UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MATHEUS YUICHI YAMASHIRO

FERRAMENTAS DE COLETA E ANÁLISE DE DADOS DE LICITAÇÕES PÚBLICAS

BAURU
Agosto/2022

MATHEUS YUICHI YAMASHIRO

FERRAMENTAS DE COLETA E ANÁLISE DE DADOS DE LICITAÇÕES PÚBLICAS

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Câmpus Bauru.
Orientador: Prof. Dr. Kelton Augusto Pontara da Costa
Coorientador: Prof. Me. Miguel José das Neves

BAURU
Agosto/2022

Y19f

Yamashiro, Matheus Yuichi
Ferramentas de coleta e análise de dados de licitações públicas / Matheus Yuichi Yamashiro. -- Bauru, 2022
36 f. : il.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências, Bauru
Orientador: Kelton Augusto Pontara da Costa
Coorientador: Miguel José das Neves

1. Processamento de linguagem natural (Computação). 2. Sites da Web. 3. Licitação Pública. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências, Bauru. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Matheus Yuichi Yamashiro

Ferramentas de Coleta e Análise de Dados de Licitações Públicas

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Câmpus Bauru.

Banca Examinadora

Prof. Dr. Kelton Augusto Pontara da Costa
Orientador
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Profª. Drª. Simone das Graças Domingues Prado
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Profª. Drª. Márcia Aparecida Zanoli Meira e Silva
Departamento de Computação
Faculdade de Ciências
Universidade Estadual Paulista "Júlio de Mesquita Filho"

Bauru, _____ de _____ de _____.

A meus pais, amigos, professores,
e todos aqueles que lutaram e ainda lutam pela educação pública

Agradecimentos

Agradeço inicialmente meus pais, por me apoiarem sempre que necessitei. Sempre com empenho para me ajudar o máximo que podem, da maneira que podem. Até além.

Agradeço também aos meus professores e amigos, do primário, do médio, do superior, e do extracurricular, por sempre me encorajar a encarar novos desafios e crescer como um acadêmico, um profissional, e, em muitas vezes, como pessoa.

Agradeço aos médicos que trataram de mim, e daqueles que tratam a todos que necessitam. Sejam nos males físicos, sejam nos males psicológicos. Seus esforços tornam as pessoas mais vivas e saudáveis para viver.

Agradeço à todos aqueles que lutaram e ainda lutam pela ciência e pela educação. Sem seus esforços, não teríamos a saúde que temos hoje. Não teríamos a tecnologia que temos hoje. Não teríamos o conforto que temos hoje. Não teríamos o que temos hoje.

Forneça opções, não dê desculpas esfarrapadas.

Andrew Hunt & David Thomas

Resumo

O mundo conectado atual vêm causando grandes mudanças no dia-a-dia das pessoas. Uma delas é a quantidade de informações consumidas e produzidas. Isso vêm crescendo de maneira tão alarmante que já não é mais possível analisar essas informações de maneira manual. É necessário automatizá-las. Contudo, a maneira que as pessoas pensam e se comunicam, seja pessoalmente ou por meios informatizados (pela internet), é fundamentalmente diferente da maneira que os computadores o fazem. Assim, é necessário gerar métodos e interfaces para que os computadores, capazes de processar um volume de dados muito maior que os humanos, entendam e utilizem os dados gerados pelas pessoas. A área de pesquisa de Processamento de Linguagem Natural (PLN) é justamente isso, desenvolver técnicas para que os computadores processem dados textuais geradas por-humanos-para-humanos, de maneira automática. Neste trabalho, foi realizado um estudo sobre esta área de pesquisa, algumas técnicas implementadas e aplicadas em dados de licitações públicas da Prefeitura de Bauru, e desenvolvida uma interface *web* para a visualização dos resultados obtidos.

Palavras-chave: Processamento de Linguagem Natural; Desenvolvimento *web*; Licitações Públicas.

Abstract

The connected world of today's have been causing great changes on the day-to-day lives of people. One of these is the amount of information being consumed and generated. This has been increasing to the point that it is not possible to analyse it manually anymore. It has to be automated. But the way humans think and communicate, be it in person or through the internet, is fundamentally different from how computers do it. Therefore, there is a need to develop methods and interfaces for computers, those of which are capaple of processing much more data than humans, to understand and use data generated by humans. The area of research of Natural Language Processing (NLP) is just that, developing techniques for computers process text data generated by-humans-for-humans, automatically. In this piece of work, a study of this area of research has been made, some of the techniques implemented and applied to Bauru's Prefecture public bidding data, and a web interface developed to visualize the data obtained.

Keywords: Natural Language Processing; Web Development; Public Bidding.

Lista de figuras

Lista de abreviaturas e siglas

API	*Application Programming Interface*
BoW	*Bag-of-Words*
DL	*Deep Learning*
DO	Diário Oficial
DOs	Diários Oficiais
HTML	*HyperText Markup Language*
JSON	*JavaScript Object Notation*
LCS	*Longest Common Substring*
ML	*Machine Learning*
PDF	*Portable Document File*
PLN	Processamento de Linguagem Natural
PCNP	Portal Nacional de Contratações Públicas
RegEx	*Regular Expression*
SQL	*Structured Query Language*
TF.IDF	*Term-Frequency times Inverse-Document-Frequency*
XHTML	*eXtended HyperText Markup Language*

Sumário

1 Introdução

Assim como toda organização, a aquisição e alienação de bens é necessária para o seu crescimento, desenvolvimento, e manutenção. No ambiente privado não existe uma "maneira correta" de se fazer isso. Muitas vezes fica a critério da própria empresa definir seus processos. Já no caso do setor público, a Constituição Federal, no Art. 37 inciso XXI, exige um processo licitatório para a aquisição de produtos, obras, e serviços, também chamados de "objetos" da licitação, de modo que "assegure igualdade de condições a todos os concorrentes", obedecendo "aos princípios da legalidade, impessoalidade, moralidade, publicidade e eficiência" (BRASIL, 1988).

Essa exigência é regulamentada pela Lei Federal nº 8.666 (BRASIL, 1993), abrangendo todos os níveis da administração pública (Federal, Estadual, Municipal). Nesta, o Art. 22 regulamenta as cinco modalidades de licitação: concorrência, tomada de preços, convite, concurso, e leilão. Além desses, há também a modalidade de pregão, regulamentada pela Lei Federal nº 10.520 (BRASIL, 2002). O escopo do presente trabalho é limitado às aquisições de bens por parte do poder público. Ou seja, são tratados apenas as modalidades de concorrência, tomada de preços, convite, e pregão, não sendo tratados contratação serviços ou alienações.

A divulgação das licitações para interessados é feita por meio de editais, que, pelo Art. 21 da Lei 8.666, devem possuir ao menos um resumo publicado no respectivo Diário Oficial (DO). Este resumo deverá também conter informações para a obtenção do texto integral.

Tanto o DO quanto os Editais são escritos utilizando a linguagem natural, que, assim como o nome diz, é utilizada naturalmente pelos humanos em geral. Essa língua também é chamada de "não estruturada", o que gera ambiguidades, dificultando sua organização e leitura automatizada. Ou seja, a análise computadorizada desses textos não é um processo trivial.

Com isso em mente, uma nova área de pesquisa surgiu para analisar a escrita e a fala humana, chamada Processamento de Linguagem Natural (PLN, ou NLP, do inglês *Natural Language Processing*). Jurafsky e Martin (2021) resumem as competências necessárias para essa tarefa: fonética (o som), morfologia (os componentes), sintaxe (a estrutura), semântica (os significados), pragmática (a intenção), e discurso (o todo).

Historicamente, os algoritmos utilizados consistiam de implementações de regras de linguagem e suas manipulações. Já atualmente se observa uma grande presença de aprendizagem de máquina (ML, do inglês *machine learning*) e aprendizagem profunda (DL, do inglês *deep learning*) (RODRIGUES, 2020).

Esta área de pesquisa têm ganhado interesse pois ela permite o processamento automático de dados gerados por pessoas, como fala e postagens em redes sociais. Isso permite uma

melhor interação humano-computador, além de ser capaz de digerir uma quantidade massiva de informações. Além disso, diversos outros textos, como textos acadêmicos e documentos pessoais, podem ser processados para auxiliar na catalogação e indexação dessas informações, agilizando a busca e o acesso à elas.

1.1 Problema

Para comprar produtos ou contratar serviços, as entidades públicas publicam esse pedido em forma de licitações nos DO competentes. Essas licitações buscam agregar interessados em providenciar o serviço ou produto para que a gestão pública escolha com quem efetuar o negócio.

