SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PAMPA

PAULO HENRIQUE SEIXAS LEITE

UMA FERRAMENTA COMPUTACIONAL PARA OBTENÇÃO DA MATRIZ TERMO OCORRÊNCIA EM CORPUS TEXTUAIS

Bagé, 2021

PAULO HENRIQUE SEIXAS LEITE

UMA FERRAMENTA COMPUTACIONAL PARA OBTENÇÃO DA MATRIZ TERMO OCORRÊNCIA EM CORPUS TEXTUAIS

Trabalho de Conclusão de Curso apresentado ao Curso de especialização Lato Sensu Modelagem Computacional em Ensino, Experimentação e Simulação da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Especialista em Modelagem Computacional.

Bagé, 2021

SERVIÇO PÚBLICO FEDERAL

MINISTÉRIO DA EDUCAÇÃO

Universidade Federal do Pampa

PAULO HENRIQUE SEIXAS LEITE

UMA FERRAMENTA COMPUTACIONAL PARA OBTENÇÃO DA MATRIZ TERMO OCORRÊNCIA EM CORPUS TEXTUAIS

Trabalho de Conclusão de Curso apresentado ao Curso de especialização Lato Sensu Modelagem Computacional em Ensino, Experimentação e Simulação da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Especialista em Modelagem Computacional.

Trabalho de Conclusão de Curso defendido e aprovado em: 06, dezembro de 2021.

Banca examinadora:

_____

Prof. Dra. Vera Lúcia Duarte Ferreira

Orientador

UNIPAMPA

Prof. Dr. Fernando Luis Dias

UNIPAMPA

Prof. Dr. Paulo Fernando Marques Duarte Filho

UNIPAMPA

Assinado eletronicamente por **VERA LUCIA DUARTE FERREIRA**, **PROFESSOR DO MAGISTERIOSUPERIOR**, em 10/01/2022, às 15:19, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **FERNANDO LUIS DIAS**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em10/01/2022, às 15:26, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **PAULO FERNANDO MARQUES DUARTE FILHO**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 19/01/2022, às 15:21, conforme horário oficial de Brasília, de acordocom as normativas legais aplicáveis.

A autenticidade deste documento pode ser conferida no site https://sei.unipampa.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0709054** eo código CRC **B7E53ADB**.

Referência: Processo nº 23100.020744/2021-51 SEI nº 0709054

Ficha catalográfica elaborada automaticamente com os dados fornecidos pelo(a) autor(a) através do Módulo de Biblioteca do
Sistema GURI (Gestão Unificada de Recursos Institucionais).

L457f Leite, Paulo Henrique Seixas
Uma ferramenta computacional para obtenção da matriz termo ocorrência em corpus textuais / Paulo Henrique Seixas Leite.
39 p.

Trabalho de Conclusão de Curso(Especialização)--Universidade Federal do Pampa, ESPECIALIZAÇÃO EM MODELAGEM COMPUTACIONAL EM ENSINO, EXPERIMENTAÇÃO E SIMULAÇÃO, 2021.
"Orientação: Vera Lúcia Duarte Ferreira".

1. Mineração Textual. 2. Ferramenta Computacional. 3.Frequência de Palavras. I. Título.

RESUMO

O presente trabalho apresenta a primeira versão de ferramenta computacional para análise de dados não estruturados, desenvolvida em linguagem Python e embasada em técnicas de mineração de texto e processamento de linguagem natural. A ferramenta proposta tem centralidade na análise lexical através de frequência das palavras e na posterior determinação da matriz termo ocorrência de um corpus textual. A aplicação do experimento valeu-se de um corpus textual do gênero notícias compostas de sites de notícias da internet. Os resultados mostraram a eficiência da referida ferramenta para análise lexicográfica dos verbos dicendi utilizados como descritores, o gênero notícias, por meio de gráficos apresentados com a frequência de palavras, nuvem de palavras, produzidos a partir da matriz de ocorrência saída da ferramenta computacional.

Palavras-chave*:* Mineração Textual; Ferramenta Computacional, Frequência de Palavras.

ABSTRACT

This work presents the first version of a computational tool for analyzing unstructured data, developed in Python language and based on text mining and natural language processing techniques. The proposed tool has centrality in the lexical analysis through the frequency of the words and in the subsequent determination of the term occurrence matrix of a textual corpus. The Application of the Experiment used a textual news corpus composed of internet news sites. The results required the efficiency of the tool for lexicographical analysis of the dicendi verbs used as descriptors, the news genre, through graphics, with the frequency of words, word cloud, bootable from the output matrix of the computational tool.

Keywords: Textual Mining; Computational Tool, Word Frequency.

LISTA DE FIGURAS

LISTA DE TABELAS

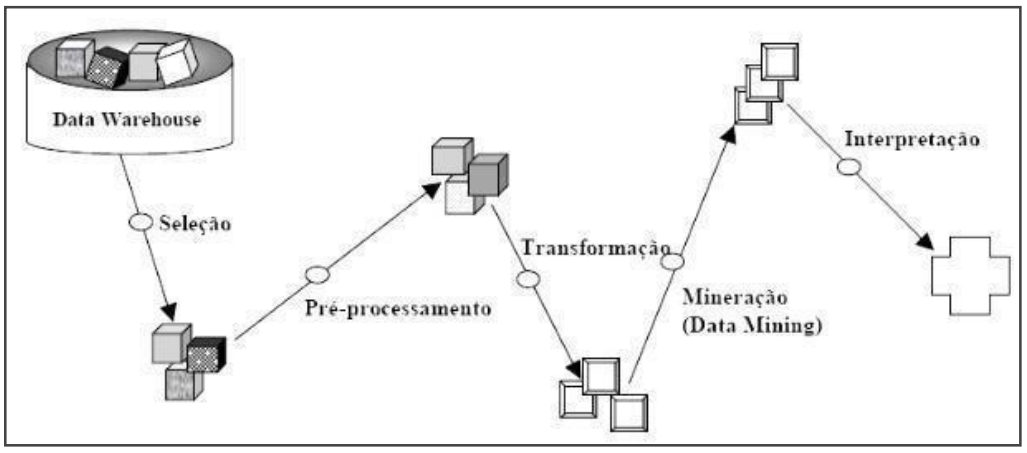
Sumário

1 INTRODUÇÃO

O termo Mineração de Dados (do inglês, data mining), denomina o processo de analisar diferentes dados, consistindo em um processo de descoberta do conhecimento a partir de uma base de dados e transformá-los em informações importantes como, por exemplo, perfis de consumidores ou colaboradores que conseguem determinar vários fatores como preços e indicadores econômicos. Outro processo chamado KDD (do inglês, "*Knowledge Discovery in Databases*") é definido como busca de conhecimentos em Banco de dados, sendo relacionado à área que tem como objetivo a descoberta de informações novas dentro do contexto da análise de grandes quantidades de dados (SOUZA 2016; CUNHA, 2018).

A Figura 1 abaixo, demonstra o processo de funcionamento do KDD.

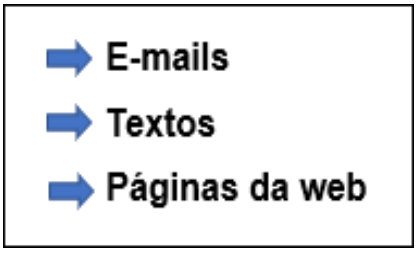
Figura 1- Etapas do KDD



Fonte: BOENTE et.al, 2007.

