais para produzirem catálogos completos. Assim, ao seguir as atividades apresentadas, o profissional pode produzir, automaticamente, uma categorização que descreve os dados contidos nesses portais.

São descritos dois processos para o alinhamento de categorias. O primeiro é a Obtenção do Subconjunto Abrangente de Categorias, descrito na seção A.1, e o segundo é o Alinhamento de Categorias, descrito na seção A.2. Esses processos podem ser executados de forma independente, respeitando as entradas e saídas de cada processo. O primeiro processo, a Obtenção do Subconjunto Abrangente, retorna um conjunto de categorias que representa os assuntos mais frequentemente abordados nos portais de dados abertos analisados. O segundo processo, o Alinhamento de Categorias, relaciona cada categoria de cada portal com o alguma categoria do Subconjunto Abrangente. Os processos podem ser utilizados de forma independente, já que, o usuário pode querer produzir apenas um Subconjunto Abrangente de Categorias, e alinhá-las manualmente, ou ainda, pode produzir um conjunto de categorias que servirá de entrada para o processo de alinhamento. Não realizamos nenhuma etapa algorítmica de avaliação na Obtenção do Subconjunto Abrangente ou no Alinhamento. Essas avaliações devem ser realizadas pelo usuário.

A.1 Processo 1: Obtenção do Subconjunto Abrangente de Categorias

Entrada: Dados dos portais.
Saída: Arquivo *JSON* com as Categorias Abrangentes.

Nessa seção, descrevemos os passos e ações necessárias para Obtenção do Subconjunto Abrangente de Categorias. O Subconjunto Abrangente é composto por categorias que ocorrem frequentemente nos diversos portais em análise. A obtenção desse conjunto é uma etapa importante, pois são essas categorias que serão alvo do alinhamento com as categorias de cada portal. No entanto, esse processo pode não ser executado, para isso, basta o usuário fornecer para o processo de Alinhamento as categorias as quais queira que sejam alinhadas.

A Figura A.1 apresenta o modelo de atividades propostas para Obtenção do Subconjunto Abrangente de Categorias. As atividades estão descritas em sequência uma das outras, mesmo que algumas possam ser realizadas em paralelo. No entanto para melhor entendimento do processo e desse Guia, sugerirmos realizar as atividades em sequência. Ainda, para melhor entendimento do processo, não são apresentadas no modelo as entradas e saídas das atividades. No entanto, as entradas e saídas das atividades são descritas neste Guia. Nas próximas sessões apresentamos detalhadamente cada atividade, que são realizadas pelo usuário, ou por algoritmos. Nesse guia disponibilizamos os códigos-fonte em linguagem de programação Python [49], versão 3.7, para cada etapa realizada por algoritmo.

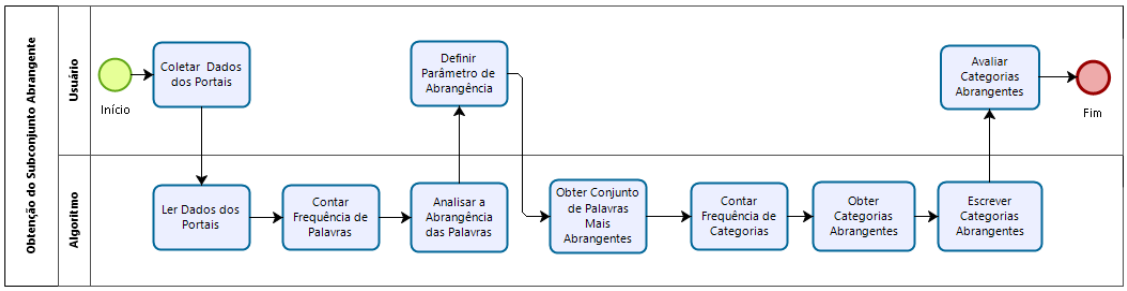

Figura A.1: Modelo de atividades para Obtenção do Subconjunto Abrangente.

A.1.1 Atividade 1: Coletar dados dos portais

Entrada: Não há entrada nessa atividade.
Saída: A saída dessa atividade é um arquivo *JSON* contendo os dados necessários dos portais.

A primeira atividade a ser realizada, pelo usuário do processo, é coletar os dados dos portais de seu interesse, ou seja, coletar as categorias dos portais onde se quer obter um Subconjunto Abrangente de Categorias. Essa coleta de dados pode ser realizada manualmente, automaticamente ou ainda, pode se utilizar alguma base de dados de portais existentes.

Coletar as categorias manualmente significa visitar cada portal que se queira analisar e então anotar todas as categorias de cada portal. Muitos dos portais de dados abertos disponibilizam seu catálogo para acesso através de *API*. Dessa forma, obtendo os catálogos por aplicação, a coleta de dados dos portais pode ser automatizada. As *URL's* dos portais devem ser conhecidas e também as chamadas das funções das *API's*. O Data Portals [59] oferece um catálogo de vários portais de dados abertos ao redor do mundo.

Os códigos fornecidos neste guia para execução das atividades realizadas por algoritmo demandam o uso de um arquivo *JSON* para entrada dos dados dos portais. Todas as formas de coleta de dados dos portais devem produzir um arquivo *JSON* semelhante para a entrada das atividades realizadas por algoritmo. O código abaixo apresenta o arquivo *JSON* utilizado na execução dos exemplos apresentados neste guia. Cada portal no arquivo *JSON* contém alguns atributos, são eles: o nome da cidade (*city*), a url do portal (*url*), as coordenadas de posicionamento da cidade (*coord*), o estilo de categorização do portal (*categorization*), a plataforma utilizada (*platform*) e todas as categorias (*categories*). Os atributos necessários para execução dos códigos-fonte fornecidos nesse guia são o nome da cidade (*city*) e as categorias dos portais (*categories*). Os demais atributos dos portais são opcionais e podem possuir valor nulo.

```
1 [
2         {
3         "city": "NYC",
4         "url": "https://opendata.cityofnewyork.us/data/",
5         "coord": "40.6643N 73.9385W",
6         "categorization": "Categories",
7         "platform": "Socrata",
8         "categories": [
9             "Business",
10            "City Government",
```

```
11              "Education",
12              "Environment",
13              "Health",
14              "Housing & Development",
15              "Public Safety",
16              "Recreation",
17              "Social Services",
18              "Transportation"
19          ]
20          },
21          {
22          "city": "Los Angeles",
23          "url": "https://data.lacity.org/browse",
24          "coord": "34.0194N 118.4108W",
25          "categorization": "Categories",
26          "platform": "Socrata",
27          "categories": [
28              "A Livable and Sustainable City",
29              "A Prosperous City",
30              "A Safe City",
31              "A Well Run City"
32          ]
33      },
34      {
35          "city": "Chicago ",
36          "url": "https://data.cityofchicago.org/browse",
37          "coord": "41.8376N 87.6818W",
38          "categorization": "Categories",
39          "platform": "Socrata",
40          "categories": [
41              "Administration & Finance",
42              "Buildings",
43              "Community",
44              "Education",
45              "Environment",
46              "Ethics",
47              "Events",
48              "FOIA",
49              "Facilities & Geo. Boundaries ",
50              "Health & Human Services",
51              "Historic Preservation",
52              "Parks & Recreation",
53              "Public Safety",
54              "Sanitation",
55              "Service Requests",
56              "Transportation"
57          ]
58      }
59  ]
```

Após preparar o arquivo *JSON* com os dados dos portais que queremos estudar, podemos realizar a leitura do arquivo para entrada das atividades realizadas por algoritmo.

A.1.2 Atividade 2: Ler Dados dos Portais

Entrada: Arquivo *JSON* contendo os dados necessários dos portais.
Saída: Uma lista de objetos do tipo Portal.

Nesta seção apresentamos, brevemente, a leitura dos dados dos portais para entrada das atividades necessárias para a Obtenção do Subconjunto Abrangente. Os códigos-fonte, escritos em linguagem *Python*, versão 3.7, são descritos e apresentados nesta seção.

Definimos uma classe denominada *Portal* para produzirmos uma lista desses objetos que contém todos os portais lidos do arquivo *JSON*. Na classe apresentada, todos os atributos estão definidos como *strings*. O atributo *categories*, que contém todas as categorias do portal, é um vetor de *strings*. Também estão definidos na classe os métodos *getters* e *setters*.

```
1  class Portal:
2
3      def __init__(self):
4          self.city = ""
5          self.url = ""
6          self.coord = ""
7          self.categorization = ""
8          self.platform = ""
9          self.categories = []
10
11     def setCity(self,city):
12         self.city = city
13
14     def getCity(self):
15         return self.city
16
17     def setUrl(self,url):
18         self.url = url
19
20     def getUrl(self):
21         return self.url
22
23     def setCoord(self,coord):
24         self.coord = coord
25
26     def getCoord(self):
27         return self.coord
28
29     def setCategorization(self, categorization):
30         self.categorization = categorization
31
32     def getCategorization(self):
33         return self.categorization
34
```

```
35     def setPlatform(self, platform):
36         self.platform = platform
37 
38     def getPlatform(self):
39         return self.platform
40 
41     def addCategorie(self,categoria):
42         self.categories.append(categoria)
43 
44     def setCategories(self, categories):
45         self.categories = categories
46 
47     def getCategories(self):
48         return self.categories
```

A leitura do arquivo *JSON* poderá ser realizada e, então, serão criados objetos do tipo *Portal* tanto quantos objetos existirem no arquivo *JSON*.

```
1 import json
2 from portal import Portal
3 
4 def readPortalsFromJsonFile():
5 
6     lstPortals = []
7     with open('portals.json', 'r') as json_file:
8         data = json.load(json_file)
9         for p in data:
10            portal = Portal()
11            portal.setCity(p["city"])
12            portal.setUrl(p["url"])
13            portal.setCoord(p["coord"])
14            portal.setCategorization(p["categorization"])
15            portal.setPlatform(p["platform"])
16            portal.setCategories(p['categories'])
17            lstPortals.append(portal)
18 
19    return lstPortals
```

A.1.3 Atividade 3: Contar a Frequência de Palavras

Entrada: Uma lista de objetos do tipo *Portal* e uma lista de palavras com pouco conteúdo semântico para o domínio abordado.

Saída: Dicionário com as palavras que compõem as categorias e suas frequências.

Uma maneira de se obter um conjunto de categorias que descrevem os conjuntos de dados de vários portais é contar a frequência de ocorrência dessas categorias nos diversos portais. Como podem ocorrer categorias semelhantes escritas de modos diferentes, uma das atividades necessárias é contar a frequência das palavras que formam as categorias. E então, obter um conjunto mais frequente de palavras, para que, cada palavra mais frequente seja associada a uma categoria, assim, produzir um conjunto de categorias mais frequentes, ou seja, o Subconjunto Abrangente. Nessa atividade de Contagem de Frequência, não realizamos nenhuma interpretação semântica das palavras, por isso, é comum aparecerem palavras com conceitos semânticos semelhantes.

Figura A.2: Categorias dos portais das três cidades americanas mais populosas.

Tabela A.1: Contagem de frequência das palavras do exemplo utilizado nesse guia.

education	2	social	1	ethics	1
environment	2	livable	1	events	1
health	2	sustainable	1	facilities	1
safety	2	prosperous	1	boundaries	1
recreation	2	safe	1	human	1
services	2	well	1	historic	1
transportation	2	run	1	preservation	1
business	1	administration	1	parks	1
government	1	finance	1	sanitation	1
housing	1	buildings	1	service	1
development	1	community	1	requests	1

A Figura A.2 apresenta as categorias dos portais das três cidades americanas mais populosas, são elas: Nova York, Los Angeles e Chicago. Vamos obter um conjunto de

Categorias Abrangentes nesses três portais. Nessa atividade, realizamos a contagem de frequência das palavras das categorias das três cidades. A Tabela A.1 apresenta a contagem de frequência de palavras desse exemplo.

Para executar a atividade de Contar a Frequência das Palavras, precisamos definir algumas funções auxiliares. A primeira delas é a função que obtém todas as categorias dos portais em estudo, denominada *allCategories*. A entrada dessa função é uma lista de objetos *Portal*.

```
1  def allCategories(lstPortals):
2
3      lstCategories = []
4
5      for portal in lstPortals:
6          categories = portal.getCategories()
7          for category in categories:
8              lstCategories.append(category)
9
10     return lstCategories
```

Definimos também a função de *tokenização* que transforma todas as categorias em palavras únicas. Também definimos a função para remoção das *stop words*, que são palavras com pouco conceito semântico no domínio abordado. Usamos a lista de *stop words* do *NLTK* para Python, e adicionalmente, são removidas palavras contidas na lista *words_to_remove*, definida pelo usuário.

```
1  from nltk.corpus import stopwords
2  from nltk.tokenize import word_tokenize
3
4  def tokenizer(lstString):
5
6      tokens = []
7      for string in lstString:
8
9          tokenize = word_tokenize(string)
10         for token in tokenize:
11             tokens.append(token)
12
13     return tokens
14
15 def removeStopWords(tokens, words_to_remove):
16
17     processed_word_list = []
18     for word in tokens:
19         word = word.lower()
20         if word not in stopwords.words("english"):
21             if word not in words_to_remove:
22                 processed_word_list.append(word)
```

```
23
24     return processed_word_list
```

Podemos agora escrever a função que faz a contagem de frequência das palavras que ocorrem nas categorias. Essa função recebe uma lista de todas as palavras que ocorrem nas categorias dos portais, incluindo palavras repetidas, já que precisamos das repetições para contagem a frequência. A função retorna um dicionário que contém a palavra como chave e a frequência de ocorrência da palavra como valor. A função ainda ordena o dicionário pelo valor da frequência.

```
1 from collections import OrderedDict
2 import operator
3
4 def frequency_word_count(lstWords):
5
6     dictWordFreq = {}
7     for word in lstWords:
8         freq = dictWordFreq.get(word)
9         if freq is None:
10            freq = 0
11        freq += 1
12        dictWordFreq.update({word : freq})
13
14    return OrderedDict(sorted(dictWordFreq.items(),
15                              key = operator.itemgetter(1),
16                              reverse = True))
```

O dicionário retornado no exemplo das três cidades americanas, utilizado nesse Guia, está impresso abaixo:

```
1 >> OrderedDict([('education', 2), ('environment', 2), ('health', 2), ('safety', 2), ('
    recreation', 2), ('services', 2), ('transportation', 2), ('business', 1), ('
    government', 1), ('housing', 1), ('development', 1), ('social', 1), ('livable', 1),
    ('sustainable', 1), ('prosperous', 1), ('safe', 1), ('well', 1), ('run', 1), ('
    administration', 1), ('finance', 1), ('buildings', 1), ('community', 1), ('ethics',
    1), ('events', 1), ('facilities', 1), ('boundaries', 1), ('human', 1), ('historic',
    1), ('preservation', 1), ('parks', 1), ('sanitation', 1), ('service', 1), ('requests
    ', 1)])
```

Devemos ressaltar que, dado que muitas palavras possuem a mesma frequência, conforme mostrado na Tabela A.1, a ordem de leitura dos portais e, consequentemente, das categorias é fator importante no algoritmo. Quando as palavras possuem a mesma

frequência, a ordenação entre elas na lista de frequência é realizada de acordo com a leitura dos dados nos portais. Assim a ordenação dos portais no arquivo de entrada é de extrema importância e deve ser levada em conta pelo usuário. No exemplo utilizado neste guia, ordenamos os portais no arquivo de entrada de acordo com a densidade populacional de cada cidade, a mais populosa aparece primeiro.

A.1.4 Atividade 4: Analisar a Abrangência das Palavras

Entrada: Uma lista de objetos *Portal*, o dicionário de palavras que compõem as categorias e suas frequências.
Saída: Dicionário que representa um conjunto de palavras como chave e a abrangência desse conjunto como valor.

A próxima atividade a ser realizada, por algoritmo, é Analisar a Abrangência das Palavras nos portais. A análise da Abrangência das Palavras nos portais será utilizada para obter um subconjunto de palavras mais frequentes, que por sua vez, será utilizado para obter o as categorias mais frequentes. Lembramos que o objetivo final desse processo é eleger um conjunto de Categorias Abrangentes, as quais serão alinhadas no processo Alinhamento de Categorias, com as categorias dos portais de interesse.

Definimos a Abrangência como a quantidade de portais onde um grupo de palavras ocorre. A Tabela A.1 mostra a lista de palavras, e suas frequências, encontradas nos portais das três cidades americanas mais populosas, utilizadas como exemplo nesse guia. Seja C_1 o conjunto formado pela primeira palavra da lista, C_2 o conjunto formado pela primeira e segunda palavras da lista, C_3, o conjunto formado pela primeira, segunda e terceira palavras da lista, e assim sucessivamente, devemos determinar um conjunto C_n que satisfaça a condição imposta pela parâmetro de abrangência definido pelo usuário. Ou seja, a partir da lista de palavras que ocorrem nos portais, em ordem de frequência e de entrada, vamos determinar o conjunto C_n que primeiro satisfaça a condição de igualdade com o parâmetro de Abrangência definido pelo usuário.

O Apêndice B apresenta a Tabela com todas os conjuntos formados pelas palavras que ocorrem nas categorias dos portais do exemplo utilizado nesse guia, em ordem de frequência e de entrada, conforme descrito na Seção A.1.3, e seus valores de abrangência nos portais do exemplo. Podemos perceber que o conjunto C_13 é o primeiro a apresentar Abrangência de 100%, o que significa, que todo este conjunto de palavras está contido nos portais utilizados no exemplo deste guia.

Para a realização dessa atividade vamos definir um conjunto de funções. A primeira delas, $portalsWithCategorias$, retorna o conjunto de portais que possuem categorias. A função recebe a lista de todos os portais e retorna uma lista de portais que possuem categorias.

```
1
2 def portalsWithCategories(portals):
3
4     portalsWithCategories = []
5
6     for portal in portals:
7
8         categories = portal.getCategories()
9         if(len(categories) > 0):
10            portalsWithCategories.append(portal)
11
12    return portalsWithCategories
```

A próxima função, *fillDictWordPortals*, retorna um dicionário que contém como chaves todas as palavras que ocorrem nos portais em estudo, ordenadas por frequência. Essas palavras e suas frequências foram obtidas na atividade de Contagem de Frequência de Palavras. Cada chave, ou palavra, tem como valor associado no dicionário uma lista de portais onde a palavra ocorre (uma ou mais vezes) nas categorias do portal. A entrada dessa função é um dicionário $palavra- > frequncia$, obtido também na atividade Contagem de Frequência de Palavras, e a lista de portais em estudo.

```
1 def fillDictWordPortals(dictWordFreq, portals):
2
3     dictWordPortals = {}
4
5     for word, freq in dictWordFreq.items():
6
7         for portal in portals:
8
9             categories = portal.getCategories()
10
11            tokens = tokenizer(categories)
12            words = removeStopWords(tokens, words_to_remove)
13
14            if word in words:
15
16                lstPortals = dictWordPortals.get(word)
17                if(lstPortals is None):
18                    lstPortals = []
19
20                lstPortals.append(portal)
21
22                dictWordPortals.update( {word : lstPortals} )
```

```
23
24     return dictWordPortals
```

As duas próximas funções definem a Análise de Abrangência das palavras encontradas nos portais.

```
1  def fillDictWordPortalsDifference(dictWordPortals):
2
3      dictWordPortalsDifference = {}
4      previousPortals = []
5
6      for word, portals in dictWordPortals.items():
7
8          differentPortals = list( set(portals) - set(previousPortals) )
9          dictWordPortalsDifference.update( {word : differentPortals} )
10
11         for portal in portals:
12             previousPortals.append(portal)
13
14     return dictWordPortalsDifference
15
16 def fillDictPortalsCoverage(dictWordFreq, portals):
17
18     dictPortalsCoverage = {}
19
20     dictWordPortals = fillDictWordPortals(dictWordFreq, portals)
21     dictWordPortalsDifference = fillDictWordPortalsDifference(dictWordPortals)
22
23     portalWithCategories = portalsWithCategories(portals)
24
25     somaPerc = 0
26     for word, portais in dictWordPortalsDifference.items():
27         perc = (len(portais) * 100 / len(portalWithCategories))
28         somaPerc = somaPerc + perc
29
30         dictPortalsCoverage.update({word : somaPerc})
31
32     return dictPortalsCoverage
```

O dicionário retornado contém a abrangência para cada conjunto de palavras formado pela lista de mais frequentes. A palavra que aparece na chave do dicionário é a única palavra diferente do conjunto representado pela chave em relação ao conjunto representado pela chave anterior.

```
1 >> {'education': 66.66666666666667, 'environment': 66.66666666666667, 'health':
      66.66666666666667, 'safety': 66.66666666666667, 'recreation': 66.66666666666667, '
      services': 66.66666666666667, 'transportation': 66.66666666666667, 'business':
      66.66666666666667, 'government': 66.66666666666667, 'housing': 66.66666666666667, '
```

```
development': 66.66666666666667, 'social': 66.66666666666667, 'livable': 100.0, '
sustainable': 100.0, 'prosperous': 100.0, 'safe': 100.0, 'well': 100.0, 'run': 100.0,
 'administration': 100.0, 'finance': 100.0, 'buildings': 100.0, 'community': 100.0, '
ethics': 100.0, 'events': 100.0, 'facilities': 100.0, 'boundaries': 100.0, 'human':
100.0, 'historic': 100.0, 'preservation': 100.0, 'parks': 100.0, 'sanitation': 100.0,
 'service': 100.0, 'requests': 100.0}
```

A.1.5 Atividade 5: Definir Parâmetro de Abrangência

Entrada: Não há entrada nessa atividade.
Saída: Parâmetro de abrangência que será utilizado para obtenção das palavras abrangentes.

O usuário deve definir, logo após verificar a análise de Abrangência das Palavras, o valor de corte na Abrangência nos portais que irá definir o conjunto de palavras mais frequentes. A definição desse valor é fundamental para estipular o conjunto de palavras mais frequentes que será utilizado na entrada da próxima etapa, Obter Conjunto de Palavras mais Abrangentes. O tamanho desse conjunto será o tamanho do conjunto de Categorias mais Abrangentes. Assim o usuário deve definir o corte de acordo com a quantidade de categorias que deseja como resultado final e também com o valor de Abrangência dos portais. Um valor alto de Abrangência produzirá um conjunto maior de categorias, um valor mais baixo produzirá um conjunto menor de categorias. No exemplo utilizado neste Guia, vamos utilizar um valor de Abrangência igual a 100%, que se mostra adequado para o tamanho do conjunto de Palavras Mais Abrangentes.

