UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Tecnologia

PEDRO ARTICO RODRIGUES

SISBASETEXT: UM SISTEMA DE BUSCA PARA CRIAÇÃO SEMIAUTOMÁTICA DE BASES DE DADOS TEXTUAIS

LIMEIRA
2019

PEDRO ARTICO RODRIGUES

SISBASETEXT: UM SISTEMA DE BUSCA PARA CRIAÇÃO SEMIAUTOMÁTICA DE BASES DE DADOS TEXTUAIS

Monografia apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Bacharel em Sistemas de Informação.

Orientadora: PROF. DRA. ANA ESTELA ANTUNES DA SILVA
Coorientador: PROF. ME. PEDRO IVO GARCIA NUNES

LIMEIRA
2019

PEDRO ARTICO RODRIGUES

SISBASETEXT: UM SISTEMA DE BUSCA PARA CRIAÇÃO SEMIAUTOMÁTICA DE BASES DE DADOS TEXTUAIS

Monografia apresentada à Faculdade de Tecnologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Bacharel em Sistemas de Informação.

BANCA EXAMINADORA

_____

Prof. Dra. Ana Estela Antunes da Silva

_____

Prof. Dr. Celmar Guimarães da Silva

_____

Prof. Dr. Guilherme Palermo Coelho

LIMEIRA
2019

AGRADECIMENTOS

Dirijo os meus agradecimentos ao Prof. Dr. Celmar Guimarães da Silva e ao Prof. Dr. Guilherme Palermo Coelho pela disposição e gentileza de aceitarem o convite para compor a banca de meu trabalho de conclusão de curso. Agradeço também à Prof.ª Dr.ª Ana Estela Antunes da Silva e ao Prof. Me. Pedro Ivo Garcia Nunes pela paciente orientação e tolerância diante de minhas dificuldades e limitações.

RESUMO

A mineração de textos, também chamada de mineração de dados textuais, consiste em extrair padrões ou tendências de textos em linguagem natural visando aquisição de conhecimento específico. O processo de descoberta de conhecimento requer diversas etapas importantes, destacando-se a etapa inicial de coleta de informações e formação de uma base de dados textual. Em determinadas situações, não é possível utilizar bases previamente disponibilizadas na Web, pois a informação requerida tem caráter específico. Assim, as buscas precisam ser feitas manualmente na Web e toda a informação coletada deve ser analisada e organizada em atributos pertinentes ao domínio de estudo para que, posteriormente, as demais etapas da Mineração de Textos possam ser realizadas. Desse modo, essa etapa, muitas vezes, é considerada dispendiosa, já que faltam ferramentas deste tipo que contemplem a língua portuguesa. Por esse motivo, o presente trabalho objetiva o desenvolvimento de um sistema de busca Web dedicado à criação semiautomática de bases de dados textuais, a fim de oferecer suporte aos profissionais de mineração de textos diminuindo o tempo gasto na coleta de informações úteis e formação da base textual.

Palavras-chave: Mineração de textos, base de dados textual, busca Web, extração de conhecimento.

ABSTRACT

Text mining, also called textual data mining, consists of extracting patterns or trends from natural language texts for specific knowledge acquisition. The process of knowledge discovery requires several important steps, especially in the initial stages of information gathering and formation of the textual dataset. In certain situations, it is not possible to use datasets previously made available on the Web, as the information required may be specific. Thus, searches need to be done manually on the Web and all information collected must be analyzed and organized into attributes relevant to the domain of study so that later algorithms can be applied. Thus, this step is often considered expensive, as there are no tools of this type that contemplate the Portuguese language. For this reason, the present work aims to develop a web search system dedicated to the semi-automatic creation of textual datasets in order to support text mining professionals reducing the time spent collecting useful information and forming the textual dataset.

Keywords: *Text mining, textual dataset, web search, knowledge extraction*

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos a mineração de textos tem atraído o interesse não apenas dos pesquisadores em Ciência da Computação, mas também das empresas, que procuram extrair conhecimento a partir de textos com o objetivo de analisar informações sobre seus clientes, de modo a conquistar um melhor posicionamento no mercado de negócios.

O processo de formação de uma base textual envolve, em primeiro lugar, a busca e coleta de informações. A Web possui bilhões de páginas indexadas englobando todo tipo de informação disponível para os mais variados interesses. Os dados disponíveis na Web não são apenas numéricos, também são textuais. Dados textuais podem ser apresentados na forma de notícias em websites, blogs, fóruns, livros, comentários e posts em redes sociais, dentre outros (Popescu, 2007).

Grande parte dos textos contidos na Web está em formato não estruturado ou semiestruturado. Essas características podem dificultar a busca de informações relevantes em determinado contexto.