A mineração de dados utiliza várias técnicas, entre elas a mineração de texto que é um conjunto de métodos para organizar, analisar e encontrar informações em bases textuais. O processo de mineração de texto tem como objetivo extrair os dados e realizar uma análise no texto (PEZZINI, 2016). Os exemplos de extração de dados, podem ser observados na Figura 2.

Figura 2- Exemplos de extração de dados



Fonte: O autor.

A Ferramenta de busca de documentos é muito importante e poderosa para os pesquisadores nas áreas do conhecimento, existem inúmeros documentos disponíveis, e encontrar aqueles que são relevantes pode contribuir para o desenvolvimento de diversos estudos (SCARPA, 2017).

Segundo Da Silva e Silva 2019, a linguagem python surgiu na cidade de Amsterdã, na capital da Holanda, um dos desenvolvedores da linguagem é Guido Van Hossum que trabalhava no CWI (Instituto de Pesquisa Nacional para Matemática e Ciência da Computação) em um sistema chamado amoeba. No entanto, como esse sistema apresentava várias falhas e era desenvolvido em linguagem C. Guido resolveu desenvolver outra linguagem para resolver os problemas que outra linguagem apresentava, logo o holandês batizou a linguagem como python devido ao seu programa favorito que era o Monty Python 's Flying Circus.

Entretanto várias linguagens são utilizadas na programação, cada uma com sua contribuição à linguagem em PHP é voltada para as aplicações web, já o Java é voltado para o desenvolvimento em desktop, o Python não possui objetivo. A linguagem oferece suporte a desktop e desenvolvimento web,

aplicação mobile, geoprocessamento, processamento de Data Scienc científico, pois, trabalha com grande número de informações e utiliza diversas bibliotecas. A Linguagem em Python é encontrada no cotidiano de muitos usuários presente em várias ferramentas digitais tais como nos buscadores de pesquisa como o Google no processamento de pesquisa de dados e em plataformas de streaming como Youtube e Netflix e entre outras grandes empresas (DA SILVA E SILVA, 2019).

As informações científicas estão crescendo em grande escala de maneira que novas demandas vão surgindo, essas informações são apresentadas em formatos como PDF ou HTML. Para aqueles que buscam por informações de características científicas como pesquisadores, é exigido um tempo maior no processo de leitura textual e com isso a necessidade de desenvolvimento de técnicas capazes de extrair conteúdo de forma rápida e objetiva (FERREIRA E CORREA, 2021).

Duas categorias definem a mineração de dados em Tarefas de Previsão que podem prever o valor de um atributo baseado nos demais atributos, na forma que o atributo a serem previsto é conhecido como o atributo alvo, já os demais atributos podem ser definidos como variáveis explicativas e Tarefas Descritivas que tem por objetivo atribuir padrões de correlações, agrupamentos e tendências, as tarefas descritivas são frequentemente utilizadas de forma exploratória, necessitando técnicas de pós processamento para a validação dos dados (TAN, STEINBACH E KUMAR, 2009).

O software Sobek está disponível tanto na versão on-line quanto na versão para *download*, essa ferramenta utiliza o processo de organização textual, inserção do conteúdo textual e geração do grafo dos termos mais relevantes extraídos dos documentos. A ferramenta foi desenvolvida pelo programa de Pós-graduação em informática na Educação, da Universidade Federal do Rio Grande do Sul – UFRGS, o destaque do *software* é já ter sido utilizado em textos em português e, com isso, gerar como resultado um gráfico de mapa mental com uma representação visual do texto (MEDEIROS et. al, 2019).

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Elaborar uma ferramenta computacional em linguagem Python no intuito de realizar análise lexicográfica.

1.1.2 Objetivos Específicos

- Elaborar o algoritmo de limpeza de dados;
- Apresentar a frequência de palavras via bag-words;
- Determinar matriz de termo de ocorrências.

1.2 JUSTIFICATIVA

A busca textual é distribuída a partir de inúmeras informações, e milhares de documentos disponíveis nas plataformas web. Tornando uma tarefa humanamente impraticável para buscar documentos com assuntos específicos baseados em palavras-chaves com agilidade, sem o auxílio de uma ferramenta de mineração de texto.

Visando facilitar a análise de busca de textos e a interação entre elementos textuais, a partir da ferramenta web em linguagem Python, desenvolvida com o intuito de agilizar a busca de conteúdos didáticos, justifica-se este estudo.

1.3 DESCRIÇÃO DO PROBLEMA

A procura textual na pesquisa acadêmica ainda impõe algumas dificuldades no acesso de assuntos específicos, existem ferramentas que auxiliam por meio de acesso aberto e outras somente por cotação.

Perante essa imposição buscou-se elaborar uma ferramenta de busca de documentos acadêmicos para auxiliar na similaridade entre textos, capaz de gerar matrizes e frequência de palavras em textos, facilitando o acesso a determinados assuntos.

2 FUNDAMENTAÇÃO TEÓRICA E ESTUDOS RELACIONADOS

2. 1 Surgimento da mineração de dados

A mineração de dados surgiu para analisar um volume de dados em grande escala. O auxílio das técnicas de mineração de dados para criar perfis de usuários ou clientes (TAN, et. al, 2009). Conforme Amo 2009, a mineração de dados é um ramo da computação que teve início na década 80, quando os profissionais de TI das empresas e organizações começaram a se preocupar com os grandes volumes de dados na empresa. Nesta época, Data Mining consistia essencialmente em extrair informação de gigantescas bases de dados da maneira mais automatizada possível.

Durante os últimos anos com avanço da tecnologia e aumento da quantidade dos dados armazenados, conforme Santos 2009, avanços recentes em várias áreas tecnológicas possibilitaram um crescimento explosivo na capacidade de gerar, coletar, armazenar e transmitir dados digitais. Na primeira década do século 21 já temos a possibilidade de armazenar vários gigabytes em dispositivos portáteis e alguns terabytes em computadores pessoais a um custo acessível. Uma quantidade quase incomensurável de informações de diversos tipos, origens, formatos e finalidades estão disponíveis na Internet, podendo ser acessadas a partir destes dispositivos comuns. O baixo custo dos dispositivos e do acesso a redes de computadores fez também com que o número de usuários destes sistemas aumentasse consideravelmente. Novas ferramentas permitem que estes usuários criem conteúdo digital de forma relativamente simples e barata, o que só faz aumentar a quantidade de informações disponíveis para outros usuários.