A.1.6 Atividade 6: Obter Conjunto de Palavras Mais Abrangentes

Entrada: Dicionário que representa um conjunto de palavras como chave, e a abrangência desse conjunto como valor. Parâmetro de Abrangência.
Saída: Lista de palavras com maior abrangência.

O objetivo dessa atividade é eleger um subconjunto de palavras encontradas nos portais, a partir das mais frequentes para as menos frequentes, as quais irão ocorrer em um quantidade de portais determinada pelo usuário. Após definir o valor de corte na

Abrangência dos portais podemos Obter o Conjunto de Palavras Mais Abrangentes. Essa atividade elege o conjunto de palavras mais frequentes baseado no valor da Abrangência. O primeiro conjunto C_n que satisfaz a condição do valor de Abrangência é eleito como Conjunto de Palavras Mais Abrangentes.

Para realizar essa atividade definimos a função $more_coverage_words$, que recebe o dicionário com os conjuntos C_n representados pela chave e como valor, a Abrangência nos portais desse conjunto. A função retorna uma lista de strings que represetam as palavras mais Abrangentes.

```
1  def more_coverage_words(dictPortalsCoverage, threshold):
2
3      more_coverage_words = []
4
5      for word, abrangencia in dictPortalsCoverage.items():
6
7          if abrangencia < threshold:
8
9              more_coverage_words.append(word)
10
11         elif abrangencia == threshold:
12
13             more_coverage_words.append(word)
14             return more_coverage_words
15
16         elif abrangencia > threshold:
17
18             return more_coverage_words
19
20     return more_coverage_words
```

A.1.7 Atividade 7: Contar Frequência de Categorias

Entrada: Lista de palavras mais abrangentes. Lista de portais. Lista de palavras com pouco conteúdo semântico para o domínio abordado.
Saída: Um dicionário contendo cada palavra mais abrangente e as categorias que contém cada palavra e suas frequências.

Nessa atividade, para cada palavra do conjunto C_n de palavras mais abrangentes são contadas as frequências das categorias dos portais onde ocorrem cada palavra. Assim, para cada palavra mais abrangente, a categoria associada deverá ser a categoria mais frequente.

A função mostrada abaixo, $fillDictWordCategoryFreq$, determina a frequência de cada categoria onde ocorre cada uma das palavras mais abrangentes. Ela recebe uma lista de strings com as palavras mais abrangentes, a lista de todos os portais em análise e a lista de palavras que possuem pouca semântica do domínio, definida pelo usuário, para serem descartadas.

```
1  def fillDictWordCategoryFreq(more_coverage_words, portals, words_to_remove):
2
3      dictWordCategoryFreq = {}
4      for word in more_coverage_words:
5
6          dictFreq = {}
7          for portal in portals:
8
9              categories = portal.getCategories()
10
11             for category in categories:
12
13                 lst = []
14                 lst.append(category)
15                 tokens = tokenizer(lst)
16                 words = removeStopWords(tokens, words_to_remove)
17
18                 if word in words:
19                     freq = dictFreq.get(category)
20
21                     if freq is None:
22                         freq = 0
23
24                     freq += 1
25
26                     dictFreq.update( {category : freq} )
27
28         dictFreqOrd = OrderedDict(sorted(dictFreq.items(),
29                                          key = operator.itemgetter(1),
30                                          reverse = True))
31         dictWordCategoryFreq.update( {word : dictFreqOrd} )
32
33     return dictWordCategoryFreq
```

A função retorna um dicionário que como chave contém cada palavra mais abrangente e como valor um outro dicionário, que contém cada categoria associada à palavra mais abrangente e suas frequências. O dicionário retornado para o exemplo neste Guia segue abaixo:

```
1 >> {'education': OrderedDict([('Education', 2)]), 'environment': OrderedDict([('
    Environment', 2)]), 'health': OrderedDict([('Health', 1), ('Health & Human Services',
     1)]), 'safety': OrderedDict([('Public Safety', 2)]), 'recreation': OrderedDict([('
    Recreation', 1), ('Parks & Recreation', 1)]), 'services': OrderedDict([('Social
```

```
    Services', 1), ('Health & Human Services', 1)]), 'transportation': OrderedDict([('
    Transportation', 2)]), 'business': OrderedDict([('Business', 1)]), 'government':
    OrderedDict([('City Government', 1)]), 'housing': OrderedDict([('Housing &
    Development', 1)]), 'development': OrderedDict([('Housing & Development', 1)]), '
    social': OrderedDict([('Social Services', 1)]), 'livable': OrderedDict([('A Livable
    and Sustainable City', 1)])}
```

A.1.8 Atividade 8: Obter Categorias Abrangentes

Entrada: Dicionário que contém cada palavra abrangente, suas categorias e frequências.
Saída: Um dicionário contendo cada palavra abrangente e a categoria mais frequente associada.

Nesse etapa simplesmente são associadas as categorias mais frequentes à cada palavra mais abrangente. A função $fillDictWordFrequentlyCategories$ recebe o dicionário retornado na etapa anterior, da função $fillDictWordCategoryFreq$, e faz a contagem de frequência das categorias para obter as mais frequentes. Caso existem categorias com a mesma frequência, o algoritmo elege as categorias com mesma frequência como Categorias Abrangentes. Na próxima etapa o usuário terá a possibilidade de avaliar o conjunto de categorias abrangentes e excluir categorias com sentidos semânticos equivalentes.

```
1  def fillDictWordFrequentlyCategories(dictWordCategoryFreq):
2
3      dictWordCategoriaFrequente = {}
4      for target, dictFreq in dictWordCategoryFreq.items():
5
6          maiorFreq = 0
7          for categoria, freqB in dictFreq.items():
8
9              if freqB > maiorFreq:
10                 maiorFreq = freqB
11
12         lstCategorias = []
13         for categoria, freqB in dictFreq.items():
14
15             if(freqB == maiorFreq):
16                 lstCategorias.append(categoria)
17
18         dictWordCategoriaFrequente.update({target : lstCategorias})
19
20     return dictWordCategoriaFrequente
```

A função retorna um dicionário que contém como chave cada palavra mais abrangente e como valor a lista de categorias eleitas como mais abrangentes para cada palavra. O dicionário retornado no exemplo utilizado neste guia segue abaixo:

```
1 >> {'education': ['Education'], 'environment': ['Environment'], 'health': ['Health', '
    Health & Human Services'], 'safety': ['Public Safety'], 'recreation': ['Recreation',
    'Parks & Recreation'], 'services': ['Social Services', 'Health & Human Services'], '
    transportation': ['Transportation'], 'business': ['Business'], 'government': ['City
    Government'], 'housing': ['Housing & Development'], 'development': ['Housing &
    Development'], 'social': ['Social Services'], 'livable': ['A Livable and Sustainable
    City']}
```

A.1.9 Atividade 9: Escrever Categorias Abrangentes

Entrada: Um dicionário contendo cada palavra abrangente e a categoria mais frequente associada.
Saída: Um arquivo *JSON* contendo todas as categorias abrangentes.

Essa etapa escreve todas as categorias abrangentes em um arquivo *JSON* para que possam ser avaliadas pelo usuário na próxima etapa. A função $write_{c}ategories$ escreve todas as categorias abrangentes no arquivo $most_{c}overage_{c}ategories.json$. O código abaixo apresenta a declaração da função.

```
1 def write_categories(dictWordFrequentlyCategories):
2
3     categories_lst = get_categories_from_dict_word_frequently(
          dictWordFrequentlyCategories)
4
5     file = open('most_coverage_categories.json', 'w')
6
7     s = json.dumps(categories_lst, indent=4, ensure_ascii=False).encode('utf8').decode('
          latin1')
8
9     file.writelines(s)
10    file.close()
```

O arquivo *JSON* com as Categorias Abrangentes obtidas com o exemplo utilizado nesse Guia, dos portais das três cidades americanas mais populosas, é apresentado abaixo:

```
1 [
2     "Education",
3     "Environment",
```

```
4     "Health",
5     "Health & Human Services",
6     "Public Safety",
7     "Recreation",
8     "Parks & Recreation",
9     "Social Services",
10    "Health & Human Services",
11    "Transportation",
12    "Business",
13    "City Government",
14    "Housing & Development",
15    "Housing & Development",
16    "Social Services",
17    "A Livable and Sustainable City"
18 ]
```

A.1.10 Atividade 10: Avaliar Categorias Abrangentes

Entrada: Um arquivo *JSON* contendo todas as categorias abrangentes.

Saída: Um arquivo *JSON* contendo todas as categorias abrangentes, após avaliação do usuário.

Tabela A.2: Categorias Abrangentes obtidas na etapa Obter Categorias Mais Abrangentes.

Palavra	Categorias
education	Education
environment	Environment
health	Health, Health & Human Services
safety	Public Safety
recreation	Recreation, Parks & Recreation
services	Social Services, Health & Human Services
transportation	Transportation
business	Business
government	City Government
housing	Housing & Development
development	Housing & Development
social	Social Services
livable	A Livable and Sustainable City

Esta etapa permite ao usuário avaliar o conjunto de categorias mais abrangentes obtido na etapa anterior. Essa etapa é importante pois nesse processo não são realizados nenhuma análise semântica das categorias dos portais e nem das categorias obtidas como

as mais abrangentes. Assim, a exclusão de categorias com sentido semântico semelhante deve ser feita pelo usuário.

A Tabela A.2 apresenta as categorias obtidas na etapa Obter Categorias Mais Abrangentes para o exemplo utilizado nesse Guia. Verificamos que existem categorias repetidas e com sentido semântico equivalentes. Podemos remover ainda categorias que não possuem um sentido semântico explícito, como por exemplo *A Livable and Sustainable City*. Isso irá produzir melhores resultados no algoritmo de alinhamento. Dado isso, podemos escolher um subconjunto dessas categorias para a formação das Categorias Abrangentes. A Tabela A.3 apresenta as categorias selecionadas nessa etapa para formação do Subconjunto Abrangente de Categorias, objetivo final desse processo. O usuário deve realizar essa avaliação de acordo com o domínio dos portais utilizados na análise.

Tabela A.3: Categorias Abrangentes obtidas após avaliação de usuário.

Categorias Abrangentes
Education
Environment
Health
Public Safety
Recreation
Social Services
Transportation
Business
City Government
Housing & Development

Sugerimos ao usuário a edição do arquivo *JSON* obtido na etapa anterior e a gravação do arquivo com um nome diferente, como por exemplo, *most_coverage_categories_edited.json*. Abaixo, apresentamos o código do arquivo *JSON* após nossa avaliação das categorias mais abrangentes.

```
1 [
2     "Education",
3     "Environment",
4     "Health",
5     "Public Safety",
6     "Parks & Recreation",
7     "Social Services",
8     "Transportation",
9     "Business",
10    "City Government",
11    "Housing & Development"
12 ]
```

Abaixo, apresentamos o código com as chamadas de todas as funções do Processo 1 e as definições das variáveis necessárias para as funções descritas nessa Guia.

```
print("\n")

lstPortal = readPortalsFromJsonFile()
print("Number of Portals: " + "{0}".format(len(lstPortal)))

lstCategories = allCategories(lstPortal)
print("Number of Categories: " + "{0}".format(len(lstCategories)))

tokens = tokenizer(lstCategories)
print("Number of tokens in categories: " + "{0}".format(len(tokens)))

words_to_remove = ["&","gis","/","kc","fy","foia","geo","city","data","go",
                   "-",",","houston","use","public","department",".]

lstWords = removeStopWords(tokens, words_to_remove)
print("Number of words in categories: " + "{0}".format(len(lstWords)))
print("\n")

dictWordFreq = frequency_word_count(lstWords)
print(dictWordFreq)
print("\n")

dictPortalsCoverage = fillDictPortalsCoverage(dictWordFreq, lstPortal)
print(dictPortalsCoverage)
print("\n")

trheshold = 100
more_coverage_words = more_coverage_words(dictPortalsCoverage, trheshold)
print(more_coverage_words)
print("\n")

dictWordCategoryFreq = fillDictWordCategoryFreq(more_coverage_words, lstPortal,
    words_to_remove)
print(dictWordCategoryFreq)
print("\n")

dictWordFrequentlyCategories = fillDictWordFrequentlyCategories(dictWordCategoryFreq)
print(dictWordFrequentlyCategories)
print("\n")

write_categories(dictWordFrequentlyCategories)
```

A.2 Processo 2: Alinhamento das Categorias dos Portais

Entrada: Dados dos portais. Categorias abrangentes.
Saída: Um arquivo *JSON* contendo todo alinhamento produzido pelo processo.

Nesse processo é realizado o Alinhamento das Categorias dos portais com as Categorias Abrangentes obtidas no processo anterior, descrito na seção A.1. Esse processo alinha todas as categorias dos portais contidos no arquivo de entrada com as categorias abrangentes, as quais queremos alinhar. O Alinhamento de Categorias é o processo de objetivo principal desse Guia. Ele pode ser executado independente do processo anterior, a Obtenção do Subconjunto Abrangente de Categorias, descrito na seção A.1, desde que o usuário forneça, na entrada do processo, o conjunto de categorias que quer utilizar como Subconjunto Abrangente.

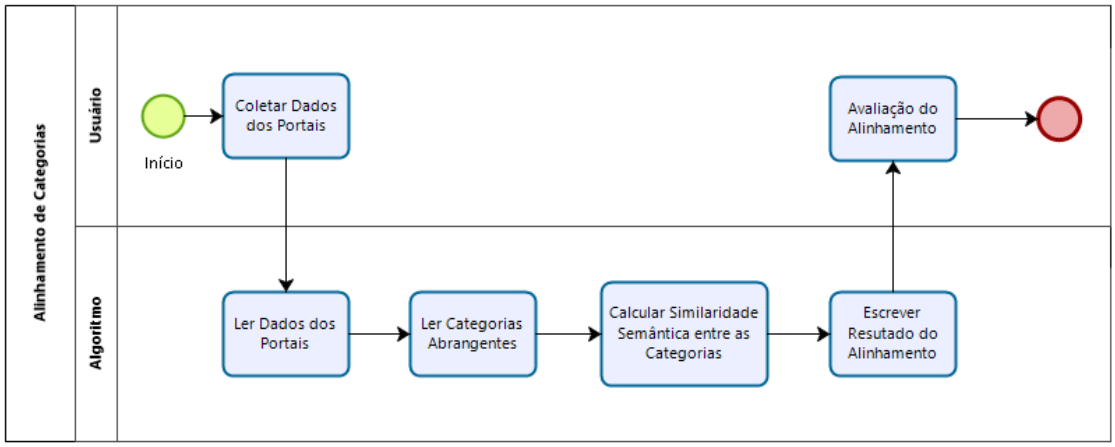

Figura A.3: Modelo de atividades para o Alinhamento de Categorias.

A Figura A.3 apresenta o modelo do processo e suas atividades. As atividades estão descritas em sequência uma das outras, mesmo que algumas possam ser realizadas em paralelo. No entanto para melhor entendimento do processo e desse Guia, sugerirmos realizar as atividades em sequência. Ainda, para melhor entendimento do processo, não são apresentadas no modelo as entradas e saídas das atividades. No entanto, as entradas e saídas das atividades são descritas neste Guia. Algumas atividades são realizadas pelo usuário e as demais são realizadas por algoritmo. Neste guia vamos apresentar todas as atividades e os códigos, escritos em *Python* 3.7, necessários para realização das atividades

algorítmicas.

A.2.1 Atividade 1: Coletar dados dos portais

Entrada: Não há entrada nessa etapa.
Saída: Arquivo *JSON* com todos os dados dos portais necessários para o alinhamento.

A primeira atividade a ser realizada, pelo usuário do processo, é coletar os dados dos portais de seu interesse, ou seja, coletar as categorias dos portais. Essa coleta de dados pode ser realizada manualmente, automaticamente ou ainda, pode se utilizar alguma base de dados de portais existentes. Essa atividade é idêntica a atividade descrita na Seção A.1.1, por isso não a descrevemos novamente. Lembramos que o objetivo dessa atividade é a criação, ou obtenção, de um arquivo *JSON* com os dados necessários dos portais. Um exemplo do arquivo gerado, como exemplo para este Guia, também é descrito na Seção A.1.1. Assim, seguiremos diretamente para a próxima atividade.

A.2.2 Atividade 2: Ler Dados dos Portais

Entrada: Arquivo *JSON* contendo os dados necessários dos portais.
Saída: Uma lista de objetos do tipo Portal.

Nessa seção apresentamos a leitura dos dados dos portais para entrada das atividades. Novamente, essa atividade é idêntica a atividade descrita na Seção A.1.2. Os códigos da classe criada, *Portal*, e para a leitura do arquivo *JSON* com os dados dos portais também são apresentados na Seção A.1.2. Dessa forma seguiremos para a descrição da próxima atividade.

A.2.3 Atividade 3: Ler Categorias Abrangentes

Entrada: Arquivo *JSON* contendo as categorias abrangentes.
Saída: Uma lista de strings contendo as categorias.

Essa atividade realiza a leitura das categorias abrangentes, obtidas no processo anterior, Obtenção do Subconjunto Abrangente de Categorias, descrito na Seção A.1. As categorias abrangentes podem também serem fornecidas pelo usuário, assim, não havendo

a necessidade de realizar o Processo 1.

O código abaixo realiza a leitura do *JSON* e cria a lista de Categorias Abrangentes que será entrada das próximas atividades.

```
import json

def readCategoriesFromJsonFile():

    lstCategories = []

    with open('most_coverage_categories_edited.json', 'r') as json_file:
        data = json.load(json_file)
        for p in data:
            lstCategories.append(p)

    return lstCategories
```

A.2.4 Atividade 4: Calcular Similaridade Semântica entre as categorias.

Entrada: Lista de objetos do tipo Portal. Lista de Categorias Abrangentes.
Saída: Dicionário contendo o nome dos portais, suas categorias e o alinhamento produzido para cada categoria do portal.

Nessa atividade são efetivamente alinhadas as categorias de cada portal com as categorias presentes no conjunto de Categorias Abrangentes. Utilizamos a similaridade semântica entre as categorias pra produzir o alinhamento. Calculamos, para cada categoria de um portal, a similaridade semântica com cada categoria da lista de Categorias Abrangentes. A Categoria Abrangente que produzir a maior similaridade semântica é alinhada com a categoria do portal.

Uma abordagem frequentemente explorada para o cálculo da Similaridade Semântica é calcular a similaridade entre as palavras que formam cada segmento de texto e então produzir um resultado para a similaridade entre os dois segmentos de texto. Dessa forma, propomos uma adaptação do método apresentado por Mihalcea et al. [38]. Assim, modelamos a semelhança semântica dos segmentos de texto como uma função da semelhança semântica de suas palavras. E então determinamos as similaridades das palavras existentes nas duas sentenças. De modo que, dadas duas sentenças T_1 e T_2, para cada palavra na sentença T_1 calcula-se a maior similaridade dessa palavra com todas as outras

palavras da sentença T_2. O mesmo processo é feito para T_2. Esse cálculo é descrito mais detalhadamente na Seção 2.3.2.

Na Figura A.4 é apresentado um fluxograma genérico que representa o cálculo da similaridade semântica entre duas categorias. Na Eq. A.1 é apresentada a equação que calcula a similaridade sim entre duas sentenças T_1 e T_2, onde $maxSim(w, T_i)$ é a similaridade máxima de uma palavra da sentença T_j com todas as palavras da sentença T_i. Assim deve-se calcular a similaridade semântica entre todas as palavras das duas categorias.

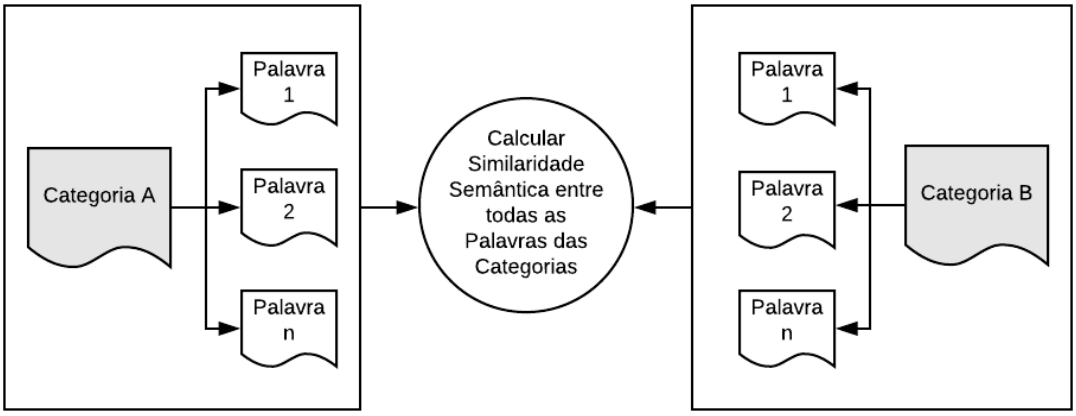

Figura A.4: Fluxograma genérico que representa o cálculo de similaridade entre duas categorias.

$$sim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{w \in T_1} maxSim(w, T_2)}{|T_1|} + \frac{\sum_{w \in T_2} maxSim(w, T_1)}{|T_2|}\right) \quad \text{(A.1)}$$

Existem diversos métodos disponíveis para o cálculo da similaridade semântica entre duas palavras. Neste trabalho utilizamos seis deles para compor o resultado final da similaridade semântica entre duas categorias, são eles: **Menor caminho (path similarity - path)** [39], **Wu and Palmer (wup similarity - wup)** [42], **Leacock and Chodorow (lch similarity - lch)** [43], **Resnik (res similiraty - res)** [44], **Jiang et al. (jcn similarity - jcn)** [46] e **Lin (lin similarity - lin)** [45]. Dessa forma, calculamos a similaridade semântica entre palavras pelos seis métodos diferentes. Na Figura A.5 é apresentado um fluxograma genérico que representa o cálculo da similaridade entre duas palavras, utilizando os seis métodos diferentes.