Uma maneira de facilitar essa busca é através do uso de *Web crawlers* (rastreadores Web, em língua portuguesa) da categoria universal. *Web crawlers* universais são o núcleo principal dos mecanismos de busca que percorrem o conteúdo da Internet automática e anonimamente, selecionando todas as páginas que encontram (Saini & Arora, 2016). Contudo, *crawlers* deste tipo apresentam limitações relacionadas às possibilidades de busca e à eficiência em buscar páginas que realmente estejam relacionadas ao interesse do usuário.

A partir das informações coletadas é possível criar as bases textuais. Existem ferramentas disponíveis, como *Google Cloud*[1] e *Microsoft Power BI*[2] que possibilitam a exportação dinâmica de dados e manipulação de atributos. Contudo, diversos recursos são limitados para versões gratuitas. Também é possível criar as bases manualmente através de planilhas do *Microsoft Excel*, porém, essa tarefa pode ser maçante se houver uma grande quantidade de informações.

Através dos pontos levantados, foi possível avaliar que um sistema que permita fazer buscas automáticas na Web de maneiras distintas acoplado à possibilidade de criar uma base textual com as informações coletadas pode trazer vários benefícios aos profissionais da área, pois além de diminuir o tempo gasto no processo de busca, cada base criada terá seus atributos devidamente organizados de maneira estruturada. Sendo assim, a

[1] Ferramenta disponibilizada pela empresa Google para criar bases textuais e armazená-las em nuvem (Google, 2019).
[2] È utilizada para criação de bases textuais e análise de dados (Microsoft, 2019).

execução de outras etapas de mineração de textos, como limpeza e transformação serão facilitadas.

1.1 Motivação

Há poucos estudos disponíveis na literatura que tratem da obtenção de informações para o desenvolvimento das bases textuais (Rouillard et al., 2016). Os estudos existentes apenas indicam que podem ser usados rastreadores Web para obter as informações mais rapidamente, entretanto, não são especificadas as técnicas de rastreamento nem como isso pode ser acoplado à criação das bases textuais.

Os demais estudos da área se concentram exclusivamente no desenvolvimento e aplicação de algoritmos de mineração de textos, assumindo que as bases textuais já existam ou possam ser obtidas trivialmente (Chen & Butte, 2016). Não foram encontrados estudos que discutam a possibilidade de realizar buscas Web associadas à criação de bases textuais.

Tendo em vista essa problemática, o estudo de uma solução que torne automática a busca Web e permita que a informação encontrada possa ser transportada para uma base textual precisou ser desenvolvido para que uma ferramenta pudesse ser criada.

Desse modo, propõe-se a união da busca e da criação de bases textuais em um único sistema Web, que chamaremos de SisBaseText. Esse sistema poderá ser utilizado por usuários com objetivos variados envolvendo tanto aplicações de mineração de textos quanto aprendizado de máquina.

1.2 Objetivos

O objetivo geral deste trabalho foi desenvolver uma versão beta de um sistema que facilite a busca de informações na Web e possibilite que elas sejam apresentadas em uma base textual. Os objetivos específicos são:

a) Utilizar métodos de rastreamento para realizar a busca Web;

b) Desenvolver uma expressão regular capaz de extrair conteúdo inteligível das páginas Web retornadas pelas buscas;

c) Aplicar técnicas de pré-processamento de texto ao conteúdo obtido por meio da busca Web.

d) Gerar uma base textual contendo as informações obtidas na busca;

e) Integrar os itens a), b), c) e d) em uma única ferramenta Web.

1.3 Estrutura do documento

Este documento está estruturado em Capítulos. No Capítulo 1, foram declarados o problema de pesquisa e os principais objetivos deste trabalho a partir da introdução de noções referentes à mineração de textos. O Capítulo 2 refere-se ao levantamento bibliográfico envolvido na pesquisa, enquanto o Capítulo 3 descreve os materiais e métodos que foram utilizados para a consecução dos objetivos. Os resultados do trabalho são apresentados no Capítulo 4, que se propõe a discuti-los. Finalmente, o Capítulo 5 se dedica à conclusão do trabalho mediante algumas considerações finais.

2 LEVANTAMENTO BIBLIOGRÁFICO

O desenvolvimento da ferramenta proposta neste trabalho demanda a compreensão de conceitos provenientes do levantamento bibliográfico em quatro temas principais: busca Web, expressões regulares, técnicas de pré-processamento de texto e bases textuais.

2.1 Busca Web

Os amplos mecanismos de busca Web, bem como muitos outros recursos de pesquisa especializada, são complexas ferramentas que permitem procurar documentos armazenados em sites, de acordo com critérios específicos (Miller & Bharat, 1998). Os mecanismos de busca são classificados da seguinte forma: baseados em humanos e em *Web crawlers*.