Assim, as empresas interessadas em participar do processo licitatório devem estar atentas às publicações de licitações nos DO. Contudo, esses cadernos podem chegar às centenas de páginas, publicados quase diariamente, dificultando o processo de busca e gerenciamento por parte da empresa.

Pela Lei nº 12.527 de 18 de Novembro de 2011 (BRASIL, 2011), as entidades administrativas devem disponibilizar esses dados de maneira legível por máquina. Contudo, a exploração inicial revelou que nem todas as entidades administrativas disponibilizavam esses dados, ou não os referenciava em seu *site*. Em outros casos os disponibilizavam através de uma interface *web*, sendo necessário a extração para outro formato mais enxuto. Mas todas tinham seus DO devidamente publicados e catalogados, apesar de estarem em linguagem natural, e não em formato legível por máquina. Este então sempre pode ser considerado como uma das fontes de dados, mesmo que de mais difícil extração.

1.2 Justificativa

As empresas interessadas em participar desse processo devem estar atentas para as licitações publicadas em cada dia, decidir de quais participar, e se manter competitiva diante os outros interessados.

Devido ao tamanho, a frequência, e a linguagem natural dos DO, torna-se difícil coletar os dados das licitações manualmente. Quando se deseja participar de licitações em mais de uma unidade administrativa, essa tarefa é multiplicada, não só pelo volume, mas também pelas diferenças de *layouts* entre os documentos.

Assim justifica-se a pesquisa de metodologias que busquem os dados referentes às licitações de maneira automatizada a partir de diversas fontes de modo a facilitar o gerenciamento e a participação da empresa nessas concorrências.

1.3 Objetivos

Nesta seção será apresentado os objetivos gerais e objetivos específicos do trabalho desenvolvido.

1.3.1 Objetivo Geral

O objetivo geral deste trabalho é agregar dados de licitações do Município de Bauru, estruturando-as de modo a disponibilizar um *dataset*.

1.3.2 Objetivos Específicos

Para atender o objetivo geral deste trabalho, e acompanhar seu progresso, os objetivos específicos apresentados a seguir foram elaborados:

- Estudar o processo licitatório;
- Estudar ferramentas de raspagem de dados;
- Desenvolver ferramentas para efetuar a raspagem e estruturação dos dados automaticamente;
- Desenvolver uma interface *web* para a interação com o usuário.

2 Fundamentação Teórica

Nesta seção serão apresentados os métodos utilizados no desenvolvimento deste trabalho.

2.1 Processamento de Linguagem Natural

O volume de dados gerados e disponibilizados na *internet* vêm crescendo de maneira alarmante, ao ponto de tornar impossível processá-los manualmente (RINO; PARDO, 2003): é necessário sistemas automáticos. Esses sistemas requerem métodos e algoritmos para que possam consumir os dados e transformá-los em informação e conhecimento. Para alguns casos, métodos já existem, como no processamento de valores numéricos, mas em outros não, como é o caso da linguagem natural.

As pessoas utilizam palavras para se comunicarem entre si, tanto diretamente quanto através de computadores. Apesar de já estarem em um computador, seu propósito lá não é para que o computador entenda esses dados, mas que outra pessoa acesse o arquivo e interprete o seu conteúdo. A tarefa do computador é apenas decodificar e exibir o que foi codificado.

Os dados armazenados pelo computador são dados de baixo nível, enquanto a interpretação humana possui um nível maior de abstração. Para um computador, todas as letras em um arquivo de texto, por exemplo, são basicamente iguais: valores entre 0 e 255, e a interpretação de cada um desses valores sempre será igual, ou seja, o *byte* `0x41` (ou 65 decimal) sempre será o caractere "`A`" maiúsculo, seja ele na palavra "LINGUAGEM", quanto na palavra "NATURAL". Analogamente, as palavras possuem significados diferentes, apesar de serem compostas por letras iguais. Fica a cargo da pessoa interpretar a cadeia de caracteres como uma palavra e a atribuir algum significado, ao invés de apenas uma sequência de *bytes*.

A área de Processamento de Linguagem Natural (PLN) visa tratar essa lacuna na interação humano-computador. Assim, o computador pode entender o que o humano diz e gerar textos (e até fala) de forma autônoma e que o humano consida entender (RODRIGUES, 2020).

2.2 Representação do Texto

Há diversas maneiras de representar um texto em um sistema informatizado. Contudo, dependendo da tarefa, algumas representações podem ser mais adequadas que outras. Por exemplo: um arquivo de configuração deve estar armazenado de uma maneira que o computador possa ler rápidamente e com o mínimo de manipulações possíveis. Já um texto altamente estilizado, como um pôster publicitário, pode ser mais adequado armazená-lo como uma imagem

ao invés de caracteres individuais.

A seguir algumas representações utilizadas neste trabalho são apresentadas.

2.2.1 Cadeia de caracteres

A cadeia de caracteres, também chamada de *string*, é a forma mais simples de armazenar texto. Muitas vezes são representadas por vetores de caracteres *ASCII*, ou *Unicode*. Arquivos que contém apenas esses caracteres são chamados de "texto-pleno" ou "texto-simples".

Por exemplo, a palavra "COMPUTADOR" é representada pelo vetor ['C', 'O', 'M', 'P', 'U', 'T', 'A', 'D', 'O', 'R'], em que não existe diferenciação entre caracteres iguais em posições diferentes. Ou seja, os caracteres 'O' da palavra "COMPUTADOR" são indistinguíveis um do outro, além da sua posição dentro da palavra ou da cadeia de caracteres.

Por causa dessa característica, sua utilidade na área de processamento de dados é limitada, muitas vezes não ultrapassando do armazenamento e transferência de dados. As informações implícitas dentro do texto-pleno deve ser lida e interpretada pelo *software* para então ser processada e manipulada.

2.2.2 *Bag-of-Words* (BoW)

Para tarefas de PLN, um passo além do texto-pleno é o uso de vetores chamados *Bag-of-Words* (BoW), traduzido literalmente em "sacola-de-palavras". Neste, a menor unidade de informação é a palavra, e não o caractere, como na cadeia de caracteres. Ou seja, as representações das palavras "`TORTA`" e "`BLUSA`" são diferentes no BoW, enquanto que no texto pleno não há diferenciação entre os "`A`" das duas palavras.

Por exemplo, a frase "A COMPUTAÇÃO E A MATEMÁTICA" pode ser representada pela sequência de vetores:

- $[1, 0, 0, 0]'$ (A)
- $[0, 1, 0, 0]'$ (COMPUTAÇÃO)
- $[0, 0, 1, 0]'$ (E)
- $[1, 0, 0, 0]'$ (A)
- $[0, 0, 0, 1]'$ (MATEMÁTICA).

Note que as palavras "A" são representadas pelo mesmo vetor, mas não têm relação com os "A"s das demais palavras. Somando os vetores, a frase inteira pode ser representada por um único vetor $[2, 1, 1, 1]$. Contudo, dessa forma perde-se as informações quanto à ordem das palavras na frase (MASSONI, 2021).

Adicionalmente, pode-se ignorar palavras como artigos e preposições, também chamadas de *stop-words*. Isso é feito para reduzir o tamanho dos vetores, já que estas palavras são tão comuns que não trazem informações novas ao texto. Dessa maneira, a frase anterior poderia ser resumida aos vetores $[1, 0]'$ (COMPUTAÇÃO) e $[0, 1]'$ (MATEMÁTICA).

2.2.3 *n-grams*

Uma extensão da representação BoW é utilizar mais de uma única palavra. Dessa forma, se for utilizado palavras adjascente, a natureza sequencial das palavras pode ser melhor preservada (MASSONI, 2021). Esse método é conhecido como *n-grams*.

Neste, ao invés de armazenar apenas *tokens* singulares, armazena-se agrupamentos com n elementos sequenciais. Por exemplo, a frase "A COMPUTAÇÃO E A MATEMÁTICA" pode ser dividida nos bigramas: (A, COMPUTAÇÃO); (COMPUTAÇÃO, E); (E, A); e (A, MATEMÁTICA), e nos trigramas (A, COMPUTAÇÃO, E); (COMPUTAÇÃO, E, A); e (E, A, MATEMÁTICA).

Aqui foi discutido o agrupamento de n palavras (ou *tokens*), mas também pode ser feito com diferentes granularidades, como caracteres ou frases (MARTINS, 2020).

2.2.4 *Word-embeddings*

Apesar de trazer benefícios no processamento de texto, a abordagem do BoW em si não traz informações sobre o significado das palavras, nem da relação entre elas. Assim, as relações entre as palavras, como sinônimos e antônimos, são perdidas.