Após calcular a similaridade semântica ente todas as palavras de duas categorias, pode-se calcular a similaridade semântica entre as duas categorias através da Equação

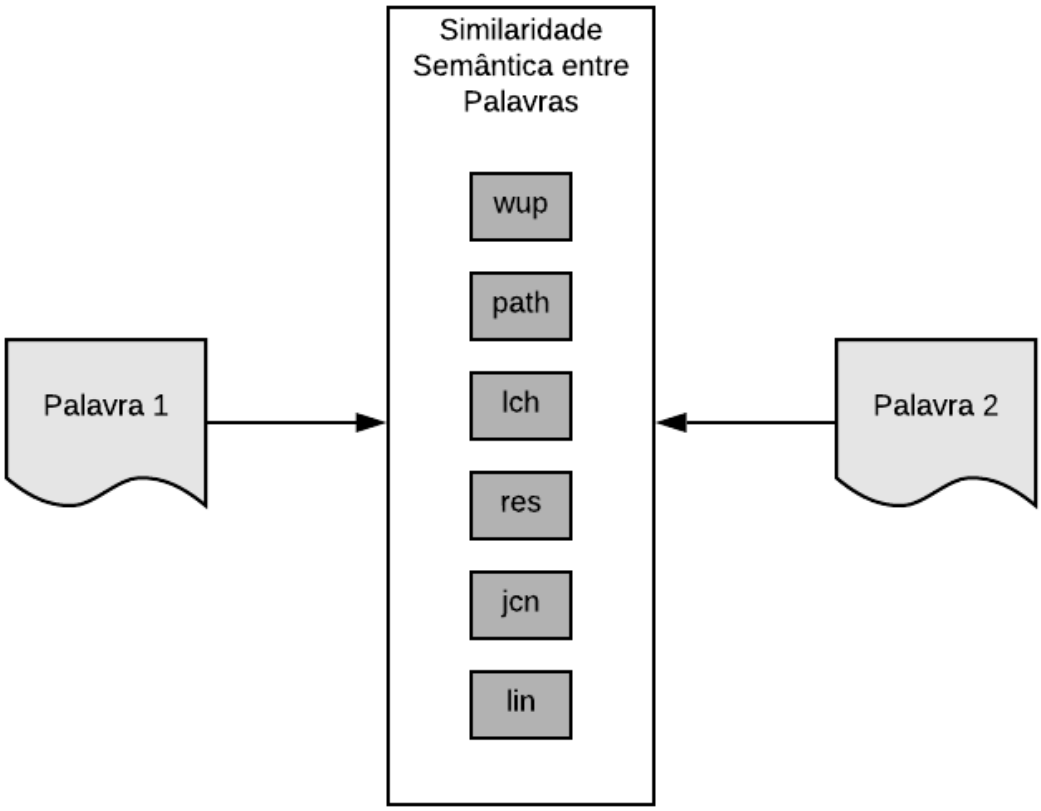

Figura A.5: Fluxograma genérico que representa o cálculo de similaridade entre duas palavras.

A.1. Nós utilizamos seis métodos diferentes para calcular a similaridade entre palavras, assim, temos seis resultados diferentes para o alinhamento de uma dada categoria de um portal. Esses seis resultados podem ser iguais ou completamente diferentes. Dessa forma definimos um modo de avaliar o resultado de todos os métodos para produzir o resultado final. O alinhamento mais frequentemente produzido pelos diferentes métodos é escolhido como resultado final. Por exemplo, se três dos seis métodos apresentam resultados iguais, e os outros três métodos apresentam resultados diferentes, o resultado final do alinhamento é o mesmo resultado produzido igualmente pelos três métodos. Quando há empate no número de resultados iguais produzidos pelos métodos, por exemplo, três deles propõem o mesmo resultado do alinhamento, e os outros três propõem um outro resultado, o resultado final é escolhido aleatoriamente entre os dois resultados possíveis.

Para realização dessa atividade, definimos várias funções. Os códigos estão listados a seguir. A primeira função apresentada *sentence_similarity* calcula a similaridade semân-

tica entre duas categorias. A função recebe as duas categorias, *sentence1* e *sentence2*, e um parâmetro, *first_synset* que indica se serão calculadas as similaridades entre todos os sinônimos de cada palavra, ou se serão usados apenas os sinônimos mais comuns de cada palavra, os primeiros da lista de *synsets*. Esses conceitos estão descritos mais profundamente na Seção 2.3.5. A função retorna os seis valores da similaridade semântica entre as duas sentenças calculados pelos seis métodos diferentes.

As funções *pos_tag* e *word_tokenizer* são funções definidas pelo pacote *NLTK* [48] do Python. A função *word_tokenizer* divide uma string em substrings, dessa forma uma categoria é dividida em palavras que compõem a categoria. A função *pos_tag* é um etiquetador de partes do discurso (substantivos, verbos, adjetivos, advérbio, etc.), ou *POS*, que processa uma string e associa um tag para a *POS* correspondente a palavra. Por convenção no *NLTK*, um token marcado é representado usando uma tupla que consiste no token e na tag.

```
1  from nltk import word_tokenize, pos_tag
2
3  #calcula a similaridade entre duas sentencas, utilizando os seis metodos
4  def sentence_similarity(sentence1, sentence2, first_synset):
5
6      """ compute the sentence similarity using Wordnet """
7
8      # Tokenize and tag
9      sentence1 = pos_tag(word_tokenize(sentence1))
10     sentence2 = pos_tag(word_tokenize(sentence2))
11
12     # Get the synsets for the tagged words
13     synsets1 = [tagged_to_synset(*tagged_word, first_synset) for tagged_word in sentence1
           ]
14     synsets2 = [tagged_to_synset(*tagged_word, first_synset) for tagged_word in sentence2
           ]
15
16     # Filter out the Nones
17     all_synsets1 = [ss for ss in synsets1 if ss]
18     all_synsets2 = [ss for ss in synsets2 if ss]
19
20     score1_path = 0.0
21     score1_wup = 0.0
22     score1_lch = 0.0
23     score1_res = 0.0
24     score1_jcn = 0.0
25     score1_lin = 0.0
26     count1 = 0
27
28     it_all_synsets1 = iter(all_synsets1)
29
30     # For each synsets of each word in the first sentence
31     for synsets1 in it_all_synsets1:
32
```

```
33          best_score_path = 0.0
34          best_score_wup = 0.0
35          best_score_lch = 0.0
36          best_score_res = 0.0
37          best_score_jcn = 0.0
38          best_score_lin = 0.0
39
40          it_all_synsets2 = iter(all_synsets2)
41
42          # For each synsets of each word in the second sentence
43          # Get the similarity value of the most similar synset in the other sentence
44          for synsets2 in it_all_synsets2:
45
46              best_score_path, best_score_wup, best_score_lch, \
47              best_score_res, best_score_jcn, best_score_lin = synsets_similarity(synsets1,
                    synsets2)
48
49              score1_path += best_score_path
50              score1_wup += best_score_wup
51              score1_lch += best_score_lch
52              score1_res += best_score_res
53              score1_jcn += best_score_jcn
54              score1_lin += best_score_lin
55
56              count1 += 1
57
58      # Average the values
59      if(score1_path != 0):
60          score1_path /= count1
61
62      if(score1_wup != 0):
63          score1_wup /= count1
64
65      if(score1_lch != 0):
66          score1_lch /= count1
67
68      if(score1_res != 0):
69          score1_res /= count1
70
71      if(score1_jcn != 0):
72          score1_jcn /= count1
73
74      if(score1_lin != 0):
75          score1_lin /= count1
76
77      score2_path = 0.0
78      score2_wup = 0.0
79      score2_lch = 0.0
80      score2_res = 0.0
81      score2_jcn = 0.0
82      score2_lin = 0.0
83      count2 = 0
84
85      it_all_synsets2 = iter(all_synsets2)
```

```
86
87      # For each word in the second sentence
88      for synsets2 in it_all_synsets2:
89
90          best_score_path = 0.0
91          best_score_wup = 0.0
92          best_score_lch = 0.0
93          best_score_res = 0.0
94          best_score_jcn = 0.0
95          best_score_lin = 0.0
96
97          it_all_synsets1 = iter(all_synsets1)
98
99          # For each synsets of each word in the first sentence
100         # Get the similarity value of the most similar synset in the other sentence
101         for synsets1 in it_all_synsets1:
102
103             best_score_path, best_score_wup, best_score_lch, \
104             best_score_res, best_score_jcn, best_score_lin = synsets_similarity(synsets1,
                    synsets2)
105
106             score2_path += best_score_path
107             score2_wup += best_score_wup
108             score2_lch += best_score_lch
109             score2_res += best_score_res
110             score2_jcn += best_score_jcn
111             score2_lin += best_score_lin
112
113             count2 += 1
114
115     # Average the values
116     if(score2_path != 0 ):
117         score2_path /= count2
118
119     if(score2_wup != 0 ):
120         score2_wup /= count2
121
122     if(score2_lch != 0 ):
123         score2_lch /= count2
124
125     if(score2_res != 0 ):
126         score2_res /= count2
127
128     if(score2_jcn != 0 ):
129         score2_jcn /= count2
130
131     if(score2_lin != 0 ):
132         score2_lin /= count2
133
134     score_path = (score1_path + score2_path) / 2
135     score_wup = (score1_wup + score2_wup) / 2
136     score_lch = (score1_lch + score2_lch) / 2
137     score_res = (score1_res + score2_res) / 2
138     score_jcn = (score1_jcn + score2_jcn) / 2
```

```
    score_lin = (score1_lin + score2_lin) / 2

    return score_path, score_wup, score_lch, score_res, score_jcn, score_lin
```

Neste trabalho, utilizamos as implementações do *WordNet* [36] para os métodos de cálculo de similaridade semântica descritos. Dessa forma, precisamos definir uma função de mapeamento para as tags retornados pela função *pos_tag* do *NLTK* e as tags reconhecidas pelo *WordNet*. A função $penn_to_wn$ define o mapeamento utilizado.

```
def penn_to_wn(tag):

    """ Convert between a Penn Treebank tag to a simplified Wordnet tag """

    if tag.startswith('N'):
        return 'n'

    if tag.startswith('V'):
        return 'v'

    if tag.startswith('J'):
        return 'a'

    if tag.startswith('R'):
        return 'r'

    if tag.startswith('D'):
        return 'd'

    return None
```

Dessa forma, podemos obter os *synsets* do *WordNet* para cada palavra que formam as categorias. A função *tagged_to_synset* retorna os *synsets* para cada palavra de acordo com a tag passada com parâmetro. Caso o parâmetro *first_synset* seja verdadeiro, a função retorna apenas o primeiro *synset* da lista de *synsets* associado a palavra. O primeiro *synset* da lista de *synsets* é o sinônimo mais comumente utilizado, de acordo com a ontologia definida no *WordNet*.

```
from nltk.corpus import wordnet as wn

def tagged_to_synset(word, tag, first_synset):

    wn_tag = penn_to_wn(tag)

    if wn_tag is None:
        return None

```

```
10     try:
11         if(first_synset is True):
12             synsets = []
13             synsets.append((wn.synsets(word, wn_tag)[0]))
14             return synsets
15         else:
16             return wn.synsets(word, wn_tag)
17     except:
18         return None
```

A seguir, definimos a função, *synsets_similarity* que realiza o cálculo da similaridade entre dois conjuntos de *synsets*, que contém todos os sinônimos de uma palavra (ou apenas o sinônimo mais comum, caso o parâmetro *first_synset* seja verdadeiro), por meio dos seis métodos diferentes, e retorna os seis resultados, calculados pelos seis métodos diferentes.

```
1 def synsets_similarity(synsets1, synsets2):
2
3     ### calcula a maior similaridade entre os dois synsets
4     best_score_path = 0.0
5     best_score_wup = 0.0
6     best_score_lch = 0.0
7     best_score_res = 0.0
8     best_score_jcn = 0.0
9     best_score_lin = 0.0
10
11    it_synsets1 = iter(synsets1)
12
13    # For each synset in the first synsets
14    for synset1 in it_synsets1:
15
16        it_synsets2 = iter(synsets2)
17
18        for synset2 in it_synsets2:
19
20            sim_path, sim_wup, sim_lch, \
21            sim_res, sim_jcn, sim_lin = synset_similarity(synset1, synset2)
22
23            if sim_path is None:
24                sim_path = 0.0
25
26            if sim_wup is None:
27                sim_wup = 0.0
28
29            if sim_lch is None:
30                sim_lch = 0.0
31
32            if sim_res is None:
33                sim_res = 0.0
34
35            if sim_jcn is None:
```

```
36                  sim_jcn = 0.0
37
38              if sim_lin is None:
39                  sim_lin = 0.0
40
41              if sim_path > best_score_path:
42                  best_score_path = sim_path
43
44              if sim_wup > best_score_wup:
45                  best_score_wup = sim_wup
46
47              if sim_lch > best_score_lch:
48                  best_score_lch = sim_lch
49
50              if sim_res > best_score_res:
51                  best_score_res = sim_res
52
53              if sim_jcn > best_score_jcn:
54                  best_score_jcn = sim_jcn
55
56              if sim_lin > best_score_lin:
57                  best_score_lin = sim_lin
58
59      return best_score_path, best_score_wup, best_score_lch, best_score_res,
            best_score_jcn, best_score_lin
```

Já a função *synset_similarity* calcula a similaridade entre dois *synsets*, ou dois sinônimos, de uma palavra, por meio dos seis métodos diferentes propostos nesse trabalho.

```
1  from nltk.corpus import wordnet_ic
2  from nltk.corpus.reader.wordnet import WordNetError
3  import logging
4  logging.basicConfig(filename='logging.log',level=logging.DEBUG)
5
6  def synset_similarity(synset1, synset2):
7
8      brown_ic = wordnet_ic.ic('ic-brown.dat')
9
10     sim_path = synset1.path_similarity(synset2)
11
12     sim_wup = synset1.wup_similarity(synset2)
13
14     try:
15         sim_lch = synset1.lch_similarity(synset2)
16     except WordNetError as err:
17         sim_lch = 0.0
18         logging.warning(err)
19
20     try:
21         sim_res = synset1.res_similarity(synset2, brown_ic)
22     except WordNetError as err:
23         sim_res = 0.0
```

```
24          logging.warning(err)
25
26      try:
27          sim_jcn = synset1.jcn_similarity(synset2, brown_ic)
28      except WordNetError as err:
29          sim_jcn = 0.0
30          logging.warning(err)
31
32      try:
33          sim_lin = synset1.lin_similarity(synset2, brown_ic)
34      except WordNetError as err:
35          sim_lin = 0.0
36          logging.warning(err)
37
38
39      return sim_path, sim_wup, sim_lch, sim_res, sim_jcn, sim_lin
```

Definidas funções para o cálculo da similaridade semântica entre duas sentenças, em nosso caso, entre duas categorias, podemos realizar o alinhamento das categorias, calculando para todas as categorias do Subconjunto Abrangente de Categorias, obtidas no Processo 1 A.1, a similaridade semântica, de acordo com seis métodos diferentes, e elegendo a categoria que tiver mais resultados iguais, entre o seis métodos.

A função get_lst_similarities recebe uma categoria e a lista de Categorias Abrangentes, e obtém a categoria mais similar entre a lista de Categorias Abrangentes, para cada método utilizado. O retorno da função é uma lista contendo o valor da maior similaridade e a categoria associada a essa maior similaridade, para todos os métodos utilizados.

```
1  def get_lst_similarities(category, lstCategoriesToMatch):
2
3      best_sim_path = 0.0
4      best_sim_wup = 0.0
5      best_sim_lch = 0.0
6      best_sim_res = 0.0
7      best_sim_jcn = 0.0
8      best_sim_lin = 0.0
9
10     best_category_path = ""
11     best_category_wup = ""
12     best_category_lch = ""
13     best_category_res = ""
14     best_category_jcn = ""
15     best_category_lin = ""
16
17     it_categories_to_match = iter(lstCategoriesToMatch)
18     for category_coverage in it_categories_to_match:
19
20         sim_path, sim_wup, sim_lch, \
```

```
21          sim_res, sim_jcn, sim_lin = sentence_similarity(category, category_coverage,
                first_synset)
22
23          if sim_path > best_sim_path:
24              best_sim_path = sim_path
25              best_category_path = category_coverage
26
27          if sim_wup > best_sim_wup:
28              best_sim_wup = sim_wup
29              best_category_wup = category_coverage
30
31          if sim_lch > best_sim_lch:
32              best_sim_lch = sim_lch
33              best_category_lch = category_coverage
34
35          if sim_res > best_sim_res:
36              best_sim_res = sim_res
37              best_category_res = category_coverage
38
39          if sim_jcn > best_sim_jcn:
40              best_sim_jcn = sim_jcn
41              best_category_jcn = category_coverage
42
43          if sim_lin > best_sim_lin:
44              best_sim_lin = sim_lin
45              best_category_lin = category_coverage
46
47      lstSimilarities = []
48      lstSimilarities.append("path"), lstSimilarities.append(best_sim_path),
            lstSimilarities.append(best_category_path)
49      lstSimilarities.append("wup"), lstSimilarities.append(best_sim_wup), lstSimilarities.
            append(best_category_wup)
50      lstSimilarities.append("lch"), lstSimilarities.append(best_sim_lch), lstSimilarities.
            append(best_category_lch)
51      lstSimilarities.append("res"), lstSimilarities.append(best_sim_res), lstSimilarities.
            append(best_category_res)
52      lstSimilarities.append("jcn"), lstSimilarities.append(best_sim_jcn), lstSimilarities.
            append(best_category_jcn)
53      lstSimilarities.append("lin"), lstSimilarities.append(best_sim_lin), lstSimilarities.
            append(best_category_lin)
54
55      return lstSimilarities
```

A função *get_most_elected_category* retorna, para uma lista de similaridades, a categoria da lista Categorias Abrangentes que obteve mais resultados iguais entre os seis métodos utilizados nos cálculos de similaridade.

```
1 from collections import OrderedDict
2 import operator
3 import random
4
```

```
5  def get_most_elected_category(lst_similarities):
6
7      dictCategoryFreq = {}
8
9      best_category_path = lst_similarities[2]
10     best_category_wup = lst_similarities[5]
11     best_category_lch = lst_similarities[8]
12     best_category_res = lst_similarities[11]
13     best_category_jcn = lst_similarities[14]
14     best_category_lin = lst_similarities[17]
15
16     freq = dictCategoryFreq.get(best_category_path)
17     if(freq is None):
18         freq = 0
19     freq += 1
20     dictCategoryFreq.update( {best_category_path : freq} )
21
22     freq = dictCategoryFreq.get(best_category_wup)
23     if(freq is None):
24         freq = 0
25     freq += 1
26     dictCategoryFreq.update( {best_category_wup : freq} )
27
28     freq = dictCategoryFreq.get(best_category_lch)
29     if(freq is None):
30         freq = 0
31     freq += 1
32     dictCategoryFreq.update( {best_category_lch : freq} )
33
34     freq = dictCategoryFreq.get(best_category_res)
35     if(freq is None):
36         freq = 0
37     freq += 1
38     dictCategoryFreq.update( {best_category_res : freq} )
39
40     freq = dictCategoryFreq.get(best_category_jcn)
41     if(freq is None):
42         freq = 0
43     freq += 1
44     dictCategoryFreq.update( {best_category_jcn : freq} )
45
46     freq = dictCategoryFreq.get(best_category_lin)
47     if(freq is None):
48         freq = 0
49     freq += 1
50     dictCategoryFreq.update( {best_category_lin : freq} )
51
52     dictFreqOrd = OrderedDict(sorted(dictCategoryFreq.items(),
53                                      key = operator.itemgetter(1),
54                                      reverse = True))
55
56     maxFreq = max(dictFreqOrd.values())
57
58     best_categories = []
```

```
59     for best_category, freq in dictFreqOrd.items():
60         if (freq == maxFreq):
61             best_categories.append(best_category)
62
63     if len(best_categories) > 1:
64         index = random.randint(0, len(best_categories) - 1)
65         return best_categories[index]
66
67     return best_categories[0]
```

Podemos então definir a função que retorna um dicionário contendo os nomes dos portais, suas categorias e o alinhamento produzido para cada categoria. A função *fillDictPortalsCategoryMatch* também retorna um dicionário contendo os nomes do portais, suas categorias e a lista de similaridades calculadas para cada categoria pelos seis métodos utilizados. Ela faz uso de uma outra função, *fillDictCategoriesMatch* que faz a chamada das funções necessárias para o cálculo de similaridade entre todas as categorias de um portal e as Categorias Abrangentes. Assim, definimos todas as funções necessárias para o alinhamento das categorias de um portal com todas as categorias do Subconjunto Abrangente.

```
1 def fillDictPortalsCategoryMatch(lstPortal, lstCategoriesCoverage):
2
3     dictPortalsCategoryMatch = {}
4     dictPortalsCategorySimilarities = {}
5
6     it_portals = iter(lstPortal)
7     for portal in it_portals:
8
9         categories = portal.getCategories()
10        dictCategorySimilarities, dictCategoryMatch = fillDictCategoriesMatch(categories,
              lstCategoriesCoverage)
11
12        dictPortalsCategorySimilarities.update( {portal.getCity() :
              dictCategorySimilarities} )
13        dictPortalsCategoryMatch.update( { portal.getCity() : dictCategoryMatch} )
14
15    return dictPortalsCategoryMatch, dictPortalsCategorySimilarities
```

```
1 def fillDictCategoriesMatch(categories, lstCategoriesToMatch):
2
3     dictCategorySimilarities = {}
4     dictCategoryMatch = {}
5
6     it_categories = iter(categories)
7     for category in it_categories:
8
```

```
9         lst_similarities = get_lst_similarities(category, lstCategoriesToMatch)
10        dictCategorySimilarities.update( {category : lst_similarities} )
11
12        best_category = get_most_elected_category(lst_similarities)
13        dictCategoryMatch.update( {category : best_category} )
14
15    return dictCategorySimilarities, dictCategoryMatch
```

A.2.5 Atividade 5: Escrever resultado do alinhamento

Entrada: Dicionário contendo o nome dos portais, suas categorias e o alinhamento produzido para cada categoria do portal.
Saída: Arquivo *JSON* contendo as categorias de todos os portais e alinhamento produzido para cada uma.