Entretanto Dias et. al, 2008, com a necessidade da utilização de infraestruturas maiores de redes, os analistas têm encontrado dificuldades para a alocação otimizada dos recursos da rede, tais como: banda, fila e outros recursos que são limitados. Além disso, as redes estão cada vez maiores e mais complexas, o que aumenta ainda mais a dificuldade de administrar prioridades em sua utilização, controlar a dinâmica de seu funcionamento e diagnosticar seus problemas. WEKA é uma das ferramentas mais utilizadas na Mineração de Dados, por ser software livre e prover um conjunto de algoritmos que

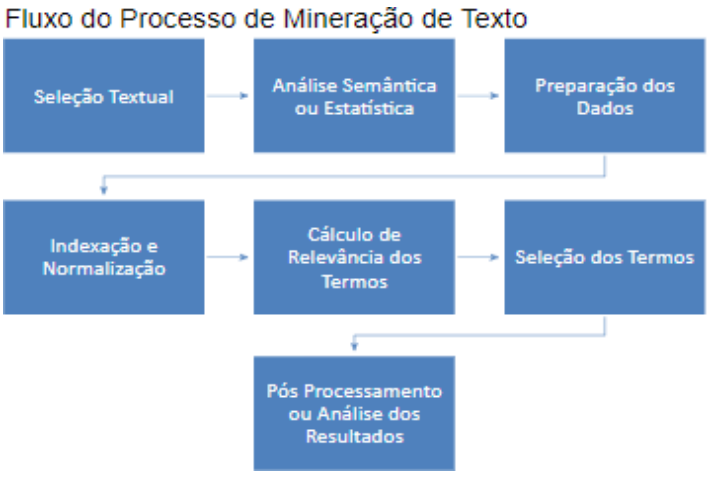
implementam diversas técnicas para resolver problemas reais de Mineração de Dados. Esta ferramenta foi implementada na linguagem Java e desenvolvida no meio acadêmico da Universidade de Waikato, na Nova Zelândia, em 1999. Suas principais características são herdadas do fato de ser uma ferramenta desenvolvida em Java, uma linguagem multi-plataforma orientada a objetos.

2.2 Mineração de Texto

Segundo Pezzini 2016, o principal objetivo da mineração de texto é buscar termos importantes em documentos de texto com grande volume de informação para estabelecer padrões e relacionamento com base na frequência dos termos encontrados. A mineração de texto tem como a principal contribuição na busca de informações específicas em documentos como na análise qualitativa e quantitativa em grandes volumes de texto e assim um melhor domínio no conteúdo (MORAIS & AMBRÓSIO, 2007).

As etapas relacionadas a Mineração de Texto está contida a partir do fluxograma do processo de mineração de texto são as seguintes: seleção do documento, a definição do tipo de escolha dos dados (análise semântica ou estatística) preparação dos dados, indexação e normalização, cálculo da relevância dos termos, seleção dos termos e pós- processamento (análise de resultados). Este processo pode ser visualizado no fluxograma na Figura 3.

Figura 3 - Fluxograma do Processo de Mineração de Texto



Fonte: O autor

São abordados dois tipos de análise de dados textuais, a análise estatística, que é baseada na frequência de vezes que os termos são encontrados no texto e análise semântica baseada na funcionalidade dos termos encontrados no texto.

- Análise Semântica é método que analisa a sequência dos termos com a condição de identificar qual sua função no texto;
- Análise Estatística é um método que tem como objetivo verificar a quantidade de vezes que um termo aparece no texto para verificar a importância dele;
- Preparação dos dados é etapa que é realizado o reconhecimento do texto é momento que seleciona o núcleo do texto que melhor expressa o conteúdo do texto. O objetivo dessa etapa é identificar a similaridade dos termos nos textos;
- Indexação e Normalização é etapa que identifica o significado e a similaridade entre as palavras do texto, verificar as variações morfológicas. A principal característica nesta etapa é a identificação dos documentos e gerar um índice, é o processo de indexação que nada mais é do que facilitar e agilizar o processo de localização e organização dos termos no texto;
- Cálculo da Relevância é etapa para verificar que nem todas as palavras que estão presentes no texto são de suma importância. Os termos mais frequentemente utilizados com a exceção das stopwords, assim como as palavras que aparecem nos títulos porque por sua vez são consideradas palavras relevantes com intuito de dar ideia do documento. A importância das palavras indica-se com o peso dela pela frequência de vezes que foi visualizada no texto, para determinar esse peso existem várias fórmulas entre elas cálculos simples de frequência: frequência absoluta, frequência relativa e frequência inversa de documentos;
- Seleção de Termos é etapa que realiza a seleção das palavras retiradas do texto após as outras etapas como de pré- processamento e cálculo da relevância, essa técnica é baseada no peso das palavras e suas posições em relação ao texto. As técnicas mais utilizadas são a filtragem baseada no peso do termo, seleção por análise de co-ocorrência, seleção por Latent Semantic Indexing Seleção por análise de linguagem natural;

- Análise de Resultados nesta etapa são aplicadas as técnicas para verificação dos resultados, nesta parte podemos aplicar métodos matemáticos e estatísticos em documentos, que possam auxiliar na prática com os resultados para resolução de problemas ou sugestões de melhorias, estas métricas podem ser utilizadas como de suma importância pelos usuários quais os textos são mais relevantes e similares além da importância do contexto do documento de cada texto.

Segundo Júnior 2007, a implementação de um fluxo com as etapas para análise do texto, permite criar e detalhar de forma clara todo o processo de mineração de texto, as etapas estão descritas abaixo:

- Seleção textual: essa etapa tem como objetivo selecionar uma base de dados textual para análise;
- Preparação dos dados: após a seleção textual, a preparação dos dados tem como objetivo utilizar técnicas de formatação com algoritmos para extração da informação;
- Indexação e normalização: é a etapa que organiza os termos do texto com um índice para garantir agilidade ao processo e rapidez.
- Análise dos dados: essa etapa remete ao conhecimento extraído e que deve se tornar uma tomada de decisão no processo de mineração de texto JÚNIOR (2007).
- A etapa da coleta: Nessa etapa é realizada a aquisição dos dados podendo ser caracterizada como a etapa mais importante para o desenvolvimento da amostra.
- A etapa do pré-processamento é realizada a preparação da informação para torná-la adaptável ao modelo de mineração, nesse momento a técnica de mineração influencia diretamente no processo.
- A etapa de mineração é o momento que são aplicados os algoritmos para extrair as informações importantes do texto com maior relevância, nessa etapa da análise dos dados são realizadas as visualizações e verificados os resultados extraídos do texto.
- A etapa da análise é a fase que o texto já está com a base das informações para ser utilizada, já está pronta para interpretação humana, nesta fase é diagnosticado um padrão ou uma tendência nos dados apresentados para que

possam ser utilizados com intuito de auxiliar nas pesquisas científicas (BARBOSA, 2018).

Conforme Silva 2004, o pré-processamento é subdividido em várias etapas correspondente aos métodos de pré-processamento, análise léxica e remoção dos termos desnecessários. Uma lista de termos como stopwords, contendo termos tais como artigos, preposições, verbos, pronomes, são extraídos dos textos, para que possa realizar a limpeza do texto.

No pré-processamento os dados são armazenados em bases de dados que ainda não estão de acordo com o formato adequado para extração da informação é necessário a aplicação de técnicas de extração, como a seleção dos dados para a redução destes volumes antes de outras etapas na mineração. A técnica de limpeza dos dados se faz necessária para deixar o processo com maior qualidade e garantir que as informações serão precisas, a remoção da mesma é realizada a retirada de valores inválidos para determinados atributos como por exemplo a redução dos dados é necessário em virtude de não ter capacidade de memória ou processamento.

Conforme Scarpa 2017, as técnicas de pré-processamento são utilizadas para aumentar a eficiência do algoritmo. A autora ainda pontua a relevância de remover do corpus textual partes do conteúdo original que não contenham informação importante que vá contribuir para a análise do texto. Nesse sentido, uma das técnicas utilizadas é a remoção das Stopwords, que consiste em remover do corpus textual palavras que são extremamente comuns, como conjunções, artigos definidos e indefinidos, preposições, pronomes e afins.

A limpeza do texto é composta por vários métodos dentre elas a Tokenização, Normalização, Extração de Entidades, Correção Ortográfica e Gramatical, Remoção de Pontuação, Remoção de Caracteres Especiais e Stemização e Lematização.