Para remediar esse problema, começou-se a representar as palavras como pontos no espaço, partindo do princípio que "palavras em contextos similares aparecem e tendem a ter significados similares" (JURAFSKY; MARTIN, 2021).

Mais recentemente, pontos com muitas dimensões, atribuindo valores automáticamente por um modelo treinado por aprendizado de máquina são utilizados. Esta abordagem é chamada de *word-embeddings*.

Um exemplo de *word-embedding* é o *word2vec*, que utiliza um *corpus* para obter a relação entre palavras próximas e atribuir características a cada palavra (BENGIO et al., 2003).

2.3 Métricas de Comparação

Para a validação dos dados obtidos, é necessário comparar documentos. Ou seja, são necessárias métricas para quantificar cada documento para então comparar essas métricas. Nesta seção, algumas métricas utilizadas neste trabalho são apresentadas.

2.3.1 Similaridade de coincidência

A maneira mais direta de responder a pergunta "duas frases são iguais?", ou em outras palavras, "duas cadeias de caracteres são 100% similares?", é verificar se todos os caracteres coincidem, ou seja, são iguais, comparando uma a uma até o final da cadeia, incluindo o próprio final. Contudo, nem sempre apenas a resposta "Não são iguais" é suficiente. Pode ser necessário uma métrica para quantificar o quão duas frases são iguais.

Para tal existem várias abordagens. Uma delas é verificar qual é o maior comprimento de caracteres coincidentes (ou sub-cadeia mais comprida, do inglês *Longest Common SubString* (LCS)). Outra é a quantidade de modificações (inserção e remoção de caracteres) necessárias para transformar uma cadeia em outra.

Vale notar que algumas dessas abordagens podem ser utilizadas para comparar outras representações como vetores ao invés de apenas caracteres.

2.3.2 Similaridade de Cosseno

Quando a representação de documentos é feita por vetores (como no BoW e no *word-embeddings*), pode-se utilizar o ângulo formado pelos vetores das palavras. Assim, quando os vetores coincidirem, ou seja, as frases forem representadas pelo mesmo vetor, o ângulo formado será nulo e o cosseno será 1. No outro extremo, quando os vetores forem perpendiculares o cosseno será 0.

Contudo, esta abordagem verifica apenas o ângulo dos vetores, não suas magnitudes. Isso deve ser levado em consideração durante o processo de escolha das métricas utilizadas.

É importante notar que vetores coincidentes não necessariamente implicam em documentos coincidentes. Por exemplo, as frases "GAVETA DE MESA" e "MESA DE GAVETA" são representadas no BoW pelo mesmo vetor $[1, 1]$, apesar de terem significados diferentes.

A similaridade de cosseno entre dois vetores $\mathbf{A} = [a_1 \dots a_n]^t$ e $\mathbf{B} = [b_1 \dots b_n]^t$ de dimensão de n características (no caso, palavras), é dada por

$$\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

2.3.3 Similaridade de Jaccard

Outra maneira de lidar com os BoWs é interpretá-las como elementos em conjuntos. Dessa maneira, cada BoW é um conjunto e cada palavra é um elemento com uma dada frequência. Assim, uma métrica utilizada é a similaridade de Jaccard, ou índice de Jaccard.

Essa métrica é a razão entre os elementos em ambos conjuntos por todos os elementos. Ou seja, sua interseção pela união (daí seu outro nome). Então, dados dois conjuntos (novamente, os BoWs) $\mathbf{A}$ e $\mathbf{B}$, a similaridade Jaccard é dada por

$$\frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A} \cup \mathbf{B}|} = \frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A}| + |\mathbf{B}| - |\mathbf{A} \cap \mathbf{B}|}$$

2.3.4 TF.IDF

Uma aplicação do processamento de texto é determinar o principal conteúdo de um documento. Ou seja, determinar suas palavras-chave. A resposta mais simples seria a frequência da palavra no documento, contudo, palavras muito comuns podem ser tão comuns que trazem pouca informação sobre o texto.

Com isso em mente, a métrica TF.IDF visa mitigar esse problema. Sua sigla é um acrônimo para *Term Frequency times Inverse Document Frequency*, e calcula o grau de importância de um termo para um documento dentro de um *Corpus*. Resumidamente, se um termo aparece em poucos documentos, ele pode ter destaque dentro dos documentos em que ele aparece (RAJARAMAN; ULLMAN, 2011). Por outro lado, se ele aparece em muitos dos documentos, não traz uma quantidade significativa de novas informações (por exemplo, artigos e preposições).

Assim, seja n_i um documento i num *corpus* N, f_{ij} a frequência do termo j no documento i, e q_i a quantidade de documentos de N em que o termo j aparece, a importância TF.IDF do termo j para o documento i no *corpus* N é dada por

$$TF \times IDF, \text{ em que } TF = \frac{f_{ij}}{\max_k f_{ik}} \text{ ou } \frac{f_{ij}}{|n_i|}, \text{ e } IDF = \log_2 \left(\frac{|N|}{q_i} \right)$$

2.4 Sumarização de textos

Uma das tarefas da área de PLN é a sumarização de textos. Ela consiste em determinar as informações mais importantes de um documento para que ele possa ser reduzido de modo que seu conteúdo seja o menos afetado possível.

Há diversas maneiras de se classificar um método de sumarização de textos, contudo, para este trabalho, o mais importante a ser considerado é o volume de informações extraídas. Rino e Pardo (2003) coloca que um sumário pode ser indicativo ou informativo:

Indicativo é aquele que "transmite somente uma ideia vaga"; e

Informativo é aquele que "contém todos os seus aspectos principais, dispensando a leitura [do original]".

3 Metodologia

Nesta seção são apresentados os métodos e ferramentas utilizados pelo projeto desenvolvido. O código desenvolvido pode ser encontrado no repositório <https://github.com/yamashirl/tcc-pln-2022>

3.1 Ferramentas utilizadas

Nesta seção são apresentadas tanto as ferramentas externas utilizadas durante a pesquisa, quanto as utilizadas na ferramenta final.

3.1.1 *Python* e bibliotecas padrão

O *Python* é uma linguagem de programação muito reconhecida por ser de fácil aprendizado e ter grande suporte de comunidade. Devido à sua popularidade, muitas bibliotecas são desenvolvidas e mantidas por voluntários e por organizações independentes para diversas finalidades.

Neste trabalho, o *Python* e seu pacote de bibliotecas padrão foi utilizado para toda a lógica de manipulação dos dados e foi escolhido devido à facilidade de uso e ampla gama de bibliotecas disponíveis. Particularmente as bibliotecas "`re`" e "`requests`" foram frequentemente utilizadas neste trabalho para a busca por *RegEx* e aquisição de páginas e documentos na *web*, respectivamente.

3.1.2 *Jupyter*

O *Jupyter* é uma plataforma interativa de desenvolvimento que permite embutir códigos executáveis à textos explicativos.

Neste trabalho, o *Jupyter* foi utilizado devido à facilidade de uso e interatividade para observar e validar a execução dos algoritmos desenvolvidos.

3.1.3 *spaCy*

O *spaCy* é uma biblioteca de processamento de linguagem natural preparada para várias tarefas como *tokenização*, marcação de partes-de-fala (ou em inglês *part-of-speech* (POS)), lematização, cálculo de similaridade semântica, entre outros. Ele pode ser carregado com modelos pré-treinados de diferentes linguagens, além de ser treinado para satisfazer uma aplicação específica.

Neste projeto foi utilizado o modelo pré-treinado para a línugua portuguesa "`pt_core_news_lg`".

3.1.4 *Apache Tika*^{TM}

O *Apache Tika* ^{TM} é um pacote de ferramentas desenvolvido para "detectar e extrair texto e metadados de mais de mil tipos de arquivos diferentes (como PPT, XLS, e PDF)" (Apache Tika, 2022). Sua tarefa principal é simplificar a extração de dados de diferentes tipos de arquivos a partir de uma interface única.

Neste projeto, a ferramenta *Parser* do pacote foi utilizado para extrair os dados dos Diários Oficiais de Bauru, que são disponibilizados em formato PDF, e extraídos em formato XHTML.

3.1.5 *BeautifulSoup*

A biblioteca *BeautifulSoup* foi desenvolvida para "extrair dados de arquivos HTML e XML" (BeautifulSoup, 2022). A partir dos rótulos de hiper-texto, conhecidos como *tags*, que são identificados pelos símbolos de "<" e ">", ela recria a hierarquia do arquivo utilizando objetos e estruturas nativas do *Python*, simplificando o processo de manuseio desse tipo de arquivo.