Essa função escreve o conteúdo dos dicionários produzidos na atividade anterior, Atividade 10 A.2.4, em dois arquivos *JSON*, um contendo o alinhamento produzido para cada categoria de cada portal, o outro contendo os valores de similaridade produzidos no alinhamento.

```
1 def write_categories_match(dictPortalsCategoryMatch, dictPortalsCategorySimilarities):
2
3     file = open('portals_category_similarities.json', 'w')
4     s = json.dumps(list(dictPortalsCategorySimilarities.items()),
5                    indent=4, ensure_ascii=False).encode('utf8').decode('latin1')
6     file.writelines(s)
7     file.close()
8
9     file = open('portals_category_match.json', 'w')
10    s = json.dumps(list(dictPortalsCategoryMatch.items()),
11                   indent=4, ensure_ascii=False).encode('utf8').decode('latin1')
12    file.writelines(s)
13    file.close()
```

A.2.6 Atividade 6: Avaliação do Alinhamento

Entrada: Arquivo JSON contendo as categorias de todos os portais e alinhamento produzido para cada uma.
Saída: Arquivo JSON editado pelo usuário contendo as categorias de todos os portais e

alinhamento produzido para cada uma após avaliação do usuário.

Nessa etapa, o usuário pode avaliar o alinhamento produzido pelo algoritmo, verificar sua consistência e corrigir algum alinhamento que considerar mal feito. Algumas categorias podem não ser alinhadas com nenhuma categoria abrangente, assim o usuário deve manualmente alinhar para uma categoria abrangente.

No código abaixo apresentamos as chamadas das funções necessárias para o processo de Alinhamento.

```
1 print("\n")
2 first_synset = True
3
4 lstPortal = readPortalsFromJsonFile()
5 print("Number of Portals: " + "{0}".format(len(lstPortal)))
6
7 lstCategoriesCoverage = readCategoriesFromJsonFile()
8 print("Number of Best Coverage Categories: " + "{0}".format(len(lstCategoriesCoverage)))
9 print("\n")
10
11 start = time.time()
12
13 dictPortalsCategoryMatch, dictPortalsCategorySimilarities = fillDictPortalsCategoryMatch(
    lstPortal, lstCategoriesCoverage)
14 write_categories_match(dictPortalsCategoryMatch, dictPortalsCategorySimilarities)
15
16 end = time.time()
17
18 print("\n")
19 time = (end - start)
20 print('duracao: {:.2f}'.format(time) + " s")
```

APÊNDICE B – ABRANGÊNCIA DAS PALAVRAS NOS PORTAIS UTILIZADOS NO EXEMPLO DO GUIA

Neste Apêndice é mostrada uma tabela com todos os conjuntos formados com a lista de palavras que ocorrem nos portais utilizados no exemplo do Guia, descrito no Apêndice A, em ordem de frequência e de entrada.

1	education	(NYC, Chicago)	67%
2	education, environment	(NYC, Chicago)	67%
3	education, environment, health	(NYC, Chicago)	67%
4	education, environment, health, safety	(NYC, Chicago)	67%
5	education, environment, health, safety, recreation	(NYC, Chicago)	67%
6	education, environment, health, safety, recreation, services	(NYC, Chicago)	67%
7	education, environment, health, safety, recreation, services, transportation	(NYC, Chicago)	67%
8	education, environment, health, safety, recreation, services, transportation, business	(NYC, Chicago)	67%
9	education, environment, health, safety, recreation, services, transportation, business, government	(NYC, Chicago)	67%
10	education, environment, health, safety, recreation, services, transportation, business, government, housing	(NYC, Chicago)	67%
11	education, environment, health, safety, recreation, services, transportation, business, government, housing, development,	(NYC, Chicago)	67%
12	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social	(NYC, Chicago)	67%
13	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable	(NYC, Los Angeles, Chicago)	100%
14	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable	(NYC, Los Angeles, Chicago)	100%
15	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous	(NYC, Los Angeles, Chicago)	100%
16	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe	(NYC, Los Angeles, Chicago)	100%
17	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well	(NYC, Los Angeles, Chicago)	100%
18	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run	(NYC, Los Angeles, Chicago)	100%
19	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration	(NYC, Los Angeles, Chicago)	100%

20	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance	(NYC, Los Angeles, Chicago)	100%
21	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings	(NYC, Los Angeles, Chicago)	100%
22	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community	(NYC, Los Angeles, Chicago)	100%
23	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics	(NYC, Los Angeles, Chicago)	100%
24	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics, events	(NYC, Los Angeles, Chicago)	100%
25	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics, events, facilities	(NYC, Los Angeles, Chicago)	100%
26	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics, events, facilities, boundaries	(NYC, Los Angeles, Chicago)	100%
27	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics, events, facilities, boundaries, human	(NYC, Los Angeles, Chicago)	100%
28	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics, events, facilities, boundaries, human, historic	(NYC, Los Angeles, Chicago)	100%

29	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics, events, facilities, boundaries, human, historic, preservation	(NYC, Los Angeles, Chicago)	100%
30	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics, events, facilities, boundaries, human, historic, preservation, parks	(NYC, Los Angeles, Chicago)	100%
31	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics, events, facilities, boundaries, human, historic, preservation, parks, sanitation	(NYC, Los Angeles, Chicago)	100%
32	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics, events, facilities, boundaries, human, historic, preservation, parks, sanitation, service	(NYC, Los Angeles, Chicago)	100%
33	education, environment, health, safety, recreation, services, transportation, business, government, housing, development, social, livable, sustainable, prosperous, safe, well, run, administration, finance, buildings, community, ethics, events, facilities, boundaries, human, historic, preservation, parks, sanitation, service, requests	(NYC, Los Angeles, Chicago)	100%