2.3 Stopwords e termos com baixo valor preditivo

As Stopwords são palavras que não possuem relevância na análise do texto, por isso são removidas normalmente para melhorar o desempenho da mineração de texto na busca da informação desejada entre as stopwords estão

as preposições, pronomes, artigos, advérbios e outras classes de palavras e também existem palavras que aparecem com muita frequência em praticamente todos os documentos e não são capazes de discriminar o documento, por isso foi criado uma lista de stopwords que chamamos de stoplist que são palavras que dificilmente são utilizadas para esses fins, por que tornaria a consulta da informação muito maior sem necessidade e deixar a consulta mais lenta e poderia não trazer os resultados desejados (SILVEIRA, 2011).

De acordo com Silveira 2011, os preditivos são palavras com baixo valor semântico que são geralmente artigos, advérbios, preposições e conjunções por isso para nível de classificação no texto são descartadas e geram uma diminuição dos termos a serem considerados para análise no texto e os termos restantes são de maior importância e quando utilizamos as técnicas o objetivo principal é utilizar os termos que sejam únicos com grande valor preditivo. As stoplist são palavras com pouco valor preditivo, e são retiradas durante a etapa de pré-processamento dos textos.

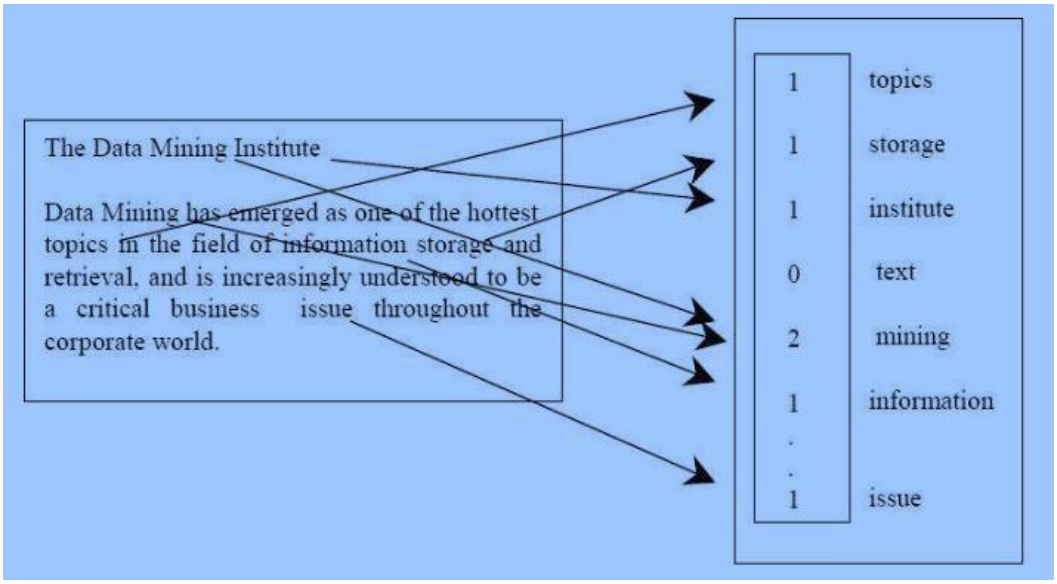
2.4 Modelo de Bag Of Words

Conforme Beysolow II 2018, o modelo Bag Of Words é um algoritmo de extração com objetivo de analisar o número de vezes que uma determinada palavra se apresenta no texto. Utilizando a técnica de associação para correlacionar as palavras de um texto a partir de padrões de combinação de palavras que sejam significativas ao usuário para o processo de mineração de texto (ROSSI, 2011).

Através de um conjunto de algoritmos que realiza a classificação das palavras em texto conforme suas características são armazenadas em um vetor. No entanto, para diminuir a quantidade de informações desnecessárias é preciso realizar técnicas de limpeza de dados, ignorando pontuações e palavras frequentes (JUSTINO; MORIMOTO; GOBBATO, 2019).

A Figura 4, apresenta a representação Bag-of-Words.

Figura 4 - Representação Bag-of-Words



Fonte: Silveira et. al, 2011.

2.5 Método de vetorização

O método de vetorização é amplamente utilizado para classificar e organizar as palavras em forma de vetores, com uma entrada para cada termo que aparece no texto, o texto é visto como um saco de palavras que é considerado apenas a quantidade de vezes que cada palavra aparece. A matriz utiliza menos memória e com isso o processamento do cálculo é bem mais rápido (SCARPA, 2017).

No que tange a transformação de texto em números, se tem o método de Bag-of- Words consiste em transformar textos em um conjunto de dados através de uma matriz. Cada linha da matriz é representada pelo texto e as ocorrências das palavras são indicadas pelo número vezes apresentado no texto (CANTO, 2020). A técnica de vetorização realiza o pré- processamento e tem como o processo de remoção de stopwords do texto e com isso realiza técnicas para transformar as palavras em vetores e com essa vetorização é feita à similaridade entre as palavras presentes no texto. O objetivo da vetorização é gerar vetores de palavras (MARQUES, 2018).

2.6 Método de Processamento de Linguagem Natural

Segundo Canto 2020, o método PLN (Processamento de Linguagem Natural) é utilizado para ensinar e auxiliar os dispositivos tecnológicos a entender a linguagem dos humanos para poder ser usado em diversas aplicações de assistentes pessoais, como OK Google, Siri, Cortana, Alexa além da utilização de aplicativos que permitem a tradução de idioma, como o Google Translator. Conforme Luques 2020, a área de Processamento de Linguagem Natural (PLN) que aborda modelos para compreender e manipular os dados na forma de texto ou fala em linguagem natural e os utilizar no dia a dia para realizar tarefas.

A inteligência artificial tem várias subáreas, como descreve Cerqueira 2021, dentre elas a PLN que tem a realização de tarefas que dependem de informações expressas em linguagem natural seja ela falada ou escrita conforme a comunicação humana. Os computadores têm um importante papel na vida humana, poder entender e gerar informações expressas em PLN se faz necessário com a atual demanda social.

2.6.1 Técnicas em Mineração

São utilizadas várias técnicas de mineração de dados, dentre elas a teoria de Redes Neurais que contribui com o desenvolvimento do algoritmo das máquinas de vetores que utilizam tarefas de classificação. As técnicas em mineração, são divididas em rede neurais e árvores de decisão.

2.6.2 Técnica de rede neurais

O processo das redes neurais é apreender um conhecimento ou relacionamento complexo e realizar a previsão de novas situações, e realizando o treinamento destas redes com sua própria arquitetura até que a mesma consiga apreender a solucionar o problema.

2.6.3 Técnica de árvores de decisão

O processo de árvores de decisão utiliza uma estrutura de classificação de nó onde cada um tem rótulo de classe, e a árvore sempre inicia por um único nodo, chamado como nodo-raiz e vai sendo dividida até levar a classe (CERQUEIRA, 2021).

2.7 Linguagem Python

A linguagem é completa e com isso aumenta a produtividade do programador, ao utilizar as bibliotecas usa-se programas desenvolvidos e testados por outros colaboradores, com isso há redução no número de erros (MENEZES, 2010).

Segundo Da Silva e Silva 2019, a linguagem Python é considerada de altíssimo nível ágil e elegante sendo desenvolvida por meio de uma metodologia focada em RAD (Rapid Application Development) desenvolvimento rápido de aplicações e na qualidade do código.