3.1.6 *MySQL*

Para armazenar e recuperar dados rapidamente, foi utilizado um banco de dados *MySQL*. Este conta com diversos mecanismos para agilizar a pesquisa e obtenção dos dados, se comparado com uma abordagem baseada em arquivos.

3.1.7 *Django*

Para o desenvolvimento *web*, foi utilizado o *framework Django*. Aqui, seu papel foi receber as requisições, buscar os dados necessários no banco de dados *MySQL*, montar a página HTML a partir do gabarito, e finalmente responder a requisição para ser exibida para o usuário em seu navegador.

3.2 Obtenção dos dados

Como já discutido nos capítulos anteriores, as licitações devem ser obrigatoriamente publicadas em algum meio. No caso da antiga Lei de Licitações no Diário Oficial, e na Nova Lei de Licitações no Portal Nacional de Contratações Públicas (PCNP). Estes sempre (exceto em casos excepcionais que não foram considerados neste trabalho, como falha de conexão, por exemplo) poderão ser consultados para obter os dados licitatórios.

O presente trabalho considera apenas as licitações publicadas em Diário Oficial, buscando os dados no documento, que é disponibilizado publicamente pela *internet*.

3.2.1 Diários Oficiais

Os Diários Oficiais são publicados no domínio *web* designado ao município. Eles são disponibilizados em forma de PDF, agregando informações como texto, layout, e imagens em um único documento, impossibilitado de ser editado sem o uso de *softwares* especializados. Adicionalmente, os arquivos disponibilizados são assinados digitalmente por uma autoridade da prefeitura, garantindo a autenticidade do documento. A Figura 1 apresenta o cabeçalho da primeira página do Diário Oficial Edição nº 3576, publicado no dia 16 de Julho de 2022, com indicação de assinatura digital no retângulo à direita, censurado para privacidade.

Figura 1 – Cabeçalho de um Diário Oficial da Prefeitura Municipal de Bauru

Fonte: Prefeitura Municipal de Bauru (2022)
Disponíveis em: <https://www2.bauru.sp.gov.br/juridico/diariooficial.aspx>

Apesar de conter informações de texto, estes dados são comprimidos e armazenados em formato binário, sendo necessário o uso de bibliotecas específicas para extraí-los. A descompressão, leitura, e extração dos dados ficou à cargo da ferramenta *Parser* do pacote de ferramentas *Apache Tika*^{TM}. Esses dados foram extraídos e salvos gerando um arquivo XHTML (*eXtended HyperText Markup Language*). Assim, este não só extrai as informações em texto, mas também é capaz de utilizar rótulos hiper-texto para separar algumas estruturas como páginas utilizando o rótulo "`div`" e parágrafos com o rótulo "`p`". Isso foi útil pois as quebras de linhas introduzidas devido ao layout de página geram quebras de linha no texto extraído, adicionando complexidade ao processo de extração de dados. O uso dos rótulos nessa tarefa se torna mais simples: a ocorrência da quebra de linha deixa de ser ambígua entre o final de um parágrafo ou o layout da página.

Esse arquivo foi então lido, processado, e manipulado utilizando a biblioteca *BeautifulSoup*. Nesta etapa também foram removidos dados que não são utilizados como metadados, cabeçalhos de páginas, e quebras de páginas.

Por fim, todos os parágrafos do DO são salvos em um arquivo de texto-simples utilizando a notação de objetos *JavaScript* (JSON, do inglês *JavaScript Object Notation*). Com isso, o acesso ao conteúdo de um DO é feito rapidamente a partir do índice numérico de uma lista de parágrafos.

3.2.2 Tabelas de licitações

A Prefeitura Municipal de Bauru já disponibiliza os dados estruturados de licitações. Contudo isso é feito por meio de páginas *web*, ao invés de uma interface de programação (API). Ou seja, os dados foram "desestruturados" durante a montagem da página, sendo necessário estruturá-los novamente.

São dois tipos de páginas utilizados: a página de listagem e a página de detalhes. A página de listagem possui uma tabela HTML (rótulos `table`, `tr` e `td`, por exemplo) com as colunas "Objeto", "Modalidade", e "Interessados", com links para a página de detalhes da licitação de cada linha. A Figura 2 apresenta uma captura de tela da página de "Licitações Abertas", idêntica à página de "Licitações Encerradas" e "Licitações Suspensas". Nela, é possível observar o início da tabela, indicado pelas palavras de cabeçalho "Objeto", "Modalidade", e "Interessados", e alguns registros com cores de fundo alternados entre cinza-claro e cinza-escuro.

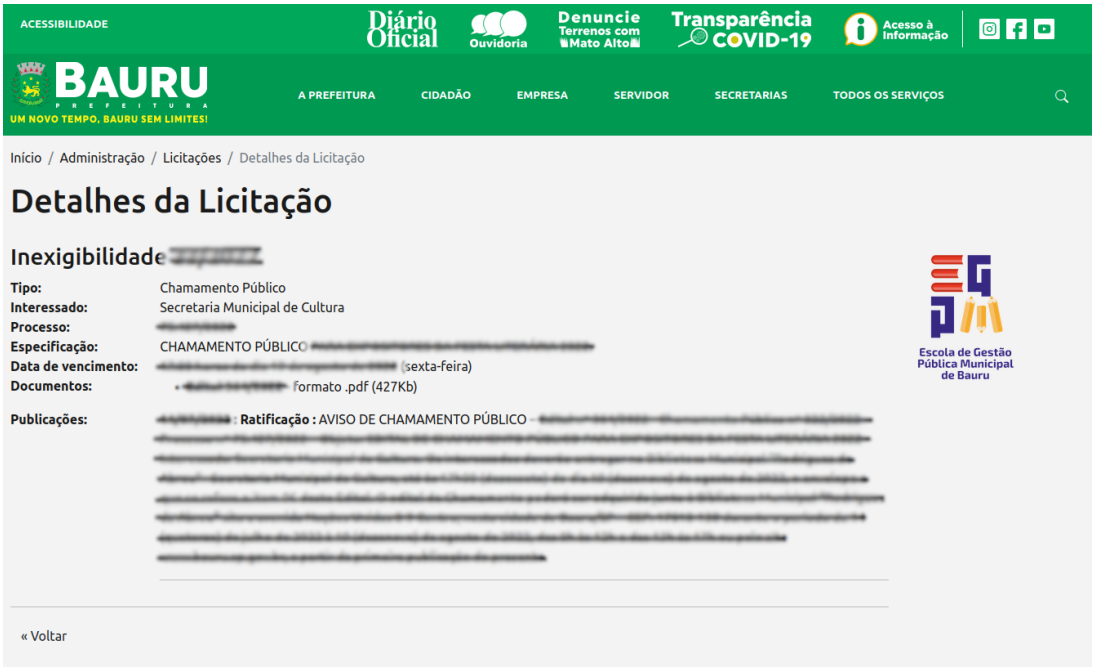
Figura 2 – Página "Licitações Abertas", apresentando a tabela com três colunas

Fonte: Prefeitura Municipal de Bauru (2022)
Disponível em: <https://www2.bauru.sp.gov.br/administracao/licitacoes/licitacoes.aspx?t=1>

Já a página de detalhes, apesar de ter a aparência de uma tabela, está estruturadas utilizando rótulos `div` aninhados com classes diferentes. A Figura 3 apresenta uma captura de tela da página de detalhes de uma licitação, acessada através dos links das tabelas de listagem. Nesta observa-se a estrutura de tabela, indicada pelos campos: "Tipo", "Interessado", "Processo", "Especificação", "Data de vencimento", "Documentos", e "Publicações", com seu conteúdo à direita.

As páginas são obtidas por meio da biblioteca `requests`, retornando o conteúdo em HTML. Novamente o *BeautifulSoup* foi utilizado para a leitura e manipulação do conteúdo

Figura 3 – Página "Detalhes da Licitação", apresentando diversos detalhes



Fonte: Prefeitura Municipal de Bauru (2022)
Disponível em:
<https://www2.bauru.sp.gov.br/administracao/licitacoes/licitacoes_detalhes.aspx?l=#>

hiper-texto. Neste estágio também foi utilizado buscas de expressões regulares, ou *RegEx* utilizando a biblioteca `re`.

Cada licitação teve seus dados coletados, estruturados, e armazenados em um *dict* e posteriormente salvos em um arquivo JSON.

3.3 Processamento e Comparação

Os dados coletados foram então estruturados e armazenados no banco de dados *MySQL* para agilizar a recuperação desses dados. Dessa forma, os dados ficam indexados através de vários dados de cada registro, ao invés de apenas seu título, como era com a abordagem em arquivos.