APÊNDICE C – DADOS DOS PORTAIS COLETADOS NA PESQUISA EXPLORATÓRIA

Neste Apêndice são apresentados os dados dos portais coletados na Pesquisa Exploratória realizada nas 100 cidades americanas mais populosas. Os dados foram compilados em um arquivo *JSON*.


```
[
        {
        "city": "NYC",
        "url": "https://opendata.cityofnewyork.us/data/",
        "coord": "40.6643 N 73.9385 W",
        "categorization": "Categories",
        "platform": "Socrata",
        "categories": [
            "Business",
            "City Government",
            "Education",
            "Environment",
            "Health",
            "Housing & Development",
            "Public Safety",
            "Recreation",
            "Social Services",
            "Transportation"
        ]
        },
        {
        "city": "Los Angeles",
        "url": "https://data.lacity.org/browse",
        "coord": "34.0194 N 118.4108 W",
        "categorization": "Categories",
        "platform": "Socrata",
        "categories": [
            "A Livable and Sustainable City",
            "A Prosperous City",
            "A Safe City",
            "A Well Run City"
        ]
```

```
33      },
34      {
35          "city": "Chicago ",
36          "url": "https://data.cityofchicago.org/browse",
37          "coord": "41.8376 N 87.6818 W",
38          "categorization": "Categories",
39          "platform": "Socrata",
40          "categories": [
41              "Administration & Finance",
42              "Buildings",
43              "Community",
44              "Education",
45              "Environment",
46              "Ethics",
47              "Events",
48              "FOIA",
49              "Facilities & Geo. Boundaries ",
50              "Health & Human Services",
51              "Historic Preservation",
52              "Parks & Recreation",
53              "Public Safety",
54              "Sanitation",
55              "Service Requests",
56              "Transportation"
57          ]
58      },
59      {
60          "city": "Houston",
61          "url": "http://data.ohouston.org/dataset",
62          "coord": "29.7805 N 95.3863 W",
63          "categorization": "Groups",
64          "platform": "CKAN",
65          "categories": [
66              "GIS",
67              "City of Houston Entreprise GIS",
68              "City of Houston Planning and Development Department",
69              "Government Boundaries",
70              "Public Works & Engineering",
71              "Planning & Development",
72              "Transportation",
73              "Permitting and Licensing",
74              "City of Houston Public Works and Engineering",
75              "Public Health & Safety",
76              "City of Houston Administration and Regulatory Affairs",
77              "Neighborhood Services",
78              "Finance",
79              "Environmental",
80              "Property",
81              "Hydrology",
82              "City of Houston Finance Department",
83              "Flood Hazard",
84              "Adresses Roads Cadastral",
85              "Houston Fire Department",
86              "Parking",
```

```
87              "Houston Police Department",
88              "City of Houston Health & Human Services Department",
89              "City of Houston Department of Neighborhood",
90              "City of Houston Solid Waste Mangement",
91              "Restaurants",
92              "Education and Schooling",
93              "Econonomic Development",
94              "City of Houston Parks and Recreation Department",
95              "City of Houston General Services Department",
96              "City of Houston City Secretary",
97              "Houston-Galveston Area Council",
98              "Demographics",
99              "City of Houston Office of Bussiness Opportunity",
100             "City of Houston Human Resources",
101             "City of Houston House & Community Development",
102             "City of Houston City Council",
103             "Houston Public Library",
104             "City of Houston Legal Department",
105             "City of Houston Information Technology Services",
106             "City of Houston Aviation / Houston Airport System"
107         ]
108     },
109     {
110         "city": "Phoenix",
111         "url": "https://www.phoenix.gov/opendata",
112         "coord": "33.5722 N 112.0880 W",
113         "categorization": "Categories",
114         "platform": "-",
115         "categories": [
116             "Census Data",
117             "CheckBook & Sales Tax",
118             "Energy & Sustainability",
119             "Neighborhood & Safety",
120             "Parks, Art & Culture",
121             "Property & Development",
122             "Staff Salaries",
123             "Transportation"
124         ]
125     },
126         {
127         "city": "Philadelphia",
128         "url": "https://www.opendataphilly.org/dataset",
129         "coord": "40.0094 N 75.1333 W",
130         "categorization": "Topics",
131         "platform": "CKAN",
132         "categories": [
133             "Transportation",
134             "Real State / Land Records",
135             "Environment",
136             "Health / Human Services",
137             "Planning / Zoning",
138             "Elections / Politics",
139             "Arts / Culture / History",
140             "Education",
```

```
141             "Public Safety",
142             "Parks / Recreation",
143             "Economy",
144             "Uncategorized",
145             "Food",
146             "Budget / Finance"
147         ]
148     },
149     {
150         "city": "San Antonio",
151         "url": "https://www.sanantonio.gov/SAPD/SAPD-Open-Data-Initiative",
152         "coord": "29.4724 N 98.5251 W",
153         "categorization": "-",
154         "platform": "-",
155         "categories": []
156     },
157     {
158         "city": "San Diego",
159         "url": "https://data.sandiego.gov/datasets/",
160         "coord": "32.8153 N 117.1350 W",
161         "categorization": "Categories",
162         "platform": "-",
163         "categories": [
164             "Economy & Community",
165             "City Infrastructure",
166             "City Management",
167             "Transportation",
168             "Public Safety",
169             "Energy & Environment",
170             "Culture & Recreation"
171         ]
172     },
173     {
174         "city": "Dallas",
175         "url": "https://www.dallasopendata.com/",
176         "coord": "32.7757 N 96.7967 W",
177         "categorization": "Categories",
178         "platform": "Socrata",
179         "categories": [
180             "Budget & Finance",
181             "City Infrastructure",
182             "City Services",
183             "Economic Development",
184             "Geography & Boundaries",
185             "Government",
186             "Public Safety"
187         ]
188     },
189     {
190         "city": "San Jose",
191         "url": "http://data.sanjoseca.gov/home",
192         "coord": "37.2969 N 121.8193 W",
193         "categorization": "Top Categories",
194         "platform": "Junar",
```

```
195         "categories": [
196             "Aiport",
197             "Auditor",
198             "City's Manager Office",
199             "City Wide",
200             "Department Of Transportation",
201             "Economic Development",
202             "Environmental Services",
203             "Finance",
204             "Fire",
205             "Housing",
206             "Human Resources",
207             "Independent Police Auditor",
208             "Information Technology",
209             "Library",
210             "Parks, Recreation & Neighborhood Services",
211             "Planning Building and Code Enforcement",
212             "Police",
213             "Public Works",
214             "Retirement"
215         ]
216     },
217     {
218         "city": "Austin",
219         "url": "https://data.austintexas.gov/",
220         "coord": "30.3072 N 97.7560 W",
221         "categorization": "Categories",
222         "platform": "Socrata",
223         "categories": [
224             "Bussiness",
225             "Capital Metro",
226             "Capital Planning",
227             "Education",
228             "Environmental",
229             "Financial",
230             "Fun",
231             "Geo Data",
232             "Government",
233             "Health",
234             "Neighborhood",
235             "Permitting",
236             "Public Safety",
237             "Utility",
238             "Workforce Development"
239         ]
240     },
241     {
242         "city": "Jacksonville",
243         "url": "https://data.jacksonms.gov/browse",
244         "coord": "30.3370 N 81.6613 W",
245         "categorization": "Categories",
246         "platform": "Socrata",
247         "categories": [
248             "Budget and Finance",
```

```
249             "City Services",
250             "Community Development",
251             "Economic Develpment",
252             "Government Accountability",
253             "Public Safety",
254             "Public Works ",
255             "Recreation and Parks",
256             "Schools and Education",
257             "Transportation and Transit"
258         ]
259     },
260     {
261         "city": "San Francisco",
262         "url": "https://data.sfgov.org/browse",
263         "coord": "37.7751 N 122.4193 W",
264         "categorization": "Categories",
265         "platform": "Socrata",
266         "categories": [
267             "City Infrastructure",
268             "City Management and Ethics",
269             "Culture and Recreation",
270             "Economy and Community",
271             "Energy and Environment",
272             "Geographic Locations and Boundaries",
273             "Health and Social Services",
274             "Housing and Buildings",
275             "Public Safety",
276             "Transportation"
277         ]
278     },
279     {
280         "city": "Columbus",
281         "url": "http://data-columbus.opendata.arcgis.com",
282         "coord": "39.9848 N 82.9850 W",
283         "categorization": "Categories",
284         "platform": "ArcGis",
285         "categories": [
286             "Business",
287             "Boundaries",
288             "Health",
289             "Infrastructure",
290             "Planning",
291             "Recreation & Parks",
292             "Safety",
293             "Schools",
294             "Transportation",
295             "All"
296         ]
297     },
298     {
299         "city": "Indianapolis",
300         "url": "http://data.indy.gov/",
301         "coord": "39.7767 N 86.1459 W",
302         "categorization": "Categories",
```

```
303         "platform": "-",
304         "categories": [
305             "Boundaries",
306             "Transportation",
307             "Recreation",
308             "Property",
309             "Disclose Indy",
310             "Political",
311             "Survey",
312             "Planning & Zonning"
313         ]
314     },
315     {
316         "city": "Fort Worth",
317         "url": "https://data.fortworthtexas.gov/",
318         "coord": "32.7795 N 97.3463 W",
319         "categorization": "Categories",
320         "platform": "Socrata",
321         "categories": [
322             "Business",
323             "City Government",
324             "Environment & Health",
325             "Financial",
326             "Property Data",
327             "Public Safety",
328             "Services & Recreation",
329             "Technology & Communications",
330             "Tranpostation"
331         ]
332     },
333     {
334         "city": "Charlotte",
335         "url": "http://clt-charlotte.opendata.arcgis.com/",
336         "coord": "35.2087 N 80.8307 W",
337         "categorization": "Categories",
338         "platform": "ArcGis",
339         "categories": [
340             "311 / Services",
341             "Arts & Education",
342             "Business & Budget",
343             "Community Safety",
344             "City Government",
345             "Neighborhoods & Housing",
346             "Transportation",
347             "Environment",
348             "Demographics",
349             "Planning & Zoning",
350             "Map Features",
351             "Historical Data"
352         ]
353     },
354         {
355         "city": "Seattle",
356         "url": "https://data.seattle.gov/browse",
```

```
357         "coord": "47.6205 N 122.3509 W",
358         "categorization": "Categories",
359         "platform": "Socrata",
360         "categories": [
361             "City Business",
362             "Community",
363             "Education",
364             "Finance",
365             "Land Base",
366             "Parks",
367             "Parks and Recreation",
368             "Permitting",
369             "Public Safety",
370             "Transportation"
371         ]
372     },
373     {
374         "city": "Denver",
375         "url": "https://www.denvergov.org/opendata/",
376         "coord": "39.7618 N 104.8806 W",
377         "categorization": "-",
378         "platform": "-",
379         "categories": []
380     },
381     {
382         "city": "Washington",
383         "url": "https://data.wa.gov/browse",
384         "coord": "38.9041 N 77.0171 W",
385         "categorization": "Categories",
386         "platform": "Socrata",
387         "categories": [
388             "Agriculture",
389             "Consumer Protection",
390             "Demographics",
391             "Economics",
392             "Education",
393             "Employment",
394             "Health",
395             "Labor",
396             "Natural Resources & Environment",
397             "Politics",
398             "Procurements and Contracts",
399             "Public Safety",
400             "Recreation",
401             "Transportation"
402         ]
403     },
404         {
405         "city": "Boston",
406         "url": "https://data.cityofboston.gov/",
407         "coord": "42.3320 N 71.0202 W",
408         "categorization": "Categories",
409         "platform": "Socrata",
410         "categories": [
```

```
411             "City Services",
412             "Facilites",
413             "Finance",
414             "Health",
415             "Permitting",
416             "Public Safety",
417             "Transportation"
418         ]
419     },
420     {
421         "city": "Detroit",
422         "url": "https://data.detroitmi.gov/browse",
423         "coord": "42.3830 N 83.1022 W",
424         "categorization": "Categories",
425         "platform": "Socrata",
426         "categories": [
427             "Business",
428             "Children & Families",
429             "Education",
430             "Fun",
431             "Government",
432             "Personal",
433             "Property & Parcels",
434             "Public Health",
435             "Public Safety",
436             "Transportation"
437         ]
438     },
439     {
440         "city": "Nashville",
441         "url": "https://data.nashville.gov/browse",
442         "coord": "36.1718 N 86.7850 W",
443         "categorization": "Categories",
444         "platform": "Socrata",
445         "categories": [
446             "Agriculture",
447             "Art",
448             "Beautification",
449             "Budget / Finance",
450             "Business, Development & Housing",
451             "Culture",
452             "Education",
453             "Elections",
454             "Emergency Management",
455             "Energy Usage",
456             "Environment",
457             "Fire ",
458             "Geneology",
459             "General Government",
460             "Health",
461             "History",
462             "Libraries",
463             "Licenses & Permits",
464             "Medical",
```

```
465             "Metro Government",
466             "Parks",
467             "Police",
468             "Public Safety",
469             "Public Services",
470             "Recycling/Conservation",
471             "Social Services",
472             "Transportation"
473         ]
474     },
475     {
476         "city": "Portland",
477         "url": "https://www.portlandoregon.gov/police/71673?",
478         "coord": "45.5370 N 122.6500 W",
479         "categorization": "-",
480         "platform": "-",
481         "categories": []
482     },
483     {
484         "city": "Oklahoma City",
485         "url": "https://data.okc.gov/portal/page/start",
486         "coord": "35.4671 N 97.5137 W",
487         "categorization": "Categories",
488         "platform": "-",
489         "categories": [
490             "General",
491             "GO Bond Projects (2017)",
492             "Planimetrics",
493             "Subdivision",
494             "Utilities",
495             "Census",
496             "Impact Fees",
497             "Public Safety",
498             "Survey Control Points",
499             "Zoning",
500             "Go Bond Projects (2007)",
501             "Parks and Trails",
502             "School",
503             "Transportation"
504         ]
505     },
506     {
507         "city": "Las Vegas",
508         "url": "https://opendata.lasvegasnevada.gov/browse",
509         "coord": "36.2277 N 115.2640 W",
510         "categorization": "Categories",
511         "platform": "Socrata",
512         "categories": [
513             "Arts and Culture",
514             "Building and Safety",
515             "Community Risk Reduction",
516             "Economic Development",
517             "Finance",
518             "General Information",
```

```
519             "Growing Economy",
520             "High Performing Government",
521             "Neighborhood Livability",
522             "Parks and Recreation",
523             "Planning",
524             "Public Safety",
525             "Public Works",
526             "Schools"
527         ]
528     },
529     {
530         "city": "Louisville",
531         "url": "https://data.louisvilleky.gov/search/type/dataset",
532         "coord": "38.1781 N 85.6667 W",
533         "categorization": "Categories",
534         "platform": "DKAN",
535         "categories": []
536     },
537     {
538         "city": "Baltimore",
539         "url": "https://data.baltimorecity.gov/browse",
540         "coord": "39.3002 N 76.6105 W",
541         "categorization": "Categories",
542         "platform": "Socrata",
543         "categories": [
544             "City Government",
545             "City Services",
546             "Crime",
547             "Culture & Arts",
548             "Financial",
549             "Geographic",
550             "Health",
551             "Housing & Development",
552             "Neighborhoods",
553             "Public Safety",
554             "Public Works",
555             "Transportation"
556         ]
557     },
558     {
559         "city": "Tucson",
560         "url": "https://data.tucsonaz.gov/",
561         "coord": "32.1543 N 110.8711 W",
562         "categorization": "-",
563         "platform": "-",
564         "categories": []
565     },
566     {
567         "city": "Sacramento",
568         "url": "http://data.cityofsacramento.org/",
569         "coord": "38.5666 N 121.4686 W",
570         "categorization": "Categories",
571         "platform": "ArcGis",
572         "categories": [
```

```
573             "Animal Care",
574             "Budget & Finance",
575             "Disclosure & Ethics",
576             "Economy & Community",
577             "Locations & Mapping",
578             "Parks & Recreation",
579             "Permits & Planning",
580             "Public Safety",
581             "Service Requests",
582             "Transportation & Infrastructure"
583         ]
584     },
585     {
586         "city": "Mesa",
587         "url": "http://open.mesaaz.gov/home",
588         "coord": "33.4019 N 111.7174 W",
589         "categorization": "-",
590         "platform": "Junar",
591         "categories": [
592             "Energy & Utilities",
593             "Permits & Licences",
594             "Financials",
595             "Recreation & Culture",
596             "Zoning & Property",
597             "Planes, Trains & Automobiles",
598             "Neighborhoods",
599             "Public Safety"
600         ]
601     },
602     {
603         "city": "Kansas City",
604         "url": "https://data.kcmo.org",
605         "coord": "39.1252 N 94.5511 W",
606         "categorization": "Categories",
607         "platform": "Socrata",
608         "categories": [
609             "311",
610             "Airport",
611             "Annual Audit Plan & Reports",
612             "Area Plans",
613             "Auction",
614             "Audit Highlights",
615             "Audits",
616             "Audits and Memos",
617             "Brownfields",
618             "Budget",
619             "Business",
620             "Census",
621             "Climate Protection Steering Com",
622             "Code Interpretations",
623             "Construction",
624             "Crime",
625             "Development",
626             "Development Review",
```

```
627             "Emergency",
628             "Environmental Management Commission",
629             "FY 2009 - 2010",
630             "FY 2010 - 2011",
631             "FY 2011 - 2012",
632             "FY 2012 - 2013",
633             "FY 2013 - 2014",
634             "FY 2014 - 2015",
635             "FY 2015 - 2016",
636             "FY 2016 - 2017",
637             "Fees",
638             "Finance",
639             "Focus",
640             "Food",
641             "Forecasts",
642             "Forms",
643             "Forms and Applications",
644             "GIS",
645             "Government",
646             "Health",
647             "Historic Preservation",
648             "Housing",
649             "Human Relations",
650             "Human Resources",
651             "Information Bulletins",
652             "Innovation",
653             "Investor Relations",
654             "KC Bizcare",
655             "KC City Energy Project",
656             "KCI Terminal Advisory",
657             "Land Development",
658             "Land Use",
659             "Legislation",
660             "Legislative Info",
661             "Monthly Status Reports",
662             "Municipal Court",
663             "Neighborhoods",
664             "New Tax Forms",
665             "Old Tax Forms",
666             "Peer Review",
667             "Property",
668             "Regulated Industries",
669             "Regulatory Codes",
670             "Safety",
671             "Scope Statement",
672             "Streetcar",
673             "Sustainability",
674             "Taxes",
675             "Traffic",
676             "Traffic Sign Changes",
677             "Transportation",
678             "Workforce"
679         ]
680     },
```

```
681         {
682         "city": "Atlanta",
683         "url": "https://atlanta.demo.socrata.com",
684         "coord": "33.7629 N 84.4227 W",
685         "categorization": "-",
686         "platform": "Socrata",
687         "categories": []
688     },
689     {
690         "city": "Long Beach",
691         "url": "http://datalb.longbeach.gov/",
692         "coord": "33.8091 N 118.1553 W",
693         "categorization": "Categories",
694         "platform": "ArcGis",
695         "categories": [
696             "Business",
697             "Boundaries",
698             "Health",
699             "Infrastructure",
700             "Planning",
701             "Recreation & Parks",
702             "Safety",
703             "Schools",
704             "Transportation",
705             "All"
706         ]
707     },
708     {
709         "city": "Colorado Springs",
710         "url": "http://coloradodata-statesales.opendata.arcgis.com",
711         "coord": "38.8673 N 104.7607 W",
712         "categorization": "Categories",
713         "platform": "ArcGis",
714         "categories": [
715             "Economy and Community",
716             "City Management and Ethics",
717             "Transportation",
718             "Public Safety",
719             "Health and Social Services",
720             "Geographic Locations and Boundaries",
721             "Energy and Environment",
722             "Housing and Buildings",
723             "City Infrastructure",
724             "Culture and Recreation"
725         ]
726     },
727     {
728         "city": "Raleigh",
729         "url": "https://data.raleighnc.gov/browse",
730         "coord": "35.8302 N 78.6414 W",
731         "categorization": "Categories",
732         "platform": "Socrata",
733         "categories": [
734             "Alt. Fuels",
```

```
735             "Budget and Management Services",
736             "Business",
737             "Downtown",
738             "Fire",
739             "Government",
740             "Government Buldings and Structures",
741             "Greenways",
742             "Housing",
743             "Infrastructure",
744             "Neighborhoods",
745             "Parking",
746             "People",
747             "Permits",
748             "Police",
749             "Public Safety",
750             "Sustainability",
751             "Transit",
752             "Urban Planning",
753             "Wake County EMS"
754         ]
755     },
756     {
757         "city": "Miami (Miami Dade)",
758         "url": "https://opendata.miamidade.gov/browse",
759         "coord": "25.7752 N 80.2086 W",
760         "categorization": "Categories",
761         "platform": "Socrata",
762         "categories": [
763             "311",
764             "Animals",
765             "Building",
766             "Corrections",
767             "Electoral",
768             "Environment",
769             "Infrastructure",
770             "Procurement",
771             "Transportation",
772             "Waste"
773         ]
774     },
775     {
776         "city": "Virginia Beach",
777         "url": "https://data.vbgov.com/browse",
778         "coord": "36.7793 N 76.0240 W",
779         "categorization": "Categories",
780         "platform": "Socrata",
781         "categories": [
782             "Business",
783             "FOIA",
784             "Finance / Government",
785             "Neighborhoods / Environment",
786             "Public Safety"
787         ]
788     },
```

```
789     {
790         "city": "Omaha (Dougla County)",
791         "url": "http://data-dogis.opendata.arcgis.com",
792         "coord": "41.2647 N 96.0419 W",
793         "categorization": "Categories",
794         "platform": "ArcGis",
795         "categories": [
796             "Planning",
797             "Transportation",
798             "Boundaries",
799             "Elevation",
800             "Points of Interest"
801         ]
802     },
803     {
804         "city": "Oakland",
805         "url": "https://data.oaklandnet.com/browse",
806         "coord": "37.7699 N 122.2256 W",
807         "categorization": "Categories",
808         "platform": "Socrata",
809         "categories": [
810             "CityGovernment",
811             "Economic Develpment",
812             "Education",
813             "Environmental",
814             "Financial",
815             "Infrastructure",
816             "Life Enrichment",
817             "Property",
818             "Public Safety",
819             "Public Services"
820         ]
821     },
822     {
823         "city": "Minneapolis",
824         "url": "http://opendata.minneapolismn.gov/",
825         "coord": "44.9633 N 93.2683 W",
826         "categorization": "Categories",
827         "platform": "ArcGis",
828         "categories": [
829             "All Layers",
830             "Public Safety",
831             "Administrative Boundaries",
832             "Planning",
833             "City Assessor",
834             "Regulatory Services",
835             "Snow Emergency",
836             "Business Licensing",
837             "311"
838         ]
839     },
840     {
841         "city": "Tulsa",
842         "url": "http://opengis.cityoftulsa.org/",
```

```
843         "coord": "36.1279 N 95.9023 W",
844         "categorization": "-",
845         "platform": "ArcGis",
846         "categories": []
847     },
848     {
849         "city": "New Orleans",
850         "url": "https://data.nola.gov/",
851         "coord": "30.0686 N 89.9390 W",
852         "categorization": "Categories",
853         "platform": "Socrata",
854         "categories": [
855             "Arts, Culture, History",
856             "City Administration",
857             "City Finance and Budget",
858             "Customer Service",
859             "Demographics",
860             "Economy and Workforce",
861             "Elections, Politics",
862             "Environment",
863             "Geographic Base Layers",
864             "Health, Education, and Social Services",
865             "Housing, Land Use, and Blight",
866             "Parks, Parkways",
867             "Planning, Zoning",
868             "Public Safety and Preparedness",
869             "Real Estate, Land Records",
870             "Recreation and Culture",
871             "Transportation and Infrastructure"
872         ]
873     },
874     {
875         "city": "Wichita",
876         "url": "http://opendata.wichita.gov/",
877         "coord": "37.6907 N 97.3427 W",
878         "categorization": "Categories",
879         "platform": "ArcGis",
880         "categories": [
881             "Utilities",
882             "Boundaries",
883             "Business",
884             "Community Safety",
885             "Education",
886             "Environment and Health",
887             "Community Services",
888             "Transportation"
889         ]
890     },
891     {
892         "city": "Tampa",
893         "url": "http://city-tampa.opendata.arcgis.com/",
894         "coord": "27.9701 N 82.4797 W",
895         "categorization": "Categories",
896         "platform": "ArcGis",
```

```
897         "categories": [
898             "Location",
899             "Boundaries",
900             "Streets and Addresses",
901             "Planning",
902             "Utilities",
903             "Law Enforcment",
904             "Fire",
905             "Cadastral"
906         ]
907     },
908         {
909         "city": "Aurora",
910         "url": "http://data-auroraco.opendata.arcgis.com/",
911         "coord": "39.7082 N 104.8235 W",
912         "categorization": "Categories",
913         "platform": "ArcGis",
914         "categories": [
915             "Business",
916             "Boundaries",
917             "Community Services",
918             "Demographics",
919             "Education",
920             "Planimetrics",
921             "Property and Land",
922             "Water"
923         ]
924     },
925     {
926         "city": "Honolulu",
927         "url": "https://data.honolulu.gov/browse",
928         "coord": "21.3259 N 157.8453 W",
929         "categorization": "Categories",
930         "platform": "Socrata",
931         "categories": [
932             "Business",
933             "Finance",
934             "Location",
935             "Public Safety",
936             "Recreation",
937             "Transportation"
938         ]
939     },
940     {
941         "city": "Anaheim",
942         "url": "http://data-anaheim.opendata.arcgis.com/",
943         "coord": "33.8555 N 117.7601 W",
944         "categorization": "Categories",
945         "platform": "ArcGis",
946         "categories": [
947             "Boundaries",
948             "Base Map Data",
949             "Public Safety",
950             "Transportation",
```

```
951             "Public Works",
952             "Planning",
953             "Education",
954             "Recreation",
955             "Environmental",
956             "Licenses and Permits"
957         ]
958     },
959     {
960         "city": "Santa Ana",
961         "url": "http://opendata-santa-ana.opendata.arcgis.com/",
962         "coord": "33.7365 N 117.8826 W",
963         "categorization": "-",
964         "platform": "ArcGis",
965         "categories": []
966     },
967     {
968         "city": "Riverside",
969         "url": "https://data.countyofriverside.us/browse",
970         "coord": "33.9381 N 117.3932 W",
971         "categorization": "Categories",
972         "platform": "Socrata",
973         "categories": [
974             "Administrative and Fiscal Services",
975             "Culture and Recreation",
976             "Economy and Community",
977             "Geographic Locations and Boundaries",
978             "Health and Social Services",
979             "Land Use and Environment",
980             "Public Safety and Emergency Management",
981             "RCData Internal",
982             "RIVCOconnect BroadBand",
983             "Transportation"
984         ]
985     },
986     {
987         "city": "Lexington",
988         "url": "https://data.lexingtonky.gov/dataset",
989         "coord": "38.0402 N 84.4584 W",
990         "categorization": "Groups",
991         "platform": "CKAN",
992         "categories": [
993             "Historical",
994             "Community",
995             "Development",
996             "Environmental",
997             "Miscellaneous",
998             "Transportation"
999         ]
1000    },
1001    {
1002        "city": "St. Louis",
1003        "url": "https://brigades.opendatanetwork.com/brigade?brigade=Open%20Data%20STL",
1004        "coord": "38.6357 N 90.2446 W",
```

```
1005        "categorization": "Categories",
1006        "platform": "Socrata",
1007        "categories": [
1008            "Transparency",
1009            "Neighborhoods",
1010            "Inequality",
1011            "Environment",
1012            "Youth",
1013            "Citizens As Field Agents"
1014        ]
1015    },
1016    {
1017        "city": "Pittsburgh",
1018        "url": "http://pittsburghpa.gov/dcp/gis/gis-data/index.html",
1019        "coord": "40.4398 N 79.9766 W",
1020        "categorization": "Categories",
1021        "platform": "ArcGis",
1022        "categories": [
1023            "Art & Culture",
1024            "Business & Economy",
1025            "Civic Vitality & Governance",
1026            "Demographics",
1027            "Education",
1028            "Energy",
1029            "Environment",
1030            "Events",
1031            "Geography",
1032            "Health",
1033            "Housing & Property",
1034            "Human & Social Services",
1035            "Public Safety & Justice",
1036            "Recreation",
1037            "Transportation",
1038            "Other"
1039        ]
1040    },
1041    {
1042        "city": "Saint Paul",
1043        "url": "https://information.stpaul.gov/",
1044        "coord": "44.9489 N 93.1039 W",
1045        "categorization": "Categories",
1046        "platform": "Socrata",
1047        "categories": [
1048            "Buildings, Housing & Economic Development",
1049            "Census",
1050            "City Administration",
1051            "City Infrastructure",
1052            "Permits & Licensing",
1053            "Public Facilities & Services",
1054            "Public Safety"
1055        ]
1056    },
1057    {
1058        "city": "Cincinnati",
```

```
1059        "url": "https://data.cincinnati-oh.gov/",
1060        "coord": "39.1399 N 84.5064 W",
1061        "categorization": "Categories",
1062        "platform": "Socrata",
1063        "categories": [
1064            "Fiscal Sustainability & Strategic Investment",
1065            "Growing Economy",
1066            "Innovative Government",
1067            "Safer Streets",
1068            "Thriving Healthy and Neighborhoods"
1069        ]
1070    },
1071    {
1072        "city": "Anchorage",
1073        "url": "https://data.muni.org/browse",
1074        "coord": "61.2176 N 149.8953 W",
1075        "categorization": "Categories",
1076        "platform": "Socrata",
1077        "categories": [
1078            "Housing and Homelessness",
1079            "Other",
1080            "Public Health",
1081            "Public Safety",
1082            "Street Maintenance"
1083        ]
1084    },
1085    {
1086        "city": "Henderson",
1087        "url": "https://opendata.cityofhenderson.com/browse",
1088        "coord": "36.0122 N 115.0375 W",
1089        "categorization": "Categories",
1090        "platform": "Socrata",
1091        "categories": [
1092            "Building Permits",
1093            "Business License",
1094            "Crime Data",
1095            "Economic / Demographic",
1096            "Parks & Trails"
1097        ]
1098    },
1099    {
1100        "city": "Greensboro",
1101        "url": "https://data.greensboro-nc.gov/",
1102        "coord": "36.0965 N 79.8271 W",
1103        "categorization": "Categories",
1104        "platform": "Socrata",
1105        "categories": [
1106            "Business and Financial",
1107            "Government",
1108            "Housing and Development",
1109            "Public Safety",
1110            "Housing and Development"
1111        ]
1112    },
```

```
1113    {
1114        "city": "Plano",
1115        "url": "https://dashboard.plano.gov/",
1116        "coord": "33.0508 N 96.7479 W",
1117        "categorization": "Categories",
1118        "platform": "Socrata",
1119        "categories": [
1120            "Business",
1121            "Education",
1122            "Finance",
1123            "Government",
1124            "Health"
1125        ]
1126    },
1127    {
1128        "city": "Newark",
1129        "url": "http://data.ci.newark.nj.us/dataset",
1130        "coord": "40.7242 N 74.1726 W",
1131        "categorization": "Groups",
1132        "platform": "CKAN",
1133        "categories": [
1134            "Infrastructure",
1135            "Government",
1136            "Demographics",
1137            "Business",
1138            "Education",
1139            "Public Safety"
1140        ]
1141    },
1142    {
1143        "city": "Orlando",
1144        "url": "https://data.cityoforlando.net/browse",
1145        "coord": "28.4159 N 81.2988 W",
1146        "categorization": "Categories",
1147        "platform": "Socrata",
1148        "categories": [
1149            "Citywide",
1150            "Economic Development",
1151            "Government / General",
1152            "Public Safety",
1153            "Public Works"
1154        ]
1155    },
1156        {
1157        "city": "Chula Vista",
1158        "url": "http://chulavista-cvgis.opendata.arcgis.com/",
1159        "coord": "32.6277 N 117.0152 W",
1160        "categorization": "Categories",
1161        "platform": "ArcGis",
1162        "categories": [
1163            "Administrative",
1164            "Environmental",
1165            "Facilities",
1166            "Land Records",
```

```
1167            "Transportation"
1168        ]
1169    },
1170    {
1171        "city": "Jersey City",
1172        "url": "http://data.jerseycitynj.gov",
1173        "coord": "40.7114 N 74.0648 W",
1174        "categorization": "Groups",
1175        "platform": "CKAN",
1176        "categories": [
1177            "Infrastructure",
1178            "Maps",
1179            "Boards and Commisions of H.E.D.C.",
1180            "Municipal Services",
1181            "Housing",
1182            "Transportation",
1183            "Safety",
1184            "Environment",
1185            "Economy",
1186            "Other / Unclassified",
1187            "Demographics",
1188            "Education"
1189        ]
1190    },
1191    {
1192        "city": "Durham",
1193        "url": "https://opendurham.nc.gov/pages/home/",
1194        "coord": "35.9810 N 78.9056 W",
1195        "categorization": "Categories",
1196        "platform": "OpenDataSoft",
1197        "categories": [
1198            "Transport, Movements",
1199            "Services, Social",
1200            "Administration, Govenment, Public Finances, Tax Office",
1201            "Economy, Business, SME, Economic Development, Employemnt",
1202            "Education, Training, Research, Teaching",
1203            "Health",
1204            "Justice, Safety, Police, Crime",
1205            "Spatial Planning, Equipment",
1206            "Sports, Leisure",
1207            "Environment"
1208        ]
1209    },
1210    {
1211        "city": "Laredo",
1212        "url": "http://laredo-txlocalgov.opendata.arcgis.com/",
1213        "coord": "27.5477 N 99.4869 W",
1214        "categorization": "Categories",
1215        "platform": "ArcGis",
1216        "categories": [
1217            "Natural Resources",
1218            "Boundaries",
1219            "Business",
1220            "Community Safety",
```

```
1221            "Education",
1222            "Health",
1223            "Housing",
1224            "Transportation"
1225        ]
1226    },
1227    {
1228        "city": "Madison",
1229        "url": "https://data.cityofmadison.com/",
1230        "coord": "43.0878 N 89.4301 W",
1231        "categorization": "Categories",
1232        "platform": "ArcGis",
1233        "categories": [
1234            "City Facilities",
1235            "Elections",
1236            "Events",
1237            "Fire",
1238            "Library",
1239            "Metro Transportation",
1240            "Parks & Recreation",
1241            "Police",
1242            "Polling Places",
1243            "Property",
1244            "Public Safety",
1245            "Service Requests"
1246        ]
1247    },
1248    {
1249        "city": "Scottsdale",
1250        "url": "http://data-cos-gis.opendata.arcgis.com/",
1251        "coord": "33.6687 N 111.8237 W",
1252        "categorization": "Categories",
1253        "platform": "ArcGis",
1254        "categories": [
1255            "Facilities",
1256            "Land Base",
1257            "Planning",
1258            "Recreation"
1259        ]
1260    },
1261    {
1262        "city": "Glendale",
1263        "url": "http://data-cog-gis.opendata.arcgis.com/",
1264        "coord": "33.5331 N 112.1899 W",
1265        "categorization": "-",
1266        "platform": "ArcGis",
1267        "categories": []
1268    },
1269    {
1270        "city": "Reno",
1271        "url": "http://opendatareno.org/dataset",
1272        "coord": "39.4745 N 119.7765 W",
1273        "categorization": "-",
1274        "platform": "CKAN",
```

```
1275        "categories": []
1276    },
1277    {
1278        "city": "Norfolk",
1279        "url": "http://data-orf.opendata.arcgis.com/",
1280        "coord": "36.9230 N 76.2446 W",
1281        "categorization": "Categories",
1282        "platform": "ArcGis",
1283        "categories": [
1284            "Arts and Education",
1285            "Schools",
1286            "Elementary School Boundaries",
1287            "Middle School Boundaries",
1288            "High School Boundaries",
1289            "Libraries",
1290            "Electoral",
1291            "Electoral Polling Locations",
1292            "Electoral Precincts",
1293            "Wards",
1294            "Super Wards",
1295            "Environmental",
1296            "Breakwater",
1297            "Category 1 Storm Surge",
1298            "Category 2 Storm Surge",
1299            "Category 3 Storm Surge",
1300            "Category 4 Storm Surge",
1301            "Water",
1302            "Wetlands",
1303            "Municipal",
1304            "City Limit",
1305            "City Facilities",
1306            "ZIP Code",
1307            "Neighborhoods",
1308            "Civic Leagues",
1309            "Neighborhood Service Areas",
1310            "Parks and Recreation",
1311            "Beach Access",
1312            "Boat Ramps",
1313            "Elizabeth River Trail",
1314            "Parks",
1315            "Recreation Centers",
1316            "Planning and Zoning",
1317            "AICUZ - Accidental Potential Zones",
1318            "AICUZ - Noise Levels",
1319            "CBPA",
1320            "Conditional Zoning",
1321            "Economic Districts",
1322            "Flood Zone",
1323            "Future Land Use",
1324            "HUBZones",
1325            "Limit of Moderate Wave Action-LiMWA",
1326            "Planning Districts",
1327            "Special Exceptions",
1328            "Zoning",
```

```
1329            "Property Information",
1330            "Addresses",
1331            "Building Footprints",
1332            "Parcels",
1333            "Public Safety",
1334            "Fire Demand Zones",
1335            "Fire Station",
1336            "Police Car Districts",
1337            "Police Facilities",
1338            "Police Precincts",
1339            "Transportation",
1340            "Light Rail Route",
1341            "Street Centerline"
1342        ]
1343    },
1344    {
1345        "city": "Chesapeake",
1346        "url": "http://public-chesva.opendata.arcgis.com/",
1347        "coord": "36.6794 N 76.3018 W",
1348        "categorization": "Categories",
1349        "platform": "ArcGis",
1350        "categories": [
1351            "Boundaries",
1352            "Environmental",
1353            "Adresses",
1354            "Civic",
1355            "Transportation",
1356            "Rasters",
1357            "All Content"
1358        ]
1359    },
1360    {
1361        "city": "Fremont",
1362        "url": "http://egis-cofgis.opendata.arcgis.com/",
1363        "coord": "37.4944 N 121.9411 W",
1364        "categorization": "Layers",
1365        "platform": "ArcGis",
1366        "categories": [
1367            "Economic Development",
1368            "Engineering",
1369            "Landmark",
1370            "Planning",
1371            "Transportation",
1372            "All Layers"
1373        ]
1374    },
1375    {
1376        "city": "Baton Rouge",
1377        "url": "https://data.brla.gov/browse",
1378        "coord": "30.4485 N 91.1259 W",
1379        "categorization": "Categories",
1380        "platform": "Socrata",
1381        "categories": [
1382            "Business and Financial",
```

```
1383            "Culture and Recreation",
1384            "Government",
1385            "Housing and Development",
1386            "Public Safety",
1387            "Transportation and Infrastructure"
1388        ]
1389    },
1390        {
1391        "city": "Richmond",
1392        "url": "https://data.richmondgov.com/browse",
1393        "coord": "37.5314 N 77.4760 W",
1394        "categorization": "Categories",
1395        "platform": "Socrata",
1396        "categories": [
1397            "Community Safety and Well-Being",
1398            "Economic Growth",
1399            "Education and Workforce Development",
1400            "Sustainability and Natural Environment",
1401            "Transportation",
1402            "Unique and Inclusive Neighborhoods",
1403            "Well-Managed Government"
1404        ]
1405    },
1406    {
1407        "city": "Boise",
1408        "url": "http://opendata.cityofboise.org/",
1409        "coord": "43.5985 N 116.2311 W",
1410        "categorization": "Categories",
1411        "platform": "ArcGis",
1412        "categories": [
1413            "Parks & Rec",
1414            "Utilities",
1415            "Parking"
1416        ]
1417    },
1418    {
1419        "city": "San Bernardino",
1420        "url": "http://open.sbcounty.gov/datasets",
1421        "coord": "34.1393 N 117.2953 W",
1422        "categorization": "Categories",
1423        "platform": "ArcGis",
1424        "categories": [
1425            "Boundaries",
1426            "Demographics",
1427            "Property",
1428            "Public Reporting",
1429            "Transportation",
1430            "Zoning"
1431        ]
1432    },
1433    {
1434        "city": "Spokane",
1435        "url": "http://data-spokane.opendata.arcgis.com/",
1436        "coord": "47.6669 N 117.4333 W",
```

```
1437        "categorization": "Categories",
1438        "platform": "ArcGis",
1439        "categories": [
1440            "Boundaries",
1441            "Demographics",
1442            "Environment",
1443            "Building and Planning",
1444            "Public Safety",
1445            "Reference Data",
1446            "Public Works",
1447            "Transportation"
1448        ]
1449    },
1450    {
1451        "city": "Birmingham",
1452        "url": "https://data.birminghamal.gov",
1453        "coord": "33.5274 N 86.7990 W",
1454        "categorization": "Groups",
1455        "platform": "CKAN",
1456        "categories": [
1457            "Fire",
1458            "Finance",
1459            "Police",
1460            "Planning Engineering",
1461            "Economic Development",
1462            "Youth Services",
1463            "Municipal Courts",
1464            "Community Development",
1465            "Birmingham"
1466        ]
1467    },
1468    {
1469        "city": "Tacoma",
1470        "url": "https://data.cityoftacoma.org/browse",
1471        "coord": "47.2522 N 122.4598 W",
1472        "categorization": "Categories",
1473        "platform": "Socrata",
1474        "categories": [
1475            "Business",
1476            "City Administration and Finance",
1477            "Community and Economic Development",
1478            "Environment and Sustainability",
1479            "Equity",
1480            "Human Services",
1481            "Infrastructure and Transportation",
1482            "Neighborhoods",
1483            "Public Safety"
1484        ]
1485    },
1486    {
1487        "city": "Oxnard",
1488        "url": "https://data.oxnard.org/",
1489        "coord": "34.2023 N 119.2046 W",
1490        "categorization": "Categories",
```

```
1491        "platform": "Socrata",
1492        "categories": [
1493            "Archive",
1494            "Building Permits",
1495            "Financial Data and Reports",
1496            "Local Business Data"
1497        ]
1498    },
1499    {
1500        "city": "Fayetteville",
1501        "url": "http://data.fayettevillenc.gov/",
1502        "coord": "35.0828 N 78.9735 W",
1503        "categorization": "Categories",
1504        "platform": "ArcGis",
1505        "categories": [
1506            "All Data Sets",
1507            "Business, Budget & Financials",
1508            "Parks & Recreation",
1509            "Calls for Service",
1510            "Planning & Zoning",
1511            "Police",
1512            "Fire",
1513            "Transportation"
1514        ]
1515    },
1516    {
1517        "city": "Montgomery",
1518        "url": "https://data.montgomeryal.gov",
1519        "coord": "32.3472 N 86.2661 W",
1520        "categorization": "Categories",
1521        "platform": "Socrata",
1522        "categories": [
1523            "311",
1524            "Development",
1525            "Paving",
1526            "Permits",
1527            "Property Maintenance",
1528            "Public Safety",
1529            "Public Works"
1530        ]
1531    },
1532    {
1533        "city": "Little Rock",
1534        "url": "https://data.littlerock.gov/",
1535        "coord": "34.7254 N 92.3586 W",
1536        "categorization": "Categories",
1537        "platform": "Socrata",
1538        "categories": [
1539            "311",
1540            "Community Services",
1541            "Government",
1542            "Parks",
1543            "Public Safety"
1544        ]
```

```
1545    },
1546    {
1547        "city": "Akron (County of Summit)",
1548        "url": "http://data-summitgis.opendata.arcgis.com/",
1549        "coord": "41.0805 N 81.5214 W",
1550        "categorization": "Categories",
1551        "platform": "ArcGis",
1552        "categories": [
1553            "Land Records",
1554            "Boundaries",
1555            "Environmental",
1556            "Sewer",
1557            "Engineer",
1558            "Development",
1559            "Transportation",
1560            "Topography"
1561        ]
1562    },
1563    {
1564        "city": "Grand Rapids",
1565        "url": "http://data.grcity.us/dataset",
1566        "coord": "42.9612 N 85.6556 W",
1567        "categorization": "-",
1568        "platform": "CKAN",
1569        "categories": []
1570    },
1571        {
1572        "city": "Salt Lake City",
1573        "url": "http://gis-slcgov.opendata.arcgis.com/",
1574        "coord": "40.7769 N 111.9310 W",
1575        "categorization": "Categories",
1576        "platform": "ArcGis",
1577        "categories": [
1578            "Administrative",
1579            "Atlas Plats",
1580            "Base Map",
1581            "Planning and Zoning",
1582            "Parks & Recreation"
1583        ]
1584    },
1585    {
1586        "city": "Huntsville",
1587        "url": "https://www.huntsvilleal.gov/development/building-construction/gis/data-
            depot/open-data/",
1588        "coord": "34.6990 N 86.6730 W",
1589        "categorization": "Categories",
1590        "platform": "ArcGis",
1591        "categories": [
1592            "All Open Data",
1593            "Boundaries",
1594            "Census Data",
1595            "Flood",
1596            "Grids",
1597            "Hydrography",
```

```
1598            "Map Indexes",
1599            "Planning",
1600            "Parks and Recreation",
1601            "Points of Interest",
1602            "Public Safety",
1603            "Schools",
1604            "Stormwater",
1605            "Transportation",
1606            "Miscellaneous Data"
1607        ]
1608    },
1609    {
1610        "city": "Mobile",
1611        "url": "http://maps.cityofmobile.org/gis/gisdata_datacatalog.aspx",
1612        "coord": "30.6684 N 88.1002 W",
1613        "categorization": "Categories",
1614        "platform": "-",
1615        "categories": []
1616    },
1617    {
1618        "city": "Tallahassee",
1619        "url": "http://talgov-tlcgis.opendata.arcgis.com/",
1620        "coord": "30.4551 N 84.2534 W",
1621        "categorization": "Categories",
1622        "platform": "ArcGis",
1623        "categories": [
1624            "All Data Sets",
1625            "Business & Budget",
1626            "Infrastructure",
1627            "Land Use & Permits",
1628            "Public Safety",
1629            "Service Requests",
1630            "Transportation",
1631            "TLCGis Open Data"
1632        ]
1633    },
1634    {
1635        "city": "Knoxville",
1636        "url": "http://www.knoxvilletn.gov/cms/One.aspx?portalId=109562&pageId=7240314",
1637        "coord": "35.9707 N 83.9493 W",
1638        "categorization": "Avalilable Data",
1639        "platform": "-",
1640        "categories": [
1641            "311 Performance Measures",
1642            "Blight Data Dashboard",
1643            "Budget",
1644            "City Council Agenda/Minutes",
1645            "Crime Map",
1646            "East Tennessee Index",
1647            "Homemaker Program",
1648            "KGIS Maps",
1649            "KnoxHMIS",
1650            "NPDES Reports",
1651            "Police Advisory and Review Committee",
```

```
1652            "Police Department",
1653            "Property Tax Database - City",
1654            "Property Tax Database - County",
1655            "Public Improvement Projects",
1656            "Rainfall Data",
1657            "TIFs and PILOTs",
1658            "Tree Inventory"
1659        ]
1660    },
1661    {
1662        "city": "Worcester",
1663        "url": "http://gisdata.worcesterma.gov/",
1664        "coord": "42.2695 N 71.8078 W",
1665        "categorization": "-",
1666        "platform": "-",
1667        "categories": []
1668    },
1669    {
1670        "city": "Tempe",
1671        "url": "http://data-tempegov.opendata.arcgis.com/",
1672        "coord": "33.3884 N 111.9318 W",
1673        "categorization": "Categories",
1674        "platform": "ArcGis",
1675        "categories": [
1676            "Accessibility",
1677            "Transportation",
1678            "Neighborhoods & Historic Preservation",
1679            "Land Use",
1680            "Public Safety",
1681            "All Datasets"
1682        ]
1683    },
1684    {
1685        "city": "Santa Clarita",
1686        "url": "http://www.santa-clarita.com/residents/open-data-portal",
1687        "coord": "34.4030 N 118.5042 W",
1688        "categorization": "Categories",
1689        "platform": "OpenGov",
1690        "categories": [
1691            "Annual Budget",
1692            "Employee Effectiveness",
1693            "Time To Respond",
1694            "Employee Courtesy",
1695            "Expectations Met",
1696            "Election Data at a Glance",
1697            "Election Data by Precinct",
1698            "Election Results by Candidate",
1699            "Film Permits Issued & Days Of Filming",
1700            "Film Revenue",
1701            "Hotel Occupancy",
1702            "Number of Businesses",
1703            "Number of Jobs",
1704            "Unemployment Rate",
1705            "Library Circulation",
```

```
1706            "Transit Ridership",
1707            "Graffiti Cleanup",
1708            "DFYinSCV Membership",
1709            "LMD Water Usage"
1710        ]
1711    },
1712    {
1713        "city": "Cape Coral",
1714        "url": "http://capecoral-capegis.opendata.arcgis.com/",
1715        "coord": "26.6432 N 81.9974 W",
1716        "categorization": "Categories",
1717        "platform": "ArcGis",
1718        "categories": [
1719            "311 Call Center",
1720            "Business & Budget",
1721            "Planning & Zoning",
1722            "Public Works",
1723            "Gis",
1724            "Fire",
1725            "Brownse All Data"
1726        ]
1727    },
1728    {
1729        "city": "Providence",
1730        "url": "https://data.providenceri.gov/",
1731        "coord": "41.8231 N 71.4188 W",
1732        "categorization": "Categories",
1733        "platform": "Socrata",
1734        "categories": [
1735            "Economy / Finance",
1736            "Neighborhoods",
1737            "Public Safety",
1738            "Reference"
1739        ]
1740    },
1741        {
1742        "city": "Chattanooga",
1743        "url": "https://data.chattlibrary.org/browse?Organization_Name=City+of+
                Chattanooga",
1744        "coord": "35.0660 N 85.2484 W",
1745        "categorization": "Categories",
1746        "platform": "Socrata",
1747        "categories": [
1748            "Buildings & Trails",
1749            "Economy",
1750            "Education",
1751            "Government",
1752            "Public Health",
1753            "Public Safety",
1754            "Recreation",
1755            "Social Services",
1756            "Transportation"
1757        ]
1758    },
```

```
1759    {
1760        "city": "Santa Rosa",
1761        "url": "https://data.srcity.org/",
1762        "coord": "38.4468 N 122.7061 W",
1763        "categorization": "Categories",
1764        "platform": "Socrata",
1765        "categories": [
1766            "Community Services",
1767            "Development",
1768            "Economy",
1769            "Finances",
1770            "Fire",
1771            "Government",
1772            "HCS Department",
1773            "HR Department",
1774            "Housing",
1775            "IT Department",
1776            "Internal Datasets",
1777            "Office of Community Engagement",
1778            "Police",
1779            "Public Safety",
1780            "Recreation and Culture",
1781            "Transportation",
1782            "Water Department"
1783        ]
1784    },
1785    {
1786        "city": "Sioux Falls",
1787        "url": "http://gisopendata.siouxfalls.org/",
1788        "coord": "43.5383 N 96.7320 W",
1789        "categorization": "Categories",
1790        "platform": "ArcGis",
1791        "categories": [
1792            "Community",
1793            "Education",
1794            "Imagery",
1795            "Infrastructure",
1796            "Parks",
1797            "Property",
1798            "Safety",
1799            "Transportation"
1800        ]
1801    },
1802    {
1803        "city": "McKinney",
1804        "url": "http://mckinneygis-mck.opendata.arcgis.com/",
1805        "coord": "33.1985 N 96.6680 W",
1806        "categorization": "Categories",
1807        "platform": "ArcGis",
1808        "categories": [
1809            "Planning & Zoning",
1810            "Addresses & Boundaries",
1811            "Schools",
1812            "Engineering",
```

```
1813            "Public Safety",
1814            "Parks & Recreation",
1815            "Transportation",
1816            "All Other Data"
1817        ]
1818    },
1819    {
1820        "city": "ElkGrove",
1821        "url": "http://gisdata.elkgrovecity.org/",
1822        "coord": "38.4146 N 121.3850 W",
1823        "categorization": "Layers",
1824        "platform": "ArcGis",
1825        "categories": [
1826            "Base Layers",
1827            "Building Footprints",
1828            "City Limits",
1829            "City Council Boundary",
1830            "Fire Stations",
1831            "CSD Parks",
1832            "EGUSD Schools",
1833            "Parking Lot Centerlines",
1834            "Public Amenity Footprints",
1835            "Land Layers",
1836            "General Plan",
1837            "Home Owners Associations",
1838            "Neighborhoods Associations",
1839            "Parcels",
1840            "Sphere of Influence",
1841            "Southeast Planning Area",
1842            "Zoning",
1843            "Drainage Layers",
1844            "Creeks and Channels",
1845            "Detention Basins",
1846            "Drainage Features",
1847            "Transportation Layers",
1848            "eTran Bus Routes",
1849            "eTran Bus Stops",
1850            "Bikeways",
1851            "Park and Ride Locations",
1852            "Rail Roads",
1853            "Street Centerlines",
1854            "Curb Ramps",
1855            "Sidewalks",
1856            "Infrastructure / Asset Layers",
1857            "Street Lights",
1858            "Landscape Areas",
1859            "Data Gateway Layers",
1860            "Building Permits",
1861            "Business Licenses",
1862            "Capital Improvement Projects"
1863        ]
1864    }
1865 ]
```