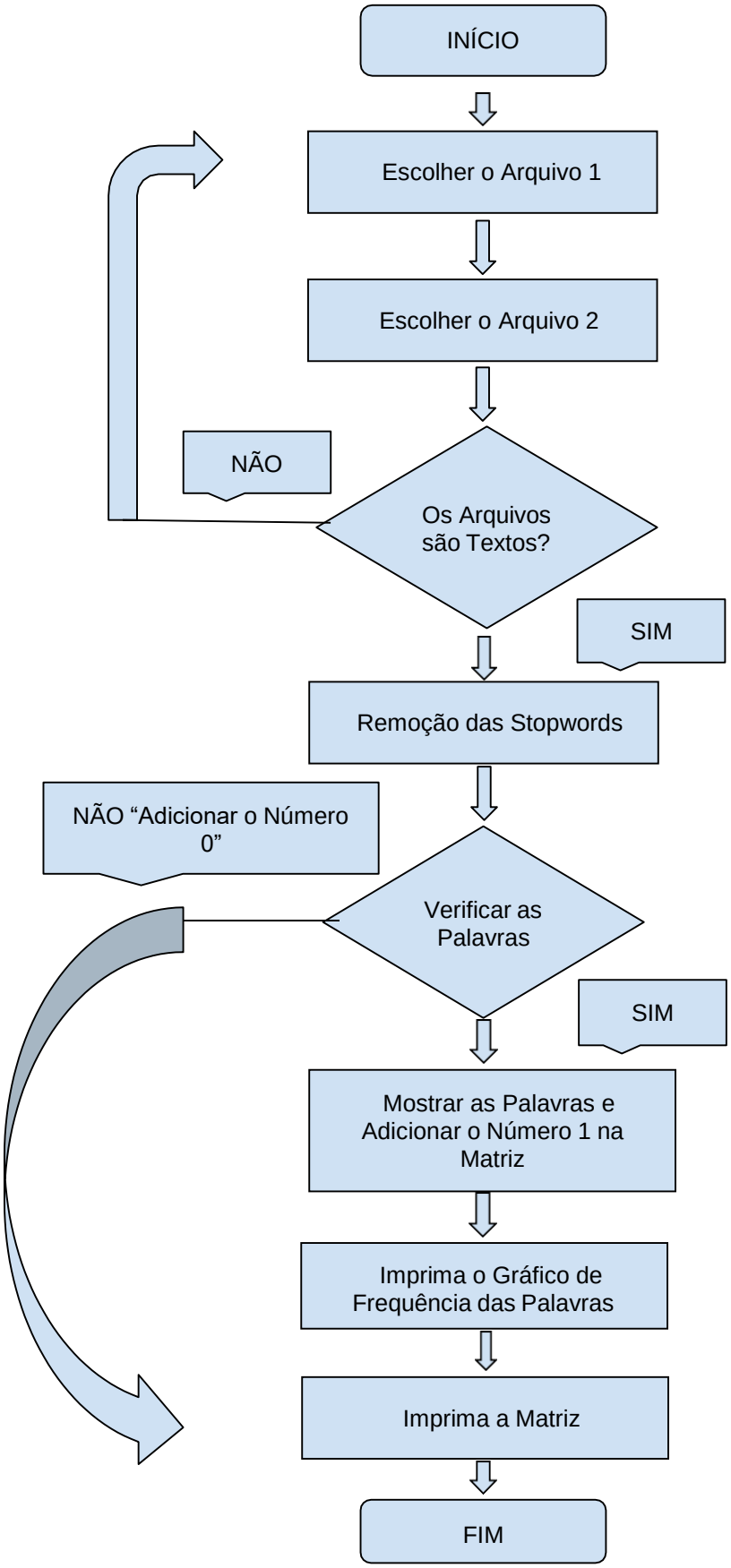
Conforme Bandeira 2019, Python utiliza a programação orientada à objetos possui uma gramática simples e dinâmica, com isso o programador tem a liberdade de decidir a melhor forma de resolver os problemas.

3 MATERIAIS E MÉTODOS

O suporte metodológico utilizado no presente trabalho iniciou com a definição dos descritores, tendo em vista a identificação de corpus textual do gênero notícias, a partir de verbos dicendi. Neste trabalho adotou-se como corpus, textos do gênero notícias, e padronização de descritores a partir da ocorrência de verbos dicendi (PEREIRA, 2016). A análise lexical (determinação da frequência das palavras) foi realizada a partir do algoritmo elaborado em linguagem Python, cuja saída fornecida é a matriz de ocorrência. A pesquisa caracteriza-se como quali-quantitativo, de cunho descritivo (GIL, 2002).

Foi utilizado o software Visual Studio Community 2019, v. 16.8.2, para o desenvolvimento do algoritmo de análise dos dados, utilizando a linguagem em Python. A interface web foi desenvolvida em linguagem PHP 7.4 e a aplicação foi realizada sob o sistema operacional Linux Ubuntu 20.04 lts. Para a realização da matriz das palavras foi utilizada a biblioteca em Python Numpy que é usada principalmente para realizar cálculos em Arrays. Neste trabalho foi utilizado o modelo de vetorização. Foram utilizadas as stopwords para a remoção de termos frequentes no texto, sem relevância para análise da mineração de texto com o objetivo de reduzir o número de palavras a serem analisadas no documento.

A definição do algoritmo está demonstrada no pseudónimo do código, abaixo.


INÍCIO
Escolher o Arquivo 1
Escolher o Arquivo 2
NÃO
Os Arquivos são Textos?
SIM
Remoção das Stopwords
NÃO “Adicionar o Número 0”
Verificar as Palavras
SIM
Mostrar as Palavras e Adicionar o Número 1 na Matriz
Imprima o Gráfico de Frequência das Palavras
Imprima a Matriz
FIM

3.1 O Corpus Textual

O corpus utilizado no presente artigo constitui-se de textos do gênero notícias coletadas por Pereira 2016, que em sua tese analisou a frequência de expressões multipalavras em um corpus de mais de 1 milhão de palavras através do uso do programa WordSmith Tools 6.0.

Perante Nascimento 2016, os verbos dicendi possuem o papel de ser um modalizador, como foi abordado nos textos o gênero notícia, o conteúdo aborda função argumentativa desses verbos já que direcionam a interpretação que o interlocutor fará do enunciado que esses verbos introduzem. Com isso os verbos dicendi expressam o real valor que realmente determina as intenções do enunciador, dentre os verbos são divididos em nove áreas semânticas, e essas áreas abrangem verbos de significados mais gerais e outros mais específicos como os citados abaixo:

- De dizer (afirmar,declarar);
- De perguntar (indagar,interrogar);
- De responder (retrucar,replicar);
- De contestar (negar,objetar);
- De concordar (assentir,anuir);
- De exclamar (gritar,bradar);
- De pedir (solicitar,rogar);
- De exortar (animar,aconselhar);
- De ordenar (mandar,determinar).