Devido ao grande volume de combinações de dados dos DOs, das tabelas, e de métodos utilizados, não é viável efetuar e registrar uma busca cartesiana dessas combinações. Logo, apenas alguns casos foram processados e apresentados na Seção 4. Estas e demais combinações podem ser calculadas sob demanda através das opções oferecidas na interface *web*.

Os métodos apresentados e discutidos neste trabalho são:

- o cálculo de confiança de licitação;
- o cálculo de similaridade cosseno, Jaccard, e de coincidência utilizando BoW e n-*grams*

de tamanhos 1 e 3; e

- a detecção de tema utilizando TF.IDF.

Um cálculo de similaridade de cosseno utilizando a técnica *word-embeddings* também foi proposta, contudo não foi utilizada devido ao seu alto custo computacional, impossibilitando executar em tempo hábil para este trabalho.

3.3.1 *Tokenização*

Como já discutido anteriormente, a unidade básica para o processamento da linguagem é a palavra. Logo, é necessário métodos para detectar os inícios, meios, e finais das palavras para que elas possam ser divididas e tratadas individualmente.

Devido à escala deste projeto, aqui foi utilizado uma simples busca por expressão regular (*RegEx*) para detectar letras e letras adjacentes. Cada grupo de letras contíguas foi considerado uma palavra. Isto foi necessário pois o uso da biblioteca até então em utilização neste projeto demandou uma quantidade de recursos computacionais maior que o autor poderia oferecer. Logo, uma abordagem mais leve foi necessária.

É necessário admitir as limitações desta abordagem, como a falta de um vocabulário, permitindo que palavras que não existam tenham o mesmo valor que palavras que existem, incluindo erros de digitação; a falta de detecção de formatos como datas, números, códigos, e emails; a falta de detecção de quebra de palavra; entre outros.

3.3.2 Estimativa de confiança

A partir da essência do TF.IDF que, relembrando, é formado por duas partes: TF, que indica quão frequente o termo é em um documento, e IDF, que indica em quantos documentos do *corpus* o termo está presente; nota-se que a segunda pode também indicar o vocabulário utilizado neste *corpus*, especialmente se esses textos forem semelhantes que, no contexto deste trabalho, são. Ou seja, esses segmentos possuem uma maneira de ser escrita que difere do resto, mas que são comuns entre si.

Assim, a estimativa apresentada nesta seção utiliza essa métrica para estimar um "nível de confiança" baseada em *scores*, e é usada para indicar o quanto um texto apresenta um vocabulário semelhante ao vocabulário utilizado nas publicações de licitações.

3.3.3 Cálculo de similaridades utilizando BoW

Para a tarefa de comparar dois documentos, foram utilizados as abordagens de similaridade de cosseno, de Jaccard, e de coincidência, com representações utilizando BoW e n-*grams*. Ou seja, dados dois documentos, $\mathbf{A}$ e $\mathbf{B}$, seus *tokens* foram extraídos e cada n-*gram* foi salvo

no BoW. Esses pares de BoWs tiveram suas similaridades calculadas com os três métodos apontados e apresentados de acordo.

3.3.4 Cálculo de similaridades utilizando *word-embeddings*

De maneira semelhante à abordagem apresentada na seção anterior, neste os parágrafos foram processados pela biblioteca *spaCy* e tiveram suas similaridades calculadas utilizando sua função de similaridade já implementada. Contudo, devido ao grande tempo de processamento e memória exigidos, comparado às abordagens anteriores, e o curto período de desenvolvimento deste projeto, este não pôde ser feito em larga escala, nem obtido dados suficientes para serem analisados.

3.3.5 Detecção de tema

Para a detecção do tema, utilizou-se a abordagem TF.IDF para cada termo encontrado no parágrafo em questão. Com essas métricas calculadas, foram selecionadas as dez maiores para serem exibidas.

3.4 Apresentação

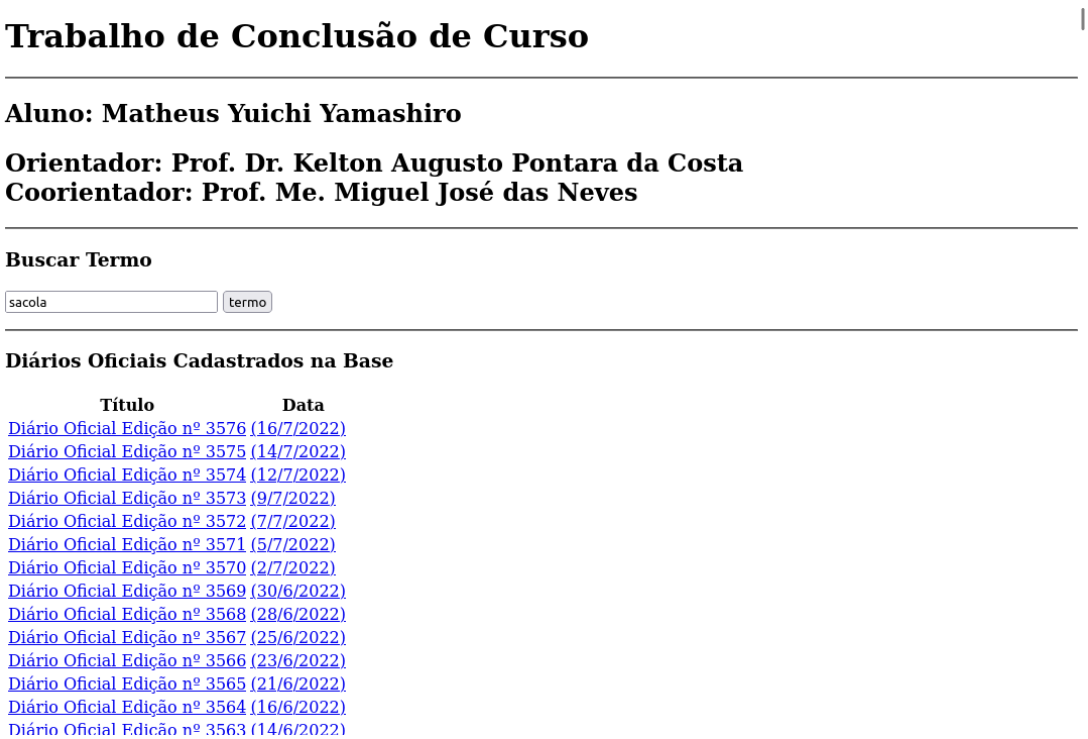
Para apresentar os dados, foi desenvolvido uma interface *web* utilizando o *framework Django*. A página inicial consiste de três partes: o cabeçalho, com o título da página e o nome dos autores do trabalho; um campo de busca; e uma listagem de DOs raspados e registrados na base. A Figura 4 apresenta uma captura da página inicial com o termo "sacola" no campo de busca.

O campo de busca permite a busca de um único termo que é utilizado para calcular seus TF.IDF em cada licitação publicada. Inicialmente o termo é buscado em toda a base de dados, seu TF.IDF é calculado em cada documento em que o termo é encontrado, e posteriormente exibido em uma outra página. A Figura 5 apresenta a página exibindo o resultado da busca pelo termo "sacola" (destaque nosso). Note que apenas o termo exato é contado (observe a coluna "Contagem"). Os termos semelhantes como "sacolas", "saco", ou "bolsa" não são, mas poderiam ser buscados se fosse usado a representação *word-embeddings*.

A listagem dos DOs raspados apenas exibe, em links, os DOs que já estão na base de dados e podem ser buscados. Cada DO é um link para uma página onde é listado todos os parágrafos raspados, junto com a métrica de confiança para aquele parágrafo. A Figura 6 apresenta essa página. Cada parágrafo também conta com um link para uma página em que os termos mais relevantes, de acordo com o TF.IDF, é exibido, como mostra a Figura 7.

O link da Figura 7 leva o usuário para uma página em que as publicações de licitações mais semelhantes são exibidas, como na Figura 8, ordenadas pela melhor similaridade de

Figura 4 – Página Inicial



Fonte: Elaborado pelo autor

coincidência, que obteve o melhor desempenho.

Todas as páginas foram geradas utilizando gabaritos que são automaticamente preenchidos pelo *Django* para responder à requisição. Dessa forma, apenas alguns gabaritos precisaram ser feitos, pois estes são completados dinâmicamente com os dados obtidos do banco de dados.