APÊNDICE D – CHAMADA DAS FUNÇÕES PARA OBTENÇÃO DO SUBCONJUNTO ABRANGENTE

Neste apêndice é apresentado o código para chamada das funções do processo de Obtenção do Subconjunto Abrangente de Categorias aplicado ao Estudo de Caso com os dados dos portais obtidos na Pesquisa Exploratória, descritos no Capítulo 4.

```
1 lstPortal = readPortalsFromJsonFile()
2 print("Number of Portals: " + "{0}".format(len(lstPortal)))
3
4 lstCategories = allCategories(lstPortal)
5 print("Number of Categories: " + "{0}".format(len(lstCategories)))
6
7 tokens = tokenizer(lstCategories)
8 print("Number of tokens in categories: " + "{0}".format(len(tokens)))
9
10 words_to_remove = ["&","gis","/","kc","fy","foia","geo","city","data","go",
11                     "-",",","houston","use","public","department",".]
12
13 lstWords = removeStopWords(tokens, words_to_remove)
14
15 dictWordFreq = frequency_word_count(lstWords)
16 print("Number of words in categories: " + "{0}".format(len(dictWordFreq)))
17 print("\n")
18 print(dictWordFreq)
19 print("\n")
20
21 dictPortalsCoverage = fillDictPortalsCoverage(dictWordFreq, lstPortal)
22 print(dictPortalsCoverage)
23 print("\n")
24
25 trheshold = 98.85057471264369
26 more_coverage_words = more_coverage_words(dictPortalsCoverage, trheshold)
27 print(more_coverage_words)
28 print("\n")
29
30 dictWordCategoryFreq = fillDictWordCategoryFreq(more_coverage_words, lstPortal,
     words_to_remove)
31 print(dictWordCategoryFreq)
```

```
print("\n")

dictWordFrequentlyCategories = fillDictWordFrequentlyCategories(dictWordCategoryFreq)
print(dictWordFrequentlyCategories)
print("\n")

write_categories(dictWordFrequentlyCategories)
```

APÊNDICE E – CHAMADA DAS FUNÇÕES PARA O ALINHAMENTO DE CATEGORIAS

Neste apêndice é apresentado o código para chamada das funções do processo de Alinhamento de Categorias aplicado ao Estudo de Caso com os dados dos portais obtidos na Pesquisa Exploratória, descritos no Capítulo 4.

```
1 print("\n")
2 first_synset = True
3
4 lstPortal = readPortalsFromJsonFile()
5 print("Number of Portals: " + "{0}".format(len(lstPortal)))
6
7 lstCategoriesCoverage = readCategoriesFromJsonFile()
8 print("Number of Best Coverage Categories: " + "{0}".format(len(lstCategoriesCoverage)))
9 print("\n")
10
11 start = time.time()
12
13 dictPortalsCategoryMatch, dictPortalsCategorySimilarities = fillDictPortalsCategoryMatch(
      lstPortal, lstCategoriesCoverage)
14 write_categories_match(dictPortalsCategoryMatch, dictPortalsCategorySimilarities)
15
16 end = time.time()
17
18 print("\n")
19 time = (end - start)
20 print('duracao: {:.2f}'.format(time) + " s")
```