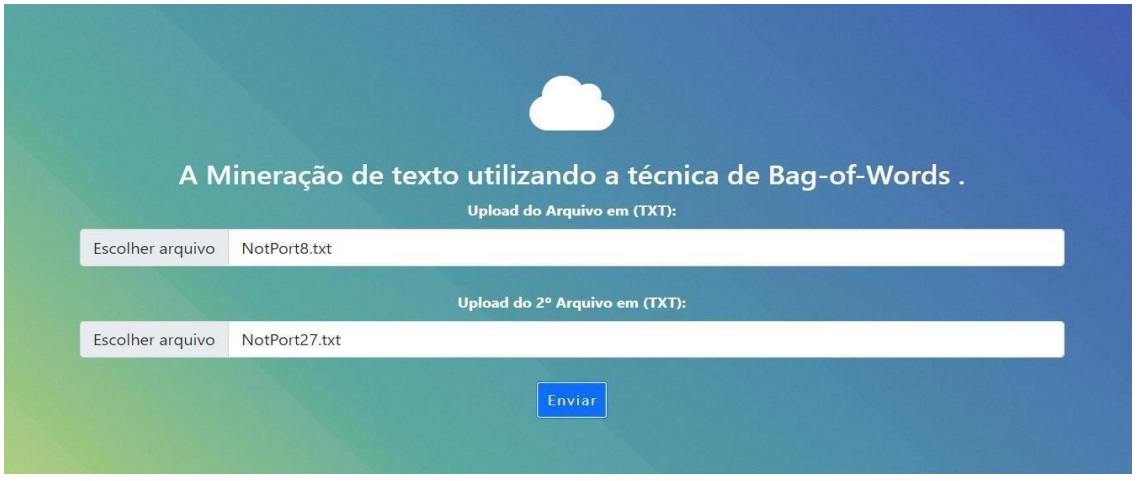
Neste trabalho, foram utilizadas nove notícias a título de amostra, porém buscando analisar a frequência de verbos dicendi que são característicos desse gênero literário, corroborando com Pereira 2016, quando pontua que sempre os profissionais de notícias buscam relatar as falas ditas pelos atores envolvidos na narrativa dos fatos. Assim, os verbos dicendi analisados foram: “falar”, “relatar”, “declarar”, “dizer”, “informar” e “concluir”. Vale ressaltar que, a busca pelavariável verbos dicendi, é utilizado como forma de detecção da utilização dessa característica para identificação dos gêneros em questão.

4 RESULTADOS E DISCUSSÃO

Como resultados, apresenta-se na Figura 5, o *layout* da interface de usuário desenvolvido como ferramenta de mineração de texto, bem como as análises realizadas nesta pesquisa piloto. Para o upload de arquivos, uma vez que o arquivo é selecionado no computador é realizado o upload deste e enviado para análise do texto. No estudo de Rosaine et. al, 2019 foi desenvolvida uma ferramenta semelhante com objetivo de extração de palavras em determinados textos.

Em um estudo semelhante de Klemann et. al, 2012, descreve uma ferramenta chamada Sobek que foi desenvolvida pela Universidade Federal do Rio Grande do Sul para busca textual, esta pode ser executada em computadores com diferentes sistemas operacionais *Linux, Windows* ou *Mac OS*, podendo ser utilizada sem maiores restrições, contudo não está disponível online.

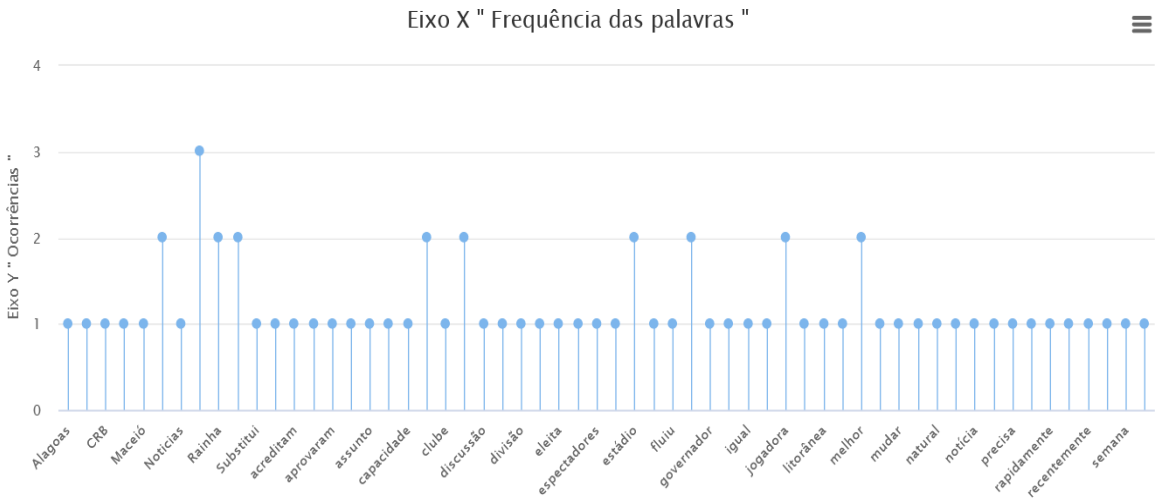
Figura 5 - Página inicial da ferramenta de busca



Fonte: O autor.

Uma vez determinada a frequência das palavras no texto pode-se verificar quais assuntos estão sendo abordados e com isso identificar o contexto presente no texto selecionado. A Figura 6, apresenta a representação gráfica da frequência de palavras utilizando um fragmento textual do corpus organizado por Pereira (2016).

Figura 6 - Representação Gráfica da Frequência De Palavras



Fonte: O autor

A Figura 6, foi elaborada para descrever a frequência das palavras em um gráfico de linhas onde o eixo “y” identifica as ocorrências e o eixo “x” a frequência das palavras. Já na Figura 7 é apresentada a nuvem de palavras escalonadas em ordem frequência.

Figura 7- Nuvem de palavras por frequência

Fonte: O autor.

A Ferramenta computacional desenvolvida encontrou 58 palavras em destaque no texto 1. Dentre os assuntos que foram selecionados, as palavras que apresentaram maior destaque são “Pelé”, “Marta” e "Futebol", configurando assim uma notícia relacionada ao exporte. Um estudo de Gil (2016), semelhante e utilizando um comparativo entre ferramentas de análise de texto através da semântica das palavras usando como exemplo o software TagCrowd com o mesmo objetivo de gerar nuvem de palavras e destacando-as.

Através da técnica da ferramenta TagCrowd encontrou a extração de 50 termos em destaque na figura 8, utilizando o texto 1.

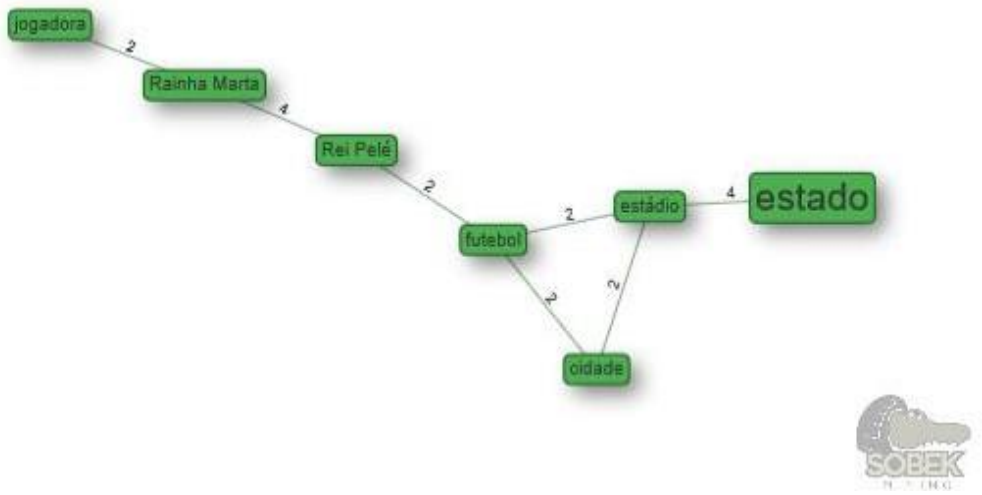
Figura 8 - Ferramenta TagCrowd

Fonte: O autor.