Figura 5 – Resultados da busca por “sacola”

paragrafo_id	Contagem	TF.IDF	Conteudo
99876984879842943	1	0.00513	NOTIFICAÇÃO DE ADJUDICAÇÃO / HOMOLOGAÇÃO – ÓRGÃO: PREFEITURA MUNICIPAL DE BAURU - SECRETARIA MUNICIPAL DE SAÚDE Processo. – Modalidade: Pregão Eletrônico SMS n. – Sistema de Registro de Preços – por meio da INTERNET – AMPLA DISPUITA E COTA RESERVADA PARA ME E EPP - Tipo Menor Preço por Lote – Objeto: aquisição estimada anual de 300.000 (trezentas mil) sacolas plásticas para medicamentos a serem distribuídos pelo Município e 264 (duzentos e sessenta e quatro) bobinas de filme plástico tipo strech. Aberto no dia: 31/07/2018 às 8 h. Notificamos aos interessados no Processo licitatório epigrafado, que o julgamento e a classificação havidos, foram adjudicados pelo pregoeiro em 24/08/2018 e devidamente Homologados pelo Sr. Secretário Municipal de Saúde em 24/08/2018, às empresas abaixo: LOTE 1: ITEM 01 - bobina com aproximadamente 4 kg de filme plástico tipo stretch, medindo 40 a 50 cm de largura x 0,025 mm de espessura; à R$ unitário; marca: LOTE 2: ITEM 02 - unidade de sacola plástica com alça cavada 30x32 para acondicionar medicamentos, impressão frente e verso (cores de impressão 4x4); à R$ unitário; marca: Bauru, 27/08/2018 - – Diretora Substituta da Divisão de Compras e Licitações – S.M.S.
99876984879842991	1	0.00452	NOTIFICAÇÃO DE ADJUDICAÇÃO / HOMOLOGAÇÃO – ÓRGÃO: PREFEITURA MUNICIPAL DE BAURU - SECRETARIA MUNICIPAL DE SAÚDE Processo: – Modalidade: Pregão Eletrônico SMS n° – Sistema de Registro de Preços – por meio da INTERNET – LICITAÇÃO EXCLUSIVA PARA ME E EPP - Tipo Menor Preço por Lote – Objeto: aquisição estimada anual de sacos de papel e sacolas plásticas para medicamentos a serem distribuídos pelo Município e bobinas de filme plástico tipo strech. Aberto no dia: 30/06/2017 às 8 h. Notificamos aos interessados no Processo licitatório epigrafado, que o julgamento e a classificação havidos, foram adjudicados pelo pregoeiro em 10/07/2017 e devidamente Homologados pelo Sr. Secretário Municipal de Saúde em 11/07/2017, à empresa abaixo: LOTE 1: ITEM 01 - bobina com aproximadamente 4 kg de filme plástico tipo stretch, medindo 40 a 50 cm de largura x 0,025 mm de espessura; à R$ unitário; marca: LOTE 2: ITEM 02 - sacos de papel kraft natural para acondicionar medicamentos, impressão frente e verso (cores de impressão 4x4); à R$ o pacote com 500 unidades; marca: LOTE 3: ITEM 03 - unidade de sacola

Fonte: Elaborado pelo autor

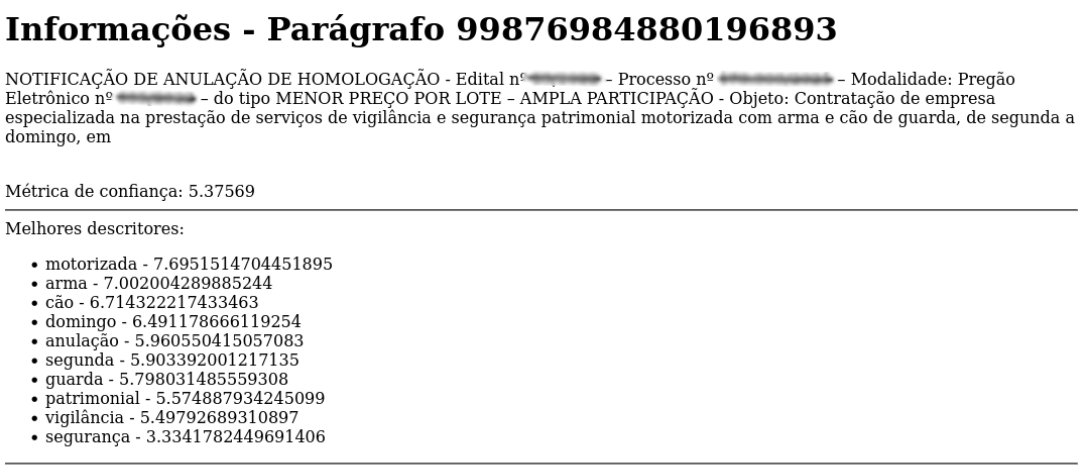
Figura 6 – Exibição dos parágrafos de um DO

Diário Oficial - Edição nº 3576

confiança	paragrafo_id	conteudo
0.00000	99876984880192429	
1.38136	99876984880192430	ANO XXVII - Edição 3.576 www.bauru.sp.gov.br SÁBADO, 16 DE JULHO DE 2.022 EDIÇÃO DIGITAL
0.02798	99876984880192431	Diário Oficial de Bauru ASSINADO DIGITALMENTE PELO DIRETOR DO DEPARTAMENTO DE COMUNICAÇÃO E
0.00000	99876984880192432	DOCUMENTAÇÃO
0.00015	99876984880192433	PODER EXECUTIVO
0.00000	99876984880192434	Prefeita Municipal
0.00394	99876984880192435	Seção I Gabinete da Prefeita
0.00349	99876984880192436	Chefe de Gabinete
0.01310	99876984880192437	LEIS MUNICIPAIS LEI Nº 7.568, DE 12 DE JULHO DE 2.022
0.00015	99876984880192438	P. 64.246/22 Autoriza a suplementação de recursos através de transposição no orçamento do exercício de 2.022.
0.61902	99876984880192439	A PREFEITA MUNICIPAL DE BAURU, nos termos do art. 51 da Lei Orgânica do Município de Bauru, faz saber que a Câmara Municipal, aprovou e ela sanciona e promulga a seguinte Lei: Art. 1º Fica autorizada a suplementação, através de transposição de recursos no Orçamento
0.12070	99876984880192440	vigente do Município de Bauru até o valor de da seguinte forma:
0.00902	99876984880192441	
0.00076	99876984880192442	Art. 2º O recurso necessário para atender o art. 1º decorrem de anulação parcial na dotação orçamentária
0.00932	99876984880192443	
0.00950	99876984880192444	Art. 3º Esta Lei entra em vigor na data da sua publicação. Bauru, 12 de julho de 2.022.
0.00076	99876984880192445	PREFEITA MUNICIPAL
0.00258	99876984880192446	SECRETÁRIO DOS NEGÓCIOS JURÍDICOS

Fonte: Elaborado pelo autor

Figura 7 – Exibição dos termos relevantes de um parágrafo



Informações - Parágrafo 99876984880196893

NOTIFICAÇÃO DE ANULAÇÃO DE HOMOLOGAÇÃO - Edital nº [redacted] – Processo nº [redacted] – Modalidade: Pregão Eletrônico nº [redacted] – do tipo MENOR PREÇO POR LOTE – AMPLA PARTICIPAÇÃO - Objeto: Contratação de empresa especializada na prestação de serviços de vigilância e segurança patrimonial motorizada com arma e cão de guarda, de segunda a domingo, em