APÊNDICE F – RESULTADO DO ALINHAMENTO DE CATEGORIAS PARA O ESTUDO DE CASO

Neste Apêndice são apresentados os resultados do processo de Alinhamento de categorias do Estudo de Caso nos 100 portais das cidades americanas densamente populosas. Os dados do alinhamento são apresentados em formato de arquivo *JSON*.


```
[
    [
        "NYC",
        {
            "Business": "Business",
            "City Government": "Government",
            "Education": "Education",
            "Environment": "Environment",
            "Health": "Health",
            "Housing & Development": "Economic Development",
            "Public Safety": "Public Safety",
            "Recreation": "Education",
            "Social Services": "Community",
            "Transportation": "Transportation"
        }
    ],
    [
        "Los Angeles",
        {
            "A Livable and Sustainable City": "City Services",
            "A Prosperous City": "City Services",
            "A Safe City": "City Services",
            "A Well Run City": "City Services"
        }
    ],
    [
        "Chicago ",
        {
            "Administration & Finance": "Finance",
            "Buildings": "Transportation",
            "Community": "Community",
            "Education": "Education",
```

```
33            "Environment": "Environment",
34            "Ethics": "Education",
35            "Events": "Education",
36            "FOIA": "",
37            "Facilities & Geo. Boundaries ": "Boundaries",
38            "Health & Human Services": "Health",
39            "Historic Preservation": "Education",
40            "Parks & Recreation": "Parks & Recreation",
41            "Public Safety": "Public Safety",
42            "Sanitation": "Environment",
43            "Service Requests": "Education",
44            "Transportation": "Transportation"
45        }
46    ],
47    [
48        "Houston",
49        {
50            "GIS": "Property",
51            "City of Houston Entreprise GIS": "City Services",
52            "City of Houston Planning and Development Department": "Economic Development"
                ,
53            "Government Boundaries": "Government",
54            "Public Works & Engineering": "Public Safety",
55            "Planning & Development": "Economic Development",
56            "Transportation": "Transportation",
57            "Permitting and Licensing": "Planning",
58            "City of Houston Public Works and Engineering": "City Services",
59            "Public Health & Safety": "Health",
60            "City of Houston Administration and Regulatory Affairs": "City Services",
61            "Neighborhood Services": "City Services",
62            "Finance": "Finance",
63            "Environmental": "",
64            "Property": "Property",
65            "Hydrology": "Education",
66            "City of Houston Finance Department": "Finance",
67            "Flood Hazard": "Transportation",
68            "Adresses Roads Cadastral": "Boundaries",
69            "Houston Fire Department": "Police",
70            "Parking": "Planning",
71            "Houston Police Department": "Police",
72            "City of Houston Health & Human Services Department": "City Services",
73            "City of Houston Department of Neighborhood": "City Services",
74            "City of Houston Solid Waste Mangement": "City Services",
75            "Restaurants": "Transportation",
76            "Education and Schooling": "Education",
77            "Econonomic Development": "Economic Development",
78            "City of Houston Parks and Recreation Department": "Parks & Recreation",
79            "City of Houston General Services Department": "City Services",
80            "City of Houston City Secretary": "Boundaries",
81            "Houston-Galveston Area Council": "Police",
82            "Demographics": "Education",
83            "City of Houston Office of Bussiness Opportunity": "Boundaries",
84            "City of Houston Human Resources": "City Services",
85            "City of Houston House & Community Development": "Economic Development",
```

```
86              "City of Houston City Council": "City Services",
87              "Houston Public Library": "Public Safety",
88              "City of Houston Legal Department": "City Services",
89              "City of Houston Information Technology Services": "City Services",
90              "City of Houston Aviation / Houston Airport System": "Transportation"
91          }
92      ],
93      [
94          "Phoenix",
95          {
96              "Census Data": "Education",
97              "CheckBook & Sales Tax": "Property",
98              "Energy & Sustainability": "Infrastructure",
99              "Neighborhood & Safety": "Public Safety",
100             "Parks, Art & Culture": "Parks & Recreation",
101             "Property & Development": "Economic Development",
102             "Staff Salaries": "Police",
103             "Transportation": "Transportation"
104         }
105     ],
106     [
107         "Philadelphia",
108         {
109             "Transportation": "Transportation",
110             "Real State / Land Records": "Property",
111             "Environment": "Environment",
112             "Health / Human Services": "Health",
113             "Planning / Zoning": "Planning",
114             "Elections / Politics": "Property",
115             "Arts / Culture / History": "Business",
116             "Education": "Education",
117             "Public Safety": "Public Safety",
118             "Parks / Recreation": "Parks & Recreation",
119             "Economy": "Economy",
120             "Uncategorized": "",
121             "Food": "Transportation",
122             "Budget / Finance": "Finance"
123         }
124     ],
125     [
126         "San Antonio",
127         {}
128     ],
129     [
130         "San Diego",
131         {
132             "Economy & Community": "Community",
133             "City Infrastructure": "Infrastructure",
134             "City Management": "City Services",
135             "Transportation": "Transportation",
136             "Public Safety": "Public Safety",
137             "Energy & Environment": "Environment",
138             "Culture & Recreation": "Education"
139         }
```

```
140     ],
141     [
142         "Dallas",
143         {
144             "Budget & Finance": "Finance",
145             "City Infrastructure": "Infrastructure",
146             "City Services": "City Services",
147             "Economic Development": "Economic Development",
148             "Geography & Boundaries": "Boundaries",
149             "Government": "Government",
150             "Public Safety": "Public Safety"
151         }
152     ],
153     [
154         "San Jose",
155         {
156             "Aiport": "",
157             "Auditor": "Transportation",
158             "City's Manager Office": "Transportation",
159             "City Wide": "City Services",
160             "Department Of Transportation": "Transportation",
161             "Economic Development": "Economic Development",
162             "Environmental Services": "City Services",
163             "Finance": "Finance",
164             "Fire": "Education",
165             "Housing": "Housing",
166             "Human Resources": "Property",
167             "Independent Police Auditor": "Police",
168             "Information Technology": "Land Use",
169             "Library": "",
170             "Parks, Recreation & Neighborhood Services": "Parks & Recreation",
171             "Planning Building and Code Enforcement": "Planning",
172             "Police": "Police",
173             "Public Works": "Public Safety",
174             "Retirement": "Environment"
175         }
176     ],
177     [
178         "Austin",
179         {
180             "Bussiness": "",
181             "Capital Metro": "Business",
182             "Capital Planning": "Education",
183             "Education": "Education",
184             "Environmental": "",
185             "Financial": "",
186             "Fun": "Parks & Recreation",
187             "Geo Data": "Business",
188             "Government": "Government",
189             "Health": "Health",
190             "Neighborhood": "Boundaries",
191             "Permitting": "Planning",
192             "Public Safety": "Public Safety",
193             "Utility": "Business",
```

```
            "Workforce Development": "Economic Development"
        }
    ],
    [
        "Jacksonville",
        {
            "Budget and Finance": "Finance",
            "City Services": "City Services",
            "Community Development": "Economic Development",
            "Economic Develpment": "",
            "Government Accountability": "Government",
            "Public Safety": "Public Safety",
            "Public Works ": "Public Safety",
            "Recreation and Parks": "Parks & Recreation",
            "Schools and Education": "Education",
            "Transportation and Transit": "Transportation"
        }
    ],
    [
        "San Francisco",
        {
            "City Infrastructure": "Infrastructure",
            "City Management and Ethics": "City Services",
            "Culture and Recreation": "Education",
            "Economy and Community": "Community",
            "Energy and Environment": "Environment",
            "Geographic Locations and Boundaries": "Boundaries",
            "Health and Social Services": "Health",
            "Housing and Buildings": "Transportation",
            "Public Safety": "Public Safety",
            "Transportation": "Transportation"
        }
    ],
    [
        "Columbus",
        {
            "Business": "Business",
            "Boundaries": "Boundaries",
            "Health": "Health",
            "Infrastructure": "Infrastructure",
            "Planning": "Planning",
            "Recreation & Parks": "Parks & Recreation",
            "Safety": "Public Safety",
            "Schools": "Business",
            "Transportation": "Transportation",
            "All": ""
        }
    ],
    [
        "Indianapolis",
        {
            "Boundaries": "Boundaries",
            "Transportation": "Transportation",
            "Recreation": "Parks & Recreation",
```

```
            "Property": "Property",
            "Disclose Indy": "",
            "Political": "",
            "Survey": "Education",
            "Planning & Zonning": "Planning"
        }
    ],
    [
        "Fort Worth",
        {
            "Business": "Business",
            "City Government": "Government",
            "Environment & Health": "Health",
            "Financial": "",
            "Property Data": "Property",
            "Public Safety": "Public Safety",
            "Services & Recreation": "Education",
            "Technology & Communications": "Education",
            "Tranpostation": ""
        }
    ],
    [
        "Charlotte",
        {
            "311 / Services": "City Services",
            "Arts & Education": "Education",
            "Business & Budget": "Business",
            "Community Safety": "Community",
            "City Government": "Government",
            "Neighborhoods & Housing": "Boundaries",
            "Transportation": "Transportation",
            "Environment": "Environment",
            "Demographics": "Education",
            "Planning & Zoning": "Planning",
            "Map Features": "Transportation",
            "Historical Data": "Business"
        }
    ],
    [
        "Seattle",
        {
            "City Business": "Business",
            "Community": "Community",
            "Education": "Education",
            "Finance": "Finance",
            "Land Base": "Property",
            "Parks": "Parks & Recreation",
            "Parks and Recreation": "Parks & Recreation",
            "Permitting": "Planning",
            "Public Safety": "Public Safety",
            "Transportation": "Transportation"
        }
    ],
    [
```

```
302         "Denver",
303         {}
304     ],
305     [
306         "Washington",
307         {
308             "Agriculture": "Finance",
309             "Consumer Protection": "Education",
310             "Demographics": "Education",
311             "Economics": "Education",
312             "Education": "Education",
313             "Employment": "Environment",
314             "Health": "Health",
315             "Labor": "Public Safety",
316             "Natural Resources & Environment": "Environment",
317             "Politics": "Property",
318             "Procurements and Contracts": "Education",
319             "Public Safety": "Public Safety",
320             "Recreation": "Parks & Recreation",
321             "Transportation": "Transportation"
322         }
323     ],
324     [
325         "Boston",
326         {
327             "City Services": "City Services",
328             "Facilites": "",
329             "Finance": "Finance",
330             "Health": "Health",
331             "Permitting": "Planning",
332             "Public Safety": "Public Safety",
333             "Transportation": "Transportation"
334         }
335     ],
336     [
337         "Detroit",
338         {
339             "Business": "Business",
340             "Children & Families": "Business",
341             "Education": "Education",
342             "Fun": "Education",
343             "Government": "Government",
344             "Personal": "",
345             "Property & Parcels": "Property",
346             "Public Health": "Health",
347             "Public Safety": "Public Safety",
348             "Transportation": "Transportation"
349         }
350     ],
351     [
352         "Nashville",
353         {
354             "Agriculture": "Finance",
355             "Art": "Transportation",
```

```
356             "Beautification": "Economic Development",
357             "Budget / Finance": "Finance",
358             "Business, Development & Housing": "Business",
359             "Culture": "Community",
360             "Education": "Education",
361             "Elections": "Education",
362             "Emergency Management": "Education",
363             "Energy Usage": "Land Use",
364             "Environment": "Environment",
365             "Fire ": "Education",
366             "Geneology": "",
367             "General Government": "Government",
368             "Health": "Health",
369             "History": "Environment",
370             "Libraries": "Transportation",
371             "Licenses & Permits": "Property",
372             "Medical": "",
373             "Metro Government": "Government",
374             "Parks": "Parks & Recreation",
375             "Police": "Police",
376             "Public Safety": "Public Safety",
377             "Public Services": "Public Safety",
378             "Recycling/Conservation": "",
379             "Social Services": "City Services",
380             "Transportation": "Transportation"
381         }
382     ],
383     [
384         "Portland",
385         {}
386     ],
387     [
388         "Oklahoma City",
389         {
390             "General": "Transportation",
391             "GO Bond Projects (2017)": "Education",
392             "Planimetrics": "",
393             "Subdivision": "Boundaries",
394             "Utilities": "Business",
395             "Census": "Education",
396             "Impact Fees": "Property",
397             "Public Safety": "Public Safety",
398             "Survey Control Points": "Education",
399             "Zoning": "Zoning",
400             "Go Bond Projects (2007)": "Education",
401             "Parks and Trails": "Parks & Recreation",
402             "School": "Business",
403             "Transportation": "Transportation"
404         }
405     ],
406     [
407         "Las Vegas",
408         {
409             "Arts and Culture": "Community",
```

```
410             "Building and Safety": "Public Safety",
411             "Community Risk Reduction": "Community",
412             "Economic Development": "Economic Development",
413             "Finance": "Finance",
414             "General Information": "Property",
415             "Growing Economy": "Economy",
416             "High Performing Government": "Government",
417             "Neighborhood Livability": "Boundaries",
418             "Parks and Recreation": "Parks & Recreation",
419             "Planning": "Planning",
420             "Public Safety": "Public Safety",
421             "Public Works": "Public Safety",
422             "Schools": "Business"
423         }
424     ],
425     [
426         "Louisville",
427         {}
428     ],
429     [
430         "Baltimore",
431         {
432             "City Government": "Government",
433             "City Services": "City Services",
434             "Crime": "Education",
435             "Culture & Arts": "Community",
436             "Financial": "",
437             "Geographic": "",
438             "Health": "Health",
439             "Housing & Development": "Economic Development",
440             "Neighborhoods": "Boundaries",
441             "Public Safety": "Public Safety",
442             "Public Works": "Public Safety",
443             "Transportation": "Transportation"
444         }
445     ],
446     [
447         "Tucson",
448         {}
449     ],
450     [
451         "Sacramento",
452         {
453             "Animal Care": "Education",
454             "Budget & Finance": "Finance",
455             "Disclosure & Ethics": "Education",
456             "Economy & Community": "Community",
457             "Locations & Mapping": "Boundaries",
458             "Parks & Recreation": "Parks & Recreation",
459             "Permits & Planning": "Education",
460             "Public Safety": "Public Safety",
461             "Service Requests": "Education",
462             "Transportation & Infrastructure": "Transportation"
463         }
```

```
464     ],
465     [
466         "Mesa",
467         {
468             "Energy & Utilities": "Business",
469             "Permits & Licences": "Environment",
470             "Financials": "",
471             "Recreation & Culture": "Education",
472             "Zoning & Property": "Property",
473             "Planes, Trains & Automobiles": "Transportation",
474             "Neighborhoods": "Boundaries",
475             "Public Safety": "Public Safety"
476         }
477     ],
478     [
479         "Kansas City",
480         {
481             "311": "",
482             "Airport": "Transportation",
483             "Annual Audit Plan & Reports": "Property",
484             "Area Plans": "Boundaries",
485             "Auction": "Education",
486             "Audit Highlights": "Property",
487             "Audits": "Property",
488             "Audits and Memos": "Property",
489             "Brownfields": "",
490             "Budget": "Planning",
491             "Business": "Business",
492             "Census": "Education",
493             "Climate Protection Steering Com": "Education",
494             "Code Interpretations": "Education",
495             "Construction": "Education",
496             "Crime": "Education",
497             "Development": "Economic Development",
498             "Development Review": "Economic Development",
499             "Emergency": "Education",
500             "Environmental Management Commission": "Police",
501             "FY 2009 - 2010": "",
502             "FY 2010 - 2011": "",
503             "FY 2011 - 2012": "",
504             "FY 2012 - 2013": "",
505             "FY 2013 - 2014": "",
506             "FY 2014 - 2015": "",
507             "FY 2015 - 2016": "",
508             "FY 2016 - 2017": "",
509             "Fees": "Property",
510             "Finance": "Finance",
511             "Focus": "Education",
512             "Food": "Transportation",
513             "Forecasts": "Property",
514             "Forms": "Property",
515             "Forms and Applications": "Land Use",
516             "GIS": "Property",
517             "Government": "Government",
```

```
518             "Health": "Health",
519             "Historic Preservation": "Education",
520             "Housing": "Housing",
521             "Human Relations": "Property",
522             "Human Resources": "Property",
523             "Information Bulletins": "Property",
524             "Innovation": "Transportation",
525             "Investor Relations": "Property",
526             "KC Bizcare": "Property",
527             "KC City Energy Project": "City Services",
528             "KCI Terminal Advisory": "Transportation",
529             "Land Development": "Economic Development",
530             "Land Use": "Land Use",
531             "Legislation": "Business",
532             "Legislative Info": "Property",
533             "Monthly Status Reports": "Environment",
534             "Municipal Court": "Community",
535             "Neighborhoods": "Boundaries",
536             "New Tax Forms": "Property",
537             "Old Tax Forms": "Property",
538             "Peer Review": "Parks & Recreation",
539             "Property": "Property",
540             "Regulated Industries": "Business",
541             "Regulatory Codes": "Property",
542             "Safety": "Public Safety",
543             "Scope Statement": "Infrastructure",
544             "Streetcar": "Transportation",
545             "Sustainability": "Infrastructure",
546             "Taxes": "Property",
547             "Traffic": "Business",
548             "Traffic Sign Changes": "Business",
549             "Transportation": "Transportation",
550             "Workforce": "Police"
551         }
552     ],
553     [
554         "Atlanta",
555         {}
556     ],
557     [
558         "Long Beach",
559         {
560             "Business": "Business",
561             "Boundaries": "Boundaries",
562             "Health": "Health",
563             "Infrastructure": "Infrastructure",
564             "Planning": "Planning",
565             "Recreation & Parks": "Parks & Recreation",
566             "Safety": "Public Safety",
567             "Schools": "Business",
568             "Transportation": "Transportation",
569             "All": ""
570         }
571     ],
```

```
572     [
573         "Colorado Springs",
574         {
575             "Economy and Community": "Community",
576             "City Management and Ethics": "City Services",
577             "Transportation": "Transportation",
578             "Public Safety": "Public Safety",
579             "Health and Social Services": "Health",
580             "Geographic Locations and Boundaries": "Boundaries",
581             "Energy and Environment": "Environment",
582             "Housing and Buildings": "Transportation",
583             "City Infrastructure": "Infrastructure",
584             "Culture and Recreation": "Education"
585         }
586     ],
587     [
588         "Raleigh",
589         {
590             "Alt. Fuels": "Property",
591             "Budget and Management Services": "Finance",
592             "Business": "Business",
593             "Downtown": "Boundaries",
594             "Fire": "Education",
595             "Government": "Government",
596             "Government Buldings and Structures": "Government",
597             "Greenways": "Boundaries",
598             "Housing": "Housing",
599             "Infrastructure": "Infrastructure",
600             "Neighborhoods": "Boundaries",
601             "Parking": "Planning",
602             "People": "Public Safety",
603             "Permits": "Property",
604             "Police": "Police",
605             "Public Safety": "Public Safety",
606             "Sustainability": "Infrastructure",
607             "Transit": "Transportation",
608             "Urban Planning": "Planning",
609             "Wake County EMS": "Boundaries"
610         }
611     ],
612     [
613         "Miami (Miami Dade)",
614         {
615             "311": "",
616             "Animals": "Transportation",
617             "Building": "Transportation",
618             "Corrections": "Police",
619             "Electoral": "",
620             "Environment": "Environment",
621             "Infrastructure": "Infrastructure",
622             "Procurement": "Education",
623             "Transportation": "Transportation",
624             "Waste": "Property"
625         }
```

```
626     ],
627     [
628         "Virginia Beach",
629         {
630             "Business": "Business",
631             "FOIA": "",
632             "Finance / Government": "Government",
633             "Neighborhoods / Environment": "Environment",
634             "Public Safety": "Public Safety"
635         }
636     ],
637     [
638         "Omaha (Dougla County)",
639         {
640             "Planning": "Planning",
641             "Transportation": "Transportation",
642             "Boundaries": "Boundaries",
643             "Elevation": "Education",
644             "Points of Interest": "Environment"
645         }
646     ],
647     [
648         "Oakland",
649         {
650             "CityGovernment": "",
651             "Economic Develpment": "",
652             "Education": "Education",
653             "Environmental": "",
654             "Financial": "",
655             "Infrastructure": "Infrastructure",
656             "Life Enrichment": "Economic Development",
657             "Property": "Property",
658             "Public Safety": "Public Safety",
659             "Public Services": "Public Safety"
660         }
661     ],
662     [
663         "Minneapolis",
664         {
665             "All Layers": "Transportation",
666             "Public Safety": "Public Safety",
667             "Administrative Boundaries": "Boundaries",
668             "Planning": "Planning",
669             "City Assessor": "City Services",
670             "Regulatory Services": "City Services",
671             "Snow Emergency": "Education",
672             "Business Licensing": "Business",
673             "311": ""
674         }
675     ],
676     [
677         "Tulsa",
678         {}
679     ],
```

```
680     [
681         "New Orleans",
682         {
683             "Arts, Culture, History": "Business",
684             "City Administration": "City Services",
685             "City Finance and Budget": "Finance",
686             "Customer Service": "Education",
687             "Demographics": "Education",
688             "Economy and Workforce": "Economy",
689             "Elections, Politics": "Property",
690             "Environment": "Environment",
691             "Geographic Base Layers": "Transportation",
692             "Health, Education, and Social Services": "Education",
693             "Housing, Land Use, and Blight": "Land Use",
694             "Parks, Parkways": "Parks & Recreation",
695             "Planning, Zoning": "Zoning",
696             "Public Safety and Preparedness": "Public Safety",
697             "Real Estate, Land Records": "Property",
698             "Recreation and Culture": "Education",
699             "Transportation and Infrastructure": "Transportation"
700         }
701     ],
702     [
703         "Wichita",
704         {
705             "Utilities": "Business",
706             "Boundaries": "Boundaries",
707             "Business": "Business",
708             "Community Safety": "Community",
709             "Education": "Education",
710             "Environment and Health": "Health",
711             "Community Services": "Community",
712             "Transportation": "Transportation"
713         }
714     ],
715     [
716         "Tampa",
717         {
718             "Location": "Boundaries",
719             "Boundaries": "Boundaries",
720             "Streets and Addresses": "Transportation",
721             "Planning": "Planning",
722             "Utilities": "Business",
723             "Law Enforcment": "Economy",
724             "Fire": "Education",
725             "Cadastral": ""
726         }
727     ],
728     [
729         "Aurora",
730         {
731             "Business": "Business",
732             "Boundaries": "Boundaries",
733             "Community Services": "Community",
```

```
734             "Demographics": "Education",
735             "Education": "Education",
736             "Planimetrics": "",
737             "Property and Land": "Property",
738             "Water": "Property"
739         }
740     ],
741     [
742         "Honolulu",
743         {
744             "Business": "Business",
745             "Finance": "Finance",
746             "Location": "Boundaries",
747             "Public Safety": "Public Safety",
748             "Recreation": "Education",
749             "Transportation": "Transportation"
750         }
751     ],
752     [
753         "Anaheim",
754         {
755             "Boundaries": "Boundaries",
756             "Base Map Data": "Transportation",
757             "Public Safety": "Public Safety",
758             "Transportation": "Transportation",
759             "Public Works": "Public Safety",
760             "Planning": "Planning",
761             "Education": "Education",
762             "Recreation": "Education",
763             "Environmental": "",
764             "Licenses and Permits": "Property"
765         }
766     ],
767     [
768         "Santa Ana",
769         {}
770     ],
771     [
772         "Riverside",
773         {
774             "Administrative and Fiscal Services": "City Services",
775             "Culture and Recreation": "Education",
776             "Economy and Community": "Community",
777             "Geographic Locations and Boundaries": "Boundaries",
778             "Health and Social Services": "Health",
779             "Land Use and Environment": "Environment",
780             "Public Safety and Emergency Management": "Public Safety",
781             "RCData Internal": "",
782             "RIVCOconnect BroadBand": "",
783             "Transportation": "Transportation"
784         }
785     ],
786     [
787         "Lexington",
```

```
788         {
789             "Historical": "",
790             "Community": "Community",
791             "Development": "Economic Development",
792             "Environmental": "",
793             "Miscellaneous": "",
794             "Transportation": "Transportation"
795         }
796     ],
797     [
798         "St. Louis",
799         {
800             "Transparency": "Transportation",
801             "Neighborhoods": "Boundaries",
802             "Inequality": "Environment",
803             "Environment": "Environment",
804             "Youth": "Transportation",
805             "Citizens As Field Agents": "Boundaries"
806         }
807     ],
808     [
809         "Pittsburgh",
810         {
811             "Art & Culture": "Business",
812             "Business & Economy": "Business",
813             "Civic Vitality & Governance": "Community",
814             "Demographics": "Education",
815             "Education": "Education",
816             "Energy": "Transportation",
817             "Environment": "Environment",
818             "Events": "Education",
819             "Geography": "Education",
820             "Health": "Health",
821             "Housing & Property": "Property",
822             "Human & Social Services": "City Services",
823             "Public Safety & Justice": "Public Safety",
824             "Recreation": "Parks & Recreation",
825             "Transportation": "Transportation",
826             "Other": ""
827         }
828     ],
829     [
830         "Saint Paul",
831         {
832             "Buildings, Housing & Economic Development": "Economic Development",
833             "Census": "Education",
834             "City Administration": "City Services",
835             "City Infrastructure": "Infrastructure",
836             "Permits & Licensing": "Property",
837             "Public Facilities & Services": "Transportation",
838             "Public Safety": "Public Safety"
839         }
840     ],
841     [
```

```
842         "Cincinnati",
843         {
844             "Fiscal Sustainability & Strategic Investment": "Finance",
845             "Growing Economy": "Economy",
846             "Innovative Government": "Government",
847             "Safer Streets": "Transportation",
848             "Thriving Healthy and Neighborhoods": "Boundaries"
849         }
850     ],
851     [
852         "Anchorage",
853         {
854             "Housing and Homelessness": "Public Safety",
855             "Other": "",
856             "Public Health": "Health",
857             "Public Safety": "Public Safety",
858             "Street Maintenance": "Economic Development"
859         }
860     ],
861     [
862         "Henderson",
863         {
864             "Building Permits": "Transportation",
865             "Business License": "Business",
866             "Crime Data": "Education",
867             "Economic / Demographic": "Education",
868             "Parks & Trails": "Parks & Recreation"
869         }
870     ],
871     [
872         "Greensboro",
873         {
874             "Business and Financial": "Business",
875             "Government": "Government",
876             "Housing and Development": "Economic Development",
877             "Public Safety": "Public Safety"
878         }
879     ],
880     [
881         "Plano",
882         {
883             "Business": "Business",
884             "Education": "Education",
885             "Finance": "Finance",
886             "Government": "Government",
887             "Health": "Health"
888         }
889     ],
890     [
891         "Newark",
892         {
893             "Infrastructure": "Infrastructure",
894             "Government": "Government",
895             "Demographics": "Education",
```

```
896             "Business": "Business",
897             "Education": "Education",
898             "Public Safety": "Public Safety"
899         }
900     ],
901     [
902         "Orlando",
903         {
904             "Citywide": "",
905             "Economic Development": "Economic Development",
906             "Government / General": "Government",
907             "Public Safety": "Public Safety",
908             "Public Works": "Public Safety"
909         }
910     ],
911     [
912         "Chula Vista",
913         {
914             "Administrative": "",
915             "Environmental": "",
916             "Facilities": "Transportation",
917             "Land Records": "Property",
918             "Transportation": "Transportation"
919         }
920     ],
921     [
922         "Jersey City",
923         {
924             "Infrastructure": "Infrastructure",
925             "Maps": "Transportation",
926             "Boards and Commisions of H.E.D.C.": "Transportation",
927             "Municipal Services": "City Services",
928             "Housing": "Housing",
929             "Transportation": "Transportation",
930             "Safety": "Public Safety",
931             "Environment": "Environment",
932             "Economy": "Economy",
933             "Other / Unclassified": "",
934             "Demographics": "Education",
935             "Education": "Education"
936         }
937     ],
938     [
939         "Durham",
940         {
941             "Transport, Movements": "Economic Development",
942             "Services, Social": "Education",
943             "Administration, Govenment, Public Finances, Tax Office": "Property",
944             "Economy, Business, SME, Economic Development, Employemnt": "Economy",
945             "Education, Training, Research, Teaching": "Education",
946             "Health": "Health",
947             "Justice, Safety, Police, Crime": "Police",
948             "Spatial Planning, Equipment": "Education",
949             "Sports, Leisure": "Education",
```

```
950             "Environment": "Environment"
951         }
952     ],
953     [
954         "Laredo",
955         {
956             "Natural Resources": "Property",
957             "Boundaries": "Boundaries",
958             "Business": "Business",
959             "Community Safety": "Community",
960             "Education": "Education",
961             "Health": "Health",
962             "Housing": "Housing",
963             "Transportation": "Transportation"
964         }
965     ],
966     [
967         "Madison",
968         {
969             "City Facilities": "Transportation",
970             "Elections": "Finance",
971             "Events": "Education",
972             "Fire": "Education",
973             "Library": "",
974             "Metro Transportation": "Transportation",
975             "Parks & Recreation": "Parks & Recreation",
976             "Police": "Police",
977             "Polling Places": "Boundaries",
978             "Property": "Property",
979             "Public Safety": "Public Safety",
980             "Service Requests": "Education"
981         }
982     ],
983     [
984         "Scottsdale",
985         {
986             "Facilities": "Transportation",
987             "Land Base": "Property",
988             "Planning": "Planning",
989             "Recreation": "Parks & Recreation"
990         }
991     ],
992     [
993         "Glendale",
994         {}
995     ],
996     [
997         "Reno",
998         {}
999     ],
1000    [
1001        "Norfolk",
1002        {
1003            "Arts and Education": "Education",
```

```
1004            "Schools": "Business",
1005            "Elementary School Boundaries": "Boundaries",
1006            "Middle School Boundaries": "Boundaries",
1007            "High School Boundaries": "Boundaries",
1008            "Libraries": "Transportation",
1009            "Electoral": "",
1010            "Electoral Polling Locations": "Boundaries",
1011            "Electoral Precincts": "Boundaries",
1012            "Wards": "Transportation",
1013            "Super Wards": "Transportation",
1014            "Environmental": "",
1015            "Breakwater": "Transportation",
1016            "Category 1 Storm Surge": "Economy",
1017            "Category 2 Storm Surge": "Economy",
1018            "Category 3 Storm Surge": "Economy",
1019            "Category 4 Storm Surge": "Economy",
1020            "Water": "Property",
1021            "Wetlands": "Transportation",
1022            "Municipal": "",
1023            "City Limit": "City Services",
1024            "City Facilities": "Transportation",
1025            "ZIP Code": "Property",
1026            "Neighborhoods": "Boundaries",
1027            "Civic Leagues": "Business",
1028            "Neighborhood Service Areas": "Boundaries",
1029            "Parks and Recreation": "Parks & Recreation",
1030            "Beach Access": "Transportation",
1031            "Boat Ramps": "Transportation",
1032            "Elizabeth River Trail": "Boundaries",
1033            "Parks": "Parks & Recreation",
1034            "Recreation Centers": "Parks & Recreation",
1035            "Planning and Zoning": "Planning",
1036            "AICUZ – Accidental Potential Zones": "Environment",
1037            "AICUZ – Noise Levels": "Infrastructure",
1038            "CBPA": "",
1039            "Conditional Zoning": "Zoning",
1040            "Economic Districts": "Boundaries",
1041            "Flood Zone": "Boundaries",
1042            "Future Land Use": "Land Use",
1043            "HUBZones": "",
1044            "Limit of Moderate Wave Action–LiMWA": "Infrastructure",
1045            "Planning Districts": "Planning",
1046            "Special Exceptions": "Education",
1047            "Zoning": "Zoning",
1048            "Property Information": "Property",
1049            "Addresses": "Property",
1050            "Building Footprints": "Transportation",
1051            "Parcels": "Transportation",
1052            "Public Safety": "Public Safety",
1053            "Fire Demand Zones": "Boundaries",
1054            "Fire Station": "Transportation",
1055            "Police Car Districts": "Police",
1056            "Police Facilities": "Police",
1057            "Police Precincts": "Police",
```

```
1058            "Transportation": "Transportation",
1059            "Light Rail Route": "Transportation",
1060            "Street Centerline": "Transportation"
1061        }
1062    ],
1063    [
1064        "Chesapeake",
1065        {
1066            "Boundaries": "Boundaries",
1067            "Environmental": "",
1068            "Adresses": "",
1069            "Civic": "",
1070            "Transportation": "Transportation",
1071            "Rasters": "Property",
1072            "All Content": "Business"
1073        }
1074    ],
1075    [
1076        "Fremont",
1077        {
1078            "Economic Development": "Economic Development",
1079            "Engineering": "Planning",
1080            "Landmark": "Boundaries",
1081            "Planning": "Planning",
1082            "Transportation": "Transportation",
1083            "All Layers": "Transportation"
1084        }
1085    ],
1086    [
1087        "Baton Rouge",
1088        {
1089            "Business and Financial": "Business",
1090            "Culture and Recreation": "Education",
1091            "Government": "Government",
1092            "Housing and Development": "Economic Development",
1093            "Public Safety": "Public Safety",
1094            "Transportation and Infrastructure": "Transportation"
1095        }
1096    ],
1097    [
1098        "Richmond",
1099        {
1100            "Community Safety and Well-Being": "Health",
1101            "Economic Growth": "Transportation",
1102            "Education and Workforce Development": "Education",
1103            "Sustainability and Natural Environment": "Environment",
1104            "Transportation": "Transportation",
1105            "Unique and Inclusive Neighborhoods": "Boundaries",
1106            "Well-Managed Government": "Government"
1107        }
1108    ],
1109    [
1110        "Boise",
1111        {
```

```
1112            "Parks & Rec": "Parks & Recreation",
1113            "Utilities": "Business",
1114            "Parking": "Planning"
1115        }
1116    ],
1117    [
1118        "San Bernardino",
1119        {
1120            "Boundaries": "Boundaries",
1121            "Demographics": "Education",
1122            "Property": "Property",
1123            "Public Reporting": "Public Safety",
1124            "Transportation": "Transportation",
1125            "Zoning": "Zoning"
1126        }
1127    ],
1128    [
1129        "Spokane",
1130        {
1131            "Boundaries": "Boundaries",
1132            "Demographics": "Education",
1133            "Environment": "Environment",
1134            "Building and Planning": "Education",
1135            "Public Safety": "Public Safety",
1136            "Reference Data": "Economy",
1137            "Public Works": "Public Safety",
1138            "Transportation": "Transportation"
1139        }
1140    ],
1141    [
1142        "Birmingham",
1143        {
1144            "Fire": "Education",
1145            "Finance": "Finance",
1146            "Police": "Police",
1147            "Planning Engineering": "Planning",
1148            "Economic Development": "Economic Development",
1149            "Youth Services": "City Services",
1150            "Municipal Courts": "Community",
1151            "Community Development": "Economic Development",
1152            "Birmingham": "City Services"
1153        }
1154    ],
1155    [
1156        "Tacoma",
1157        {
1158            "Business": "Business",
1159            "City Administration and Finance": "Finance",
1160            "Community and Economic Development": "Economic Development",
1161            "Environment and Sustainability": "Environment",
1162            "Equity": "Property",
1163            "Human Services": "City Services",
1164            "Infrastructure and Transportation": "Transportation",
1165            "Neighborhoods": "Boundaries",
```

```
1166            "Public Safety": "Public Safety"
1167        }
1168    ],
1169    [
1170        "Oxnard",
1171        {
1172            "Archive": "",
1173            "Building Permits": "Transportation",
1174            "Financial Data and Reports": "Economy",
1175            "Local Business Data": "Business"
1176        }
1177    ],
1178    [
1179        "Fayetteville",
1180        {
1181            "All Data Sets": "Economy",
1182            "Business, Budget & Financials": "Business",
1183            "Parks & Recreation": "Parks & Recreation",
1184            "Calls for Service": "Education",
1185            "Planning & Zoning": "Planning",
1186            "Police": "Police",
1187            "Fire": "Education",
1188            "Transportation": "Transportation"
1189        }
1190    ],
1191    [
1192        "Montgomery",
1193        {
1194            "311": "",
1195            "Development": "Economic Development",
1196            "Paving": "Planning",
1197            "Permits": "Property",
1198            "Property Maintenance": "Property",
1199            "Public Safety": "Public Safety",
1200            "Public Works": "Public Safety"
1201        }
1202    ],
1203    [
1204        "Little Rock",
1205        {
1206            "311": "",
1207            "Community Services": "Community",
1208            "Government": "Government",
1209            "Parks": "Parks & Recreation",
1210            "Public Safety": "Public Safety"
1211        }
1212    ],
1213    [
1214        "Akron (County of Summit)",
1215        {
1216            "Land Records": "Property",
1217            "Boundaries": "Boundaries",
1218            "Environmental": "",
1219            "Sewer": "Transportation",
```

```
1220            "Engineer": "Transportation",
1221            "Development": "Economic Development",
1222            "Transportation": "Transportation",
1223            "Topography": "Infrastructure"
1224        }
1225    ],
1226    [
1227        "Grand Rapids",
1228        {}
1229    ],
1230    [
1231        "Salt Lake City",
1232        {
1233            "Administrative": "",
1234            "Atlas Plats": "Transportation",
1235            "Base Map": "Transportation",
1236            "Planning and Zoning": "Planning",
1237            "Parks & Recreation": "Parks & Recreation"
1238        }
1239    ],
1240    [
1241        "Huntsville",
1242        {
1243            "All Open Data": "Boundaries",
1244            "Boundaries": "Boundaries",
1245            "Census Data": "Education",
1246            "Flood": "Transportation",
1247            "Grids": "Education",
1248            "Hydrography": "Education",
1249            "Map Indexes": "Transportation",
1250            "Planning": "Planning",
1251            "Parks and Recreation": "Parks & Recreation",
1252            "Points of Interest": "Environment",
1253            "Public Safety": "Public Safety",
1254            "Schools": "Business",
1255            "Stormwater": "",
1256            "Transportation": "Transportation",
1257            "Miscellaneous Data": "Business"
1258        }
1259    ],
1260    [
1261        "Mobile",
1262        {}
1263    ],
1264    [
1265        "Tallahassee",
1266        {
1267            "All Data Sets": "Business",
1268            "Business & Budget": "Business",
1269            "Infrastructure": "Infrastructure",
1270            "Land Use & Permits": "Land Use",
1271            "Public Safety": "Public Safety",
1272            "Service Requests": "Education",
1273            "Transportation": "Transportation",
```

```
1274            "TLCGis Open Data": "Economy"
1275        }
1276    ],
1277    [
1278        "Knoxville",
1279        {
1280            "311 Performance Measures": "Education",
1281            "Blight Data Dashboard": "Business",
1282            "Budget": "Planning",
1283            "City Council Agenda/Minutes": "City Services",
1284            "Crime Map": "Education",
1285            "East Tennessee Index": "Property",
1286            "Homemaker Program": "Parks & Recreation",
1287            "KGIS Maps": "Transportation",
1288            "KnoxHMIS": "",
1289            "NPDES Reports": "Property",
1290            "Police Advisory and Review Committee": "Police",
1291            "Police Department": "Police",
1292            "Property Tax Database - City": "Property",
1293            "Property Tax Database - County": "Property",
1294            "Public Improvement Projects": "Education",
1295            "Rainfall Data": "Economy",
1296            "TIFs and PILOTs": "Transportation",
1297            "Tree Inventory": "Property"
1298        }
1299    ],
1300    [
1301        "Worcester",
1302        {}
1303    ],
1304    [
1305        "Tempe",
1306        {
1307            "Accessibility": "Environment",
1308            "Transportation": "Transportation",
1309            "Neighborhoods & Historic Preservation": "Education",
1310            "Land Use": "Land Use",
1311            "Public Safety": "Public Safety",
1312            "All Datasets": ""
1313        }
1314    ],
1315    [
1316        "Santa Clarita",
1317        {
1318            "Annual Budget": "Property",
1319            "Employee Effectiveness": "Environment",
1320            "Time To Respond": "Education",
1321            "Employee Courtesy": "Economic Development",
1322            "Expectations Met": "Education",
1323            "Election Data at a Glance": "Education",
1324            "Election Data by Precinct": "Finance",
1325            "Election Results by Candidate": "Education",
1326            "Film Permits Issued & Days Of Filming": "Property",
1327            "Film Revenue": "Property",
```

```
1328            "Hotel Occupancy": "Education",
1329            "Number of Businesses": "Business",
1330            "Number of Jobs": "Education",
1331            "Unemployment Rate": "Environment",
1332            "Library Circulation": "Education",
1333            "Transit Ridership": "Transportation",
1334            "Graffiti Cleanup": "Property",
1335            "DFYinSCV Membership": "Business",
1336            "LMD Water Usage": "Land Use"
1337        }
1338    ],
1339    [
1340        "Cape Coral",
1341        {
1342            "311 Call Center": "Boundaries",
1343            "Business & Budget": "Business",
1344            "Planning & Zoning": "Planning",
1345            "Public Works": "Public Safety",
1346            "Gis": "Property",
1347            "Fire": "Education",
1348            "Brownse All Data": "Economy"
1349        }
1350    ],
1351    [
1352        "Providence",
1353        {
1354            "Economy / Finance": "Finance",
1355            "Neighborhoods": "Boundaries",
1356            "Public Safety": "Public Safety",
1357            "Reference": "Property"
1358        }
1359    ],
1360    [
1361        "Chattanooga",
1362        {
1363            "Buildings & Trails": "Transportation",
1364            "Economy": "Economy",
1365            "Education": "Education",
1366            "Government": "Government",
1367            "Public Health": "Health",
1368            "Public Safety": "Public Safety",
1369            "Recreation": "Parks & Recreation",
1370            "Social Services": "Education",
1371            "Transportation": "Transportation"
1372        }
1373    ],
1374    [
1375        "Santa Rosa",
1376        {
1377            "Community Services": "Community",
1378            "Development": "Economic Development",
1379            "Economy": "Economy",
1380            "Finances": "Property",
1381            "Fire": "Education",
```

```
1382            "Government": "Government",
1383            "HCS Department": "Police",
1384            "HR Department": "Police",
1385            "Housing": "Housing",
1386            "IT Department": "Police",
1387            "Internal Datasets": "",
1388            "Office of Community Engagement": "Community",
1389            "Police": "Police",
1390            "Public Safety": "Public Safety",
1391            "Recreation and Culture": "Education",
1392            "Transportation": "Transportation",
1393            "Water Department": "Police"
1394        }
1395    ],
1396    [
1397        "Sioux Falls",
1398        {
1399            "Community": "Community",
1400            "Education": "Education",
1401            "Imagery": "Education",
1402            "Infrastructure": "Infrastructure",
1403            "Parks": "Parks & Recreation",
1404            "Property": "Property",
1405            "Safety": "Public Safety",
1406            "Transportation": "Transportation"
1407        }
1408    ],
1409    [
1410        "McKinney",
1411        {
1412            "Planning & Zoning": "Planning",
1413            "Addresses & Boundaries": "Boundaries",
1414            "Schools": "Business",
1415            "Engineering": "Planning",
1416            "Public Safety": "Public Safety",
1417            "Parks & Recreation": "Parks & Recreation",
1418            "Transportation": "Transportation",
1419            "All Other Data": "Business"
1420        }
1421    ],
1422    [
1423        "ElkGrove",
1424        {
1425            "Base Layers": "Transportation",
1426            "Building Footprints": "Transportation",
1427            "City Limits": "City Services",
1428            "City Council Boundary": "Boundaries",
1429            "Fire Stations": "Education",
1430            "CSD Parks": "Parks & Recreation",
1431            "EGUSD Schools": "Business",
1432            "Parking Lot Centerlines": "Environment",
1433            "Public Amenity Footprints": "Public Safety",
1434            "Land Layers": "Property",
1435            "General Plan": "Parks & Recreation",
```

```
1436            "Home Owners Associations": "Business",
1437            "Neighborhoods Associations": "Business",
1438            "Parcels": "Transportation",
1439            "Sphere of Influence": "Environment",
1440            "Southeast Planning Area": "Education",
1441            "Zoning": "Zoning",
1442            "Drainage Layers": "Economic Development",
1443            "Creeks and Channels": "Education",
1444            "Detention Basins": "Transportation",
1445            "Drainage Features": "Economic Development",
1446            "Transportation Layers": "Transportation",
1447            "eTran Bus Routes": "Boundaries",
1448            "eTran Bus Stops": "Education",
1449            "Bikeways": "",
1450            "Park and Ride Locations": "Boundaries",
1451            "Rail Roads": "Transportation",
1452            "Street Centerlines": "Transportation",
1453            "Curb Ramps": "Transportation",
1454            "Sidewalks": "Transportation",
1455            "Infrastructure / Asset Layers": "Infrastructure",
1456            "Street Lights": "Transportation",
1457            "Landscape Areas": "Boundaries",
1458            "Data Gateway Layers": "Transportation",
1459            "Building Permits": "Transportation",
1460            "Business Licenses": "Business",
1461            "Capital Improvement Projects": "Education"
1462        }
1463    ]
1464 ]
```