Segundo Klemann et. al, 2012, TagCrowd é uma ferramenta online que permite criar nuvens de palavras em diferentes línguas como o inglês e foi desenvolvida por Daniel Steinbock (Stanford University Califórnia – North América).

Através da técnica da ferramenta Sobek encontrou-se a extração de 7 termos em destaque na Figura 9, utilizando o texto 1. Apresentando os resultados obtidos pela aplicação da ferramenta Sobek. Observa-se que o resultado possui quarenta três palavras a menos que a ferramenta TagCrowd e cinquenta e um referente a ferramenta computacional desenvolvida.

Figura 9 - Ferramenta Sobek



Fonte: O autor.

Observa-se na Tabela 1, as ferramentas Sobek e a ferramenta computacional apresentam-se uma série de estatística relativas ao uso de frequência de palavras e termos e geração de gráficos e já TagCrowd não realiza a frequência dos termos.

Tabela 1 - Análise comparativa de Mineradores

Ferramentas	Online	Contagem de termos	Frequência dos termos	Apresentar os termos relevantes	Visualizar gráfico dos termos
TagCrowd	X			X	X
Sobek	X		X	X	X

Ferramenta Computacional	X	X	X	X	X

Fonte: O autor.

Ao analisar as Figuras 7, 8 e 9, alguns resultados demonstraram-se importantes quando comparados a diferentes ferramentas de mineração, como visualizados na Tabela 2, referente ao texto1.

Tabela 2 - Frequência de palavras em diferentes ferramentas.

PALAVRAS T1	**TAGCLOWD**	**SOBEK**	**FERRAMENTA COMPUTACIONAL**
Rei Pelé	3	2	3
Rainha Marta	2	4	2
Jogadora	2	2	2
Futebol	2	2	2
Estádio	2	2	2
Estado	3	4	3
Cidade	2	2	2
Decisão	2	0	2
Aprovada	2	0	2

PALAVRAS T1 – Palavras analisadas no Texto 1*.

Fonte: O autor.

Nota-se que dois termos que constam na ferramenta TagClowd e na Ferramenta Computacional desenvolvida, não constam na ferramenta Sobek, não sendo detectados em comparação a ferramenta desenvolvida. Já a ferramenta TagClowd apresentou os mesmos resultados com os termos com maior relevância no texto 1 e o número de frequência das palavras.

Tabela 3 - Matriz de termos / frases

Matriz para a frequência das palavras do texto 1:

	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
Alagoas	0	0	0	0	1	0	0	0	0	0	0	0
Brasil	0	0	0	0	0	0	1	0	0	0	0	0
CRB	0	0	0	0	0	0	0	0	0	0	1	0
FIFA	0	0	0	0	1	0	0	0	0	0	0	0
Maceió	0	0	0	0	1	0	0	0	0	0	0	0
Marta	1	0	0	0	1	0	0	0	0	0	0	0
Noticias	0	0	1	0	0	0	0	0	0	0	0	0

Fonte: O autor.

A Tabela 3, demonstra a criação de uma matriz de termos por frases, nesta matriz cada palavra tem um índice que corresponde a uma linha e cada frase corresponde a uma coluna. E cada célula corresponde ao número de ocorrência das palavras.

Tabela 4 - Matriz Termo/Documento

	DENOMINAÇÃO DOS DOCUMENTOS UTILIZADOS								
TERMOS	**D1**	**D2**	**D3**	**D6**	**D12**	**D13**	**D17**	**D18**	**D24**
Falar	0	1	0	0	0	0	0	0	0
Relatar	0	0	1	0	0	0	0	0	0
Declarar	0	0	0	0	1	0	0	0	0
Disse	1	1	1	1	1	0	0	0	1
Informou	0	0	0	0	0	1	0	0	0
Concluiu	0	0	0	0	0	0	1	1	0
Afirmou	0	0	0	0	0	0	0	0	0
Comentou	0	0	0	0	0	0	0	0	0
Contou	0	0	0	0	0	0	0	0	0
Apresentou	0	0	0	0	1	0	0	0	0
Terminou	1	0	0	0	0	0	0	0	0

Numeração do documento utilizado $D^{1\text{-}24*}$.

Fonte: O autor.

A Tabela 4, analisa os verbos dicendi comparando a aparição dos verbos diante das nove notícias selecionadas, com isso na tabela pode-se identificar a variação do número de vezes em que o verbo apareceu em cada documento analisado.

5 CONCLUSÃO

Conclui-se desta forma que a ferramenta apresenta uma interface computacional amigável, capaz de realizar mineração textual, com foco na obtenção da matriz termo/ocorrências. Desenvolvida em linguagem Python, a aplicação baseia-se na análise de textos. Sua interface gráfica tem como finalidade produzir como saídas uma matriz de ocorrências e a representação em nuvem de palavras, gráficos de frequência de palavras, além de realizar a limpeza de dados, a fim de otimizar a busca textual.

No que tange às características, por se web, de forma gratuita esta pode ser utilizável em qualquer plataforma e apresentar compatibilidade com qualquer sistema operacional, podendo tornar mais viável e precisa a busca lexicográfica. Quando comparada com outras ferramentas desenvolvidas de forma semelhante, a ferramenta em questão desenvolvida demonstrou-se equivalente, no entanto apresentou um diferencial ao gerar matriz de ocorrências de palavras/termos.

Por se tratar de uma fase de protótipo, estudos comparativos de desempenho relativos a outros algoritmos serão objeto de estudos futuros.

REFERÊNCIAS

ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster Analysis**. Beverly Hills, CA: Sage, 1984. 88 p.

AMO, Sandra de. **Técnicas de Mineração de Dados.** In XXIV Congresso da Sociedade Brasileira de Computação, vol. 1–1, Jul/Ago, 2004. 38, 39, 40.

BARBOSA, Júlio César. **Mineração de texto: uso de técnicas de processamento de linguagem natural para suporte à geração de projeções baseadas em opiniões do consumidor**. 2018. Tese de Doutorado. Mestrado em Sistemas de Informação e Gestão do Conhecimento. 2018.

BARBOSA, Maria Lúcia; SEVERO, Carlos Emilio Padilla; REATEGUI, Eliseo. Mineração de padrões no gênero textual blog. **RENOTE**, v. 7, n. 3, p. 581-590, 2009.

BOENTE, A. N. P.; OLIVEIRA, F. S. G. e ROSA, J. L. A. Utilização de Ferramenta de KDD para Integração de Aprendizagem e Tecnologia em Busca da Gestão Estratégica do Conhecimento na Empresa. **Anais do Simpósio de Excelência em Gestão e Tecnologia**, 1, 123-132, 2007.

BANDEIRA, Lucas Gabriel Coliado. **Fluxo de distribuição e documentação da biblioteca Python Magpylib**. 2019. 66 p. Trabalho de Conclusão de Curso (Eng. Eletrônica) - Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina. 2019.

BEYSOLOW II, Taweh. **Applied Natural Language Processing with Python: Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing**. Apress, 2018.

CANTO, Lucas Gama. **Análise de notícias do mercado financeiro utilizando processamento de linguagem natural e aprendizado de máquina para decisões de swing trade**. 2020. 54 p. Tese de Doutorado. Universidade Federal do Rio de Janeiro/RJ. 2020.