Métrica de confiança: 5.37569

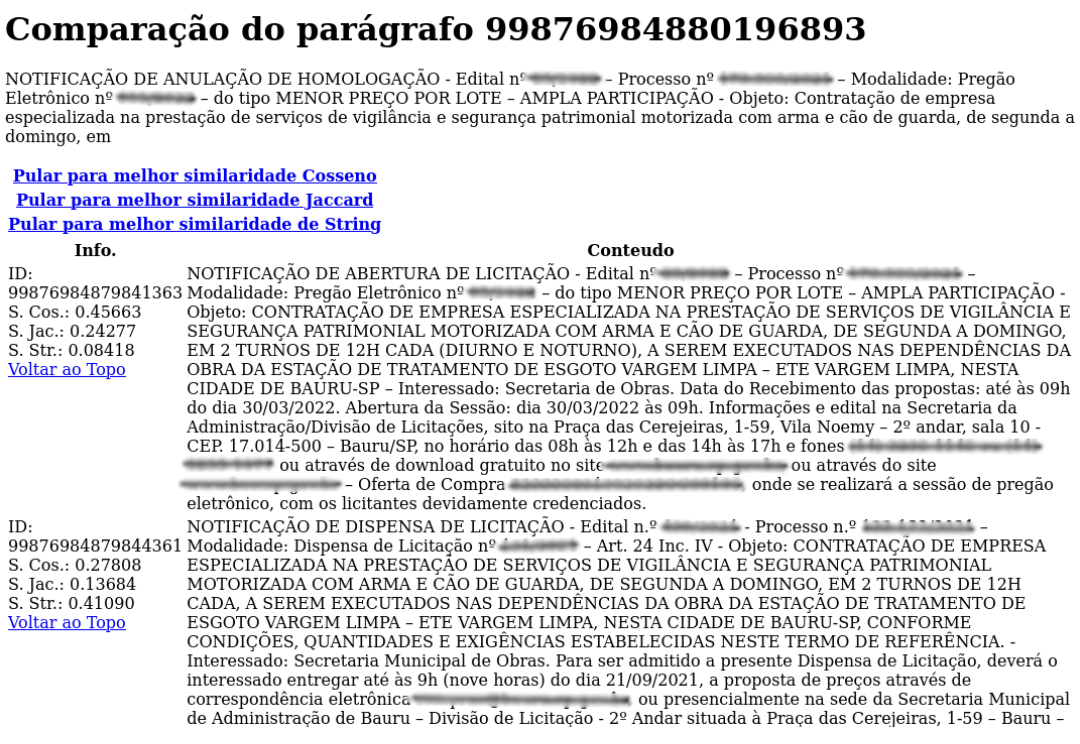
Melhores descritores:

- motorizada - 7.6951514704451895
- arma - 7.002004289885244
- cão - 6.714322217433463
- domingo - 6.491178666119254
- anulação - 5.960550415057083
- segunda - 5.903392001217135
- guarda - 5.798031485559308
- patrimonial - 5.574887934245099
- vigilância - 5.49792689310897
- segurança - 3.3341782449691406

Buscar na base

Fonte: Elaborado pelo autor

Figura 8 – Exibição das publicações mais similares



Comparação do parágrafo 99876984880196893

NOTIFICAÇÃO DE ANULAÇÃO DE HOMOLOGAÇÃO - Edital nº [redacted] – Processo nº [redacted] – Modalidade: Pregão Eletrônico nº [redacted] – do tipo MENOR PREÇO POR LOTE – AMPLA PARTICIPAÇÃO - Objeto: Contratação de empresa especializada na prestação de serviços de vigilância e segurança patrimonial motorizada com arma e cão de guarda, de segunda a domingo, em

Pular para melhor similaridade Cosseno
Pular para melhor similaridade Jaccard
Pular para melhor similaridade de String

Info.	Conteudo
ID: 99876984879841363 S. Cos.: 0.45663 S. Jac.: 0.24277 S. Str.: 0.08418 Voltar ao Topo	NOTIFICAÇÃO DE ABERTURA DE LICITAÇÃO - Edital nº [redacted] – Processo nº [redacted] – Modalidade: Pregão Eletrônico nº [redacted] – do tipo MENOR PREÇO POR LOTE – AMPLA PARTICIPAÇÃO - Objeto: CONTRATAÇÃO DE EMPRESA ESPECIALIZADA NA PRESTAÇÃO DE SERVIÇOS DE VIGILÂNCIA E SEGURANÇA PATRIMONIAL MOTORIZADA COM ARMA E CÃO DE GUARDA, DE SEGUNDA A DOMINGO, EM 2 TURNOS DE 12H CADA (DIURNO E NOTURNO), A SEREM EXECUTADOS NAS DEPENDÊNCIAS DA OBRA DA ESTAÇÃO DE TRATAMENTO DE ESGOTO VARGEM LIMPA – ETE VARGEM LIMPA, NESTA CIDADE DE BAURU-SP – Interessado: Secretaria de Obras. Data do Recebimento das propostas: até às 09h do dia 30/03/2022. Abertura da Sessão: dia 30/03/2022 às 09h. Informações e edital na Secretaria da Administração/Divisão de Licitações, sito na Praça das Cerejeiras, 1-59, Vila Noemy – 2º andar, sala 10 - CEP. 17.014-500 – Bauru/SP, no horário das 08h às 12h e das 14h às 17h e fones [redacted] ou através de download gratuito no site [redacted] ou através do site [redacted] – Oferta de Compra [redacted], onde se realizará a sessão de pregão eletrônico, com os licitantes devidamente credenciados.
ID: 99876984879844361 S. Cos.: 0.27808 S. Jac.: 0.13684 S. Str.: 0.41090 Voltar ao Topo	NOTIFICAÇÃO DE DISPENSA DE LICITAÇÃO - Edital n.º [redacted] - Processo n.º [redacted] – Modalidade: Dispensa de Licitação nº [redacted] – Art. 24 Inc. IV - Objeto: CONTRATAÇÃO DE EMPRESA ESPECIALIZADA NA PRESTAÇÃO DE SERVIÇOS DE VIGILÂNCIA E SEGURANÇA PATRIMONIAL MOTORIZADA COM ARMA E CÃO DE GUARDA, DE SEGUNDA A DOMINGO, EM 2 TURNOS DE 12H CADA, A SEREM EXECUTADOS NAS DEPENDÊNCIAS DA OBRA DA ESTAÇÃO DE TRATAMENTO DE ESGOTO VARGEM LIMPA – ETE VARGEM LIMPA, NESTA CIDADE DE BAURU-SP, CONFORME CONDIÇÕES, QUANTIDADES E EXIGÊNCIAS ESTABELECIDAS NESTE TERMO DE REFERÊNCIA. - Interessado: Secretaria Municipal de Obras. Para ser admitido a presente Dispensa de Licitação, deverá o interessado entregar até às 9h (nove horas) do dia 21/09/2021, a proposta de preços através de correspondência eletrônica [redacted], ou presencialmente na sede da Secretaria Municipal de Administração de Bauru – Divisão de Licitação - 2º Andar situada à Praça das Cerejeiras, 1-59 – Bauru –

Fonte: Elaborado pelo autor

4 Resultados e Discussão

Nesta seção serão apresentados as estatísticas dos dados coletados e dos métodos utilizados.

4.1 Coleta de dados

Apesar de ser capaz de coletar todos os dados disponíveis, apenas os dados de 1º de janeiro de 2016 à 18 de julho de 2022 foram coletados. Isso foi feito para reduzir o tamanho do conjunto de dados a ser processado devido ao curto período de desenvolvimento deste projeto. Isso resultou em 950 edições dos DOs de Bauru, totalizando 2,009,957 parágrafos para serem lidos; e 3,555 licitações disponíveis no site, totalizando 6,593 publicações de licitações em DO, sendo elas notificações, homologações, retificações, entre outros.

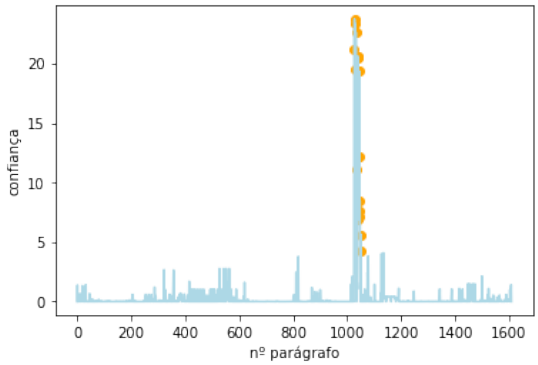
4.2 Comparação dos Métodos

Como descrito anteriormente, segmentos desses dois conjuntos de dados foram comparados utilizando diferentes técnicas para localizar suas intersecções. Ou seja, foram utilizadas técnicas para localizar um parágrafo de um corpo, em outras palavras um segmento de texto, dentro do outro.

A Figura 9 apresenta um DO escolhido aleatoriamente para exemplificar o comportamento do nível de confiança. Nela, há a curva em azul-claro e pontos em laranja. Ambas apresentam o nível de confiança do parágrafo no eixo x, mas os pontos dão destaque para os maiores valores, no caso, os valores maiores que o 99º percentil. Observa-se que estes se apresentam todos agrupados por volta do parágrafo 1050. Ou seja, possuem média em 1050 e desvio padrão pequeno, que é o comportamento esperado, já que as publicações se encontram agrupadas nos próprios DOs, na "Seção III - Editais", sob o título "Avisos".