APÊNDICE G – PLANILHA DE AVALIAÇÃO DO PROCESSO DE ALINHAMENTO DE CATEGORIAS

Neste Apêndice é mostrada a planilha criada para se produzir a avaliação do processo de Alinhamento de Categorias, descrito no Capítulo 3.

Pesquisa para avaliação de método de alinhamento de categorias

Somos um grupo de pesquisa do Instituto de Computação da Universidade Federal Fluminense interessado em estudos relacionados a portais de dados abertos. Diversas cidades ao redor do mundo vem disponibilizando dados em portais na web. Muitas das vezes, as cidades agrupam os datasets nesses portais através de categorias. No entanto, cada cidade apresenta seu próprio conjunto de categorias, o que pode dificultar a busca de datasets em diferentes portais.

Esta é uma pesquisa que tem como objetivo avaliar um método de categorização automática para conjuntos de dados em portais de dados urbanos. O método é baseado no alinhamento semântico entre categorias diferentes.

Nessa planilha de pesquisa existem 3 tabelas para alinhamento de categorias. Para cada conjunto de itens apresentados nas linhas 18, 43 e 68, você deve escolher (marcar com a letra X) entre as categorias oferecidas nas linhas subsequentes a que considerar que seja mais similar em termos semânticos. Não existe resposta correta, apenas escolha a opção que considerar que seja mais semelhante ao item apresentado. Antes de começar, gostaríamos que você respondesse algumas poucas questões na tabela abaixo.

Pergunta	Sua Resposta
Você concorda em participar dessa pesquisa? (Sim ou Não)	
Numa escala de 1 a 5, como você avalia sua fluência na língua inglesa ?	
Qual a sua idade ?	
Qual o seu nível de escolaridade ? (Fundamental, Médio, Superior ou Pós-Graduação)	
Se possui nível superior, qual o curso ?	

Para cada item apresentado nas colunas da linha 18, você deve escolher (marcar com a letra X) entre as categorias oferecidas nas linhas subsequentes a que considerar que seja mais similar em termos semânticos. Por exemplo, para o item na célula B18, você deve marcar com a letra X em apenas uma das linhas entre B19 e B39, de acordo com as categorias apresentadas na coluna A. Dessa mesma forma deve-se proceder para os outros itens da linha 18, as colunas C18 a K18.

	City Government	Housing & Development	Administration & Finance	Facilities & Geo. Boundaries	Government Boundaries	Planning & Development	City Infrastructure	Energy & Environment	Budget and Finance	Community Development
Public Safety										
Transportation										
City Services										
Parks & Recreation										
Economic Development										
Planning										
Business										
Government										
Education										
Health										
Boundaries										
Community										
Environment										
Finance										
Housing										
Infrastructure										
Property										
Police										
Land Base										
Economy										
Zoning										

Para cada item apresentado nas colunas da linha 43, você deve escolher (marcar com a letra X) entre as categorias oferecidas nas linhas subsequentes a que considerar que seja mais similar em termos semânticos. Por exemplo, para o item na célula B43, você deve marcar com a letra X em apenas uma das linhas entre B44 e B64, de acordo com as categorias apresentadas na coluna A. Dessa mesma forma deve-se proceder para os outros itens da linha 43, as colunas C43 a K43.

	Sanitation	A Livable and Sustainable City	A Safe City	Service Requests	Permitting and Licensing	Flood Hazard	Staff Salaries	Auditor	Public Works	City Management and Ethics
Public Safety										
Transportation										
City Services										
Parks & Recreation										
Economic Development										
Planning										
Business										
Government										
Education										
Health										
Boundaries										
Community										
Environment										
Finance										
Housing										
Infrastructure										
Property										
Police										
Land Base										
Economy										
Zoning										

Para cada item apresentado nas colunas da linha 68, você deve escolher (marcar com a letra X) entre as categorias oferecidas nas linhas subsequentes a que considerar que seja mais similar em termos semânticos. Por exemplo, para o item na célula B68, você deve marcar com a letra X em apenas uma das linhas entre B69 e B89, de acordo com as categorias apresentadas na coluna A. Dessa mesma forma deve-se proceder para os outros itens da linha 68, as colunas C68 a K68.

	Planning Building and Code Enforcement	Map Features	Culture & Arts	Recreation & Culture	Area Plans	Code Interpretations	Peer Review	Scope Statement	Housing, Land Use, and Blight	Recreation and Culture
Public Safety										
Transportation										
City Services										
Parks & Recreation										
Economic Development										
Planning										
Business										
Government										
Education										
Health										
Boundaries										
Community										
Environment										
Finance										
Housing										
Infrastructure										
Property										
Police										
Land Base										
Economy										
Zoning										