CERQUEIRA, Sarah Pereira. ESTUDOS DE TÉCNICAS PARA PROCESSAMENTO DE LINGUAGEM NATURAL. **Anais dos Seminários de Iniciação Científica**, n. 23, 2021.

DA SILVA, Rogério Oliveira; SILVA, Igor Rodrigues Sousa. Linguagem de Programação Python. **TECNOLOGIAS EM PROJEÇÃO**, v. 10, n. 1, p. 55-71, 2019.

DA SILVA, Cassiana F. **Uso de Informações Linguísticas na etapa de préprocessamento em Mineração de Texto**. 2004. 109 p. Tese de Doutorado. Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio do Sinos, São Leopoldo (RS). 2004.

DE MEDEIROS, Wagner Oliveira; PINHO, Fabio Assis; CORREA, Renato Fernandes. APLICAÇÃO DE SOFTWARE DE MINERAÇÃO DE TEXTO NA REPRESENTAÇÃO DA INFORMAÇÃO DE OBRAS ARTÍSTICO-

PICTÓRICAS. **Logeion: Filosofia da Informação**, v. 6, n. 1, p. 146-170, 2019.

EVERITT, B.; LANDAU, S.; LEESE, M. Cluster Analysis. A Hodder Arnold Publication. **Willey, London**, 2001.

FERREIRA, Márcio Henrique Wanderley; CORREA, Renato Fernandes. Mineração de textos científicos: análise de artigos de periódicos científico brasileiros da área de Ciência da Informação. **Em Questão**, v. 27, n. 1, p. 237-262, 2021.

FIORIO, Rosaine et al. Linguisticun: Uma Ferramenta de Auxílio ao Ensino da Língua Portuguesa e à Linguística Computacional. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2019. p. 11.

GIL, A. C. Como elaborar projetos de pesquisa brasileiros da área de Ciência da Informação. **Em Questão**, v. 27, n 1, p. 237 – 262, 2021. 4. ed. São Paulo: **Atlas**, 2002.

GIL, Carmem Zeli Vargas; SEFFNER, Fernando. Dois monólogos não fazem um diálogo: jovens e ensino médio. **Educação & Realidade**, v. 41, p. 175-192, 2016.

GIL, Ricardo Dacól. **Desenvolvimento de um sistema de análise de semântica latente para avaliar produções textuais**. 2017. 70 p. Trabalho de Conclusão de Curso (Ciência da Computação). Universidade de Caxias do Sul/RS, 2017.

JUNIOR, João Ribeiro Carrilho. **Desenvolvimento de uma metodologia para mineração de textos.** 2007. 96 f. Dissertação (Mestrado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

JUSTINO, HARUAN MOSSATO; MORIMOTO, LUANA GUERRA DE OLIVEIRA; GOBBATO, LUANNA. **Sistema de classificação de notícias.** 2019. 111 p. Trabalho de conclusão de curso (Tecnologia em Análise e Desenvolvimento de Sistemas). Universidade Federal do Paraná/PR. 2019.

KLEMANN, Miriam; REATEGUI, Eliseo; RAPKIEWICZ, Clevi. Análise de ferramentas de mineração de textos para apoio a produção textual. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. 2012.

KULTZAK, Adriano Francisco. **Categorização de textos utilizando algoritmos de aprendizagem de máquina com WEKA.** 2016. 74 f. Trabalho de Conclusão de Curso (Graduação) - Universidade Tecnológica Federal do Paraná, Ponta Grossa. 2016.

LUQUES, Ivair Nobrega. **Inteligência computacional aplicada a detecção intrínseca de plágio em documentos textuais**. 2020. 59 p. Dissertação (Mestrado). Centro Federal de Educação Tecnológica Celso Suckow da

Fonseca. Rio de Janeiro/RJ. 2020.

MARQUES, Elaine Cristina Moreira. **Redução de características baseada em grupos semânticos aplicados à classificação de textos.** 2018. 101 p. Dissertação (Programa de Pós-Graduação em Biometria e Estatística Aplicada) - Universidade Federal Rural de Pernambuco, Recife. 2018.

MARCUSCHI, Luiz Antônio. Gêneros textuais: definição e funcionalidade. **In: Gêneros textuais e ensino. 2. ed**. Ângela Paiva Dionísio, Ana Rachel Machado, Maria Auxiliadora Bezerra (Orgs). São Paulo: Parábola Editorial, 2003.

MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007.

MENEZES, Nilo Ney Coutinho. Introdução a programação com Python. **São Paulo: Novatec**, 2010.

NASCIMENTO, P.A; CANOSSA, I.A. A função modalizadora dos verbos dicendi no gênero textual notícia. **Revista Philologus**, Ano 22, N° 65. Rio de Janeiro: CiFEFiL, maio/ago.2016.
Disponível em: <http://www.filologia.org.br/rph/ANO22/65/002.pdf>.
Acesso em: 03 jan. 2022.

PEZZINI, Anderson. Mineração de textos: Conceito, processo e aplicações. **REAVI-Revista Eletrônica do Alto Vale do Itajaí**, v. 5, n. 8, p. 58-61, 2016.

PEREIRA, Aden R. **Análise contrastiva de verbos dicendi em textos jornalísticos de corpus paralelo português-espanhol à luz da Linguística de Corpus.** In: NADIN, Odair FERREIRA, Anise A. G. D.; FARGETI, Cristina M. (orgs.) Léxico e suas interfaces: descrição, reflexão e ensino. São Paulo/: Cultura Acadêmica, 2016. pp. 177-197.

PEREIRA, Aden Rodrigues. **Análise de base em córpus da tradução de expressões multipalavra no par linguístico português-espanhol.** 2016. 173 p. Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Programa de Pós-Graduação em Estudos da Tradução, Florianópolis, 2016.

ROSSI, Rafael Geraldeli. **Representação de coleções de documentos textuais por meio de regras de associação**. 2011. 159 p. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2011.
DOI: 10.11606/D.55.2011.tde-31082011-125648.

SCARPA, Alice Duarte. **Técnicas de processamento de linguagem natural aplicadas às ciências sociais**. 2017. 86 p. Dissertação (mestrado) – Fundação Getúlio Vargas, Escola de Matemática Aplicada. 2017.

SILVEIRA, Brunno Athayde; MURAMATSU, Thiago Yusuke; REVOREDO, Kate Cerqueira. **Análise do perfil de uma comunidade científica através de**

mineração de texto. 2011. 68 p. Monografia em Informática, Departamento de Informática Aplicada, Universidade Federal do Estado do Rio de Janeiro (Unirio), Rio de Janeiro/RJ, 2011.

SANTOS, Rafael et al. Conceitos de Mineração de dados na web. **XV Simpósio Brasileiro de Sistemas Multimídia e Web, VI Simpósio Brasileiro de Sistemas Colaborativos–Anais, MM Teixeira, CAC Teixeira, FAM Trinta, e P. PM Farias, Eds**, v. 1, n. 1, p. 81-124, 2009.

SOUZA, Adriano; FORTES, Reinaldo; LIMA, Joubert. OLAP Textual com Múltiplas Hierarquias de Tópicos e Rankings Segmentados. In: **Anais do XIII Simpósio Brasileiro de Sistemas de Informação**. SBC, 2017. p. 480-487.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. Introdução ao datamining: mineração de dados. 1ª edição. Rio de Janeiro. **Ed. Ciência Moderna**, 2009.