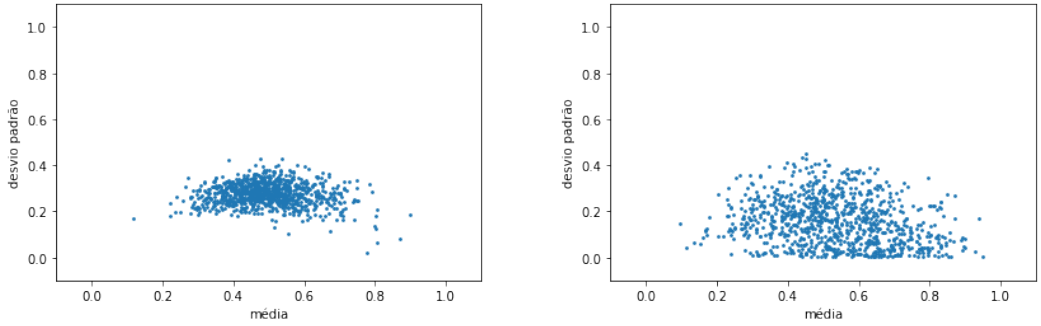
Quando se observa a posição normalizada dos parágrafos, nota-se a formação de uma tendência. A Figura 10 apresenta uma representação da média e do desvio padrão da posição de todos os pontos em 10a, e apenas os com confiança acima do 99º percentil em 10b. Nota-se que quando se considera apenas os parágrafos com maior confiança de serem licitações, a média dos parágrafos tende a ser encontrada mais no final do que no começo do DO, geralmente com desvios padrão reduzidos, entre 0.0 e 0.2. Ou seja, há um indício de que as licitações ocorrem agrupadas.

Figura 9 – Exemplo dos níveis de confiança de um DO



Fonte: Elaborado pelo autor

Figura 10 – Representação da média e desvio padrão da posição dos pontos



(a) Pontos acima do 95º percentil (b) Pontos acima do 99º percentil

Fonte: Elaborado pelo autor

4.2.1 Similaridades

Como apresentado anteriormente, foram utilizados três abordagens para calcular a similaridade entre dois parágrafos:

S. Cosseno que calcula o ângulo entre os vetores;

S. Jaccard que calcula a razão entre a intersecção e a união dos conjuntos; e

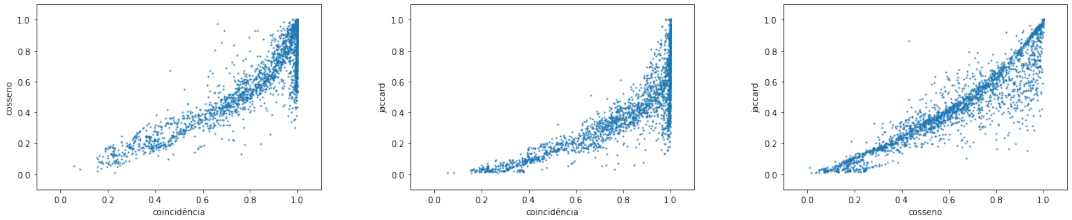
S. de coincidência que compara unidade a unidade.

As duas primeiras utilizaram bolsas de n-*grams* com 3 termos. Já a última com um único termo.

Apesar de utilizar apenas um único termo, a similaridade de coincidência obteve a melhor performance devido ao fato dos dois documentos, o publicado em DO e o publicado no sistema, já serem iguais, já que um deriva do outro.

A Figura 11 compara os pontos em que os três métodos concordaram em um único parágrafo como o mais similar de uma publicação. Os métodos sendo comparados estão indicados nos subtítulos.

Figura 11 – Comparação da performance dos métodos de comparação de texto



(a) Coincidência × Cosseno (b) Coincidência × Jaccard (c) Cosseno × Jaccard
Fonte: Elaborado pelo autor

Nota-se que em ambos os casos em que o método de coincidência é comparado, enquanto este indica que os textos são idênticos (Coincidência = 1.0), em muitas vezes o método sendo comparado acusa similaridades bem menores. Isso pode ser causado pela discordância na formação das bolsas, já que as de 3 termos são mais sensíveis à quebras de linha e outros artifícios de formatação, por exemplo.

Isso deixa de acontecer na Figura 11c, em que se compara dois métodos que utilizam bolsas feitas com 3 termos. Dessa maneira, ambas são afetadas pelos artifícios anteriormente citados de forma semelhante.

5 Conclusão

Neste trabalho diversas técnicas para representar, processar, e comparar textos escritos em linguagem natural foram estudadas e utilizadas. Os resultados obtidos foram apresentados utilizando uma *framework* de desenvolvimento de páginas *web*.

Com o crescimento do volume de conteúdo gerados por usuários na internet, o processamento desses dados se torna impossível se não forem automatizados. Para isso, técnicas de processamento de linguagem natural são estudadas, implementadas, desenvolvidas, e aprimoradas. Este trabalho apresenta um estudo introdutório para a área de pesquisa.

Além disso, a apresentação utilizando tecnologias *web* acompanha seu crescimento global, que por sua vez acompanha a necessidade de agilidade e versatilidade no mundo digital e conectado. Apesar deste trabalho apresentar interfaces simples, é também um estudo introdutório à área de atuação de desenvolvimento *back-end*.

Como a área de processamento de linguagem natural é, não só extensa, mas em desenvolvimento e de muito interesse e valor, ainda existem muitos aspectos em que a ferramenta deste trabalho pode ser aprimorado. Muitos segmentos desta área de pesquisa poderiam aprimorar a performance e o escopo desta, ou de outras ferramentas, mas não foram explorados neste trabalho. Alguns desses incluem: a detecção e correção de erros de digitação (utilizando o método de Lavenshtein, por exemplo), o processamento de gírias, a detecção automática de dados numéricos como datas e contadores (números de processos e modalidades), a sugestão de termos correlatos (utilizando representações como *word-embeddings*), e muitos outros.

Além disso, a interface de usuário também pode ser aprimorada utilizando *frameworks* e bibliotecas *front-end* especializadas, tornar a interface responsiva ao dispositivo, e apresentar os resultados de forma gráfica. Isso tornará a ferramenta mais agradável e intuitiva de ser utilizada.

Referências

Apache Tika. *Apache Tika - a content analysis toolkit*. 2022. Disponível em: <https://tika.apache.org/>. Acesso em: 20 jul. 2022.

BeautifulSoup. *Beautiful Soup Documentation*. 2022. Disponível em: <https://beautiful-soup-4.readthedocs.io/en/latest/>. Acesso em: 20 jul. 2022.

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JAUVIN, C. A neural probabilistic language model. *Journal of Machine Learning Research*, v. 3, p. 1137–1155, 2003.

BRASIL. Constituição da república federativa do brasil de 1988. Brasília, DF: Presidência da República, 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 05 mai. 2022.

BRASIL. Lei nº 8.666, de 21 de junho de 1993. *Diário Oficial União*, Brasília, DF: Presidência da República, 1993. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L8666compilado.htm>. Acesso em: 05 mai. 2022.

BRASIL. Lei nº 10.520, de 17 de julho de 2002. *Diário Oficial União*, Brasília, DF: Presidência da República, 2002. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/2002/l10520.htm>. Acesso em: 05 mai. 2022.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. *Diário Oficial União*, Brasília, DF: Presidência da República, 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 05 mai. 2022.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*. 3. ed. Pearson Prentice Hall, 2021. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>. Acesso em: 20 jul. 2022.

MARTINS, J. S. Similaridade léxica. In: SAGAH. *Processamentos de Linguagem Natural*. 1. ed. Porto Alegre: Grupo A, 2020. cap. 1, p. 89–102. ISBN 978-6-55-690057-5. Disponível em: <https://online.vitalsource.com/books/9786556900575>. Acesso em: 05 mai. 2022.

MASSONI, G. *Análise de textos por meio de processos estocásticos na representação word2vec*. 59 p. Dissertação (Mestrado em Estatística - Programa Interinstitucional de Pós-Graduação em Estatística) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, 2021. Disponível em: <https://www.teses.usp.br/teses/disponiveis/104/104131/tde-31032021-123649/pt-br.php>. Acesso em: 20 jul. 2022.

RAJARAMAN, A.; ULLMAN, J. D. *Mining of massive datasets*. Cambridge University Press, 2011. Disponível em: <http://infolab.stanford.edu/~ullman/mmds.html>. Acesso em: 20 jul. 2022.

RINO, L. H. M.; PARDO, T. A. S. A sumarização automática de textos: principais características e metodologias. In: *Anais do XXIII Congresso da Sociedade Brasileira de Computação*. [S.l.: s.n.], 2003. v. 8, p. 203–245.

RODRIGUES, S. M. A. F. Introdução ao processamento de linguagem natural. In: SAGAH. *Processamentos de Linguagem Natural*. 1. ed. Porto Alegre: Grupo A, 2020. cap. 1, p. 13–34. ISBN 978-6-55-690057-5. Disponível em: <https://online.vitalsource.com/books/9786556900575>. Acesso em: 05 mai. 2022.