Os mecanismos de busca baseados em humanos indexam um conjunto de sites de interesse escolhidos manualmente por usuários através do envio de breves descrições de cada site ao servidor. Nesse tipo de mecanismo os resultados são afetados diretamente pela intervenção humana.

Por outro lado, os mecanismos de busca baseados em *Web crawlers* iniciam o processo de rastreamento a partir de algum localizador uniforme de recursos (*Uniform Resource Locator* – URL) conhecido, adicionando todos os hiperlinks da página a uma lista de URLs que serão visitados na ordem em que foram encontrados. O objetivo desses mecanismos é coletar o máximo de páginas úteis no menor tempo possível.

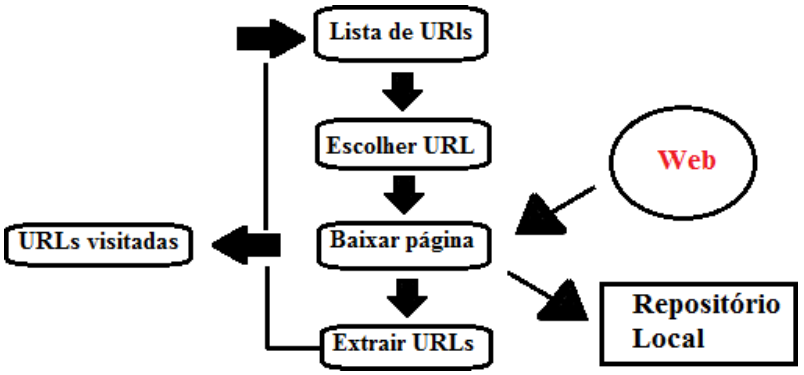
Em algumas aplicações, os *crawlers* gastam muito tempo procurando páginas em milhares de servidores, ou seja, questões de flexibilidade, robustez e capacidade de gerenciamento devem ser levadas em consideração. Além disso, desempenho, recursos de rede e os limites do sistema operacional são de grande importância para a obtenção de alto desempenho em termos de tempo computacional (Miller & Bharat, 1998).

O entendimento sobre *Web crawlers* envolve o estudo e compreensão de alguns conceitos relacionados à sua arquitetura e tipos de *crawler*. Esses conceitos são apresentados a seguir.

2.1.1 Arquitetura básica de Web crawlers

Crawlers necessitam de um conjunto de URLs como entrada e obtêm uma coleção de páginas como saída. Não é necessário ter apenas boas estratégias de rastreamento, mas também uma arquitetura eficiente. A Figura 2.1 apresenta a arquitetura básica de *Web crawlers.*

Figura 2.1 – Arquitetura básica de *Web crawlers*



Fonte: Adaptado de Kumar, Bhatia e Rattan (2017)

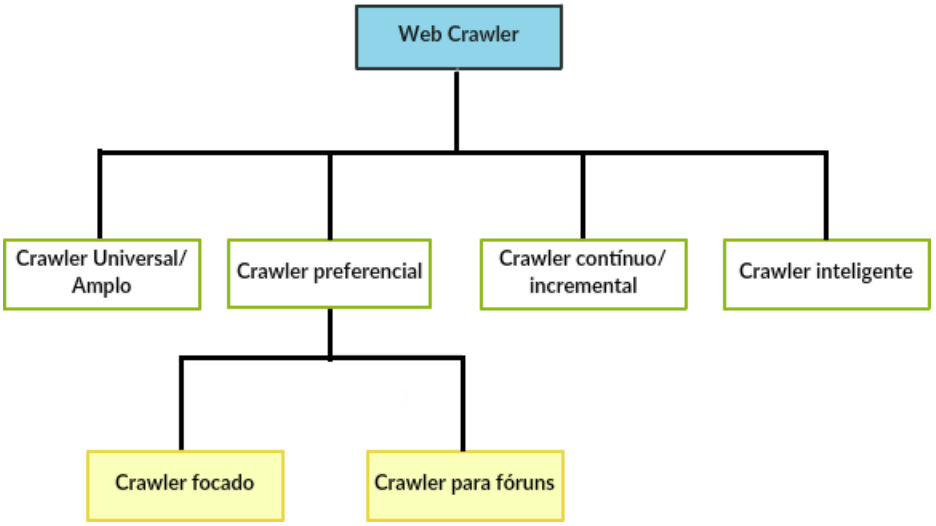
Conforme mostrado na Figura 2.1, um *Web crawler* inicia o processo de rastreamento buscando páginas e, ao visitá-las, baixa seu conteúdo e o armazena no repositório local. As páginas visitadas têm seus links extraídos e armazenados na lista de URLs. Os URLs da lista são visitados em uma ordem de escolha pré-estabelecida e, após isso, colocados na lista de URLs visitados (Baeza-Yates et al., 2005). Este processo é repetido recursivamente até que a lista de URLs esteja vazia.

A arquitetura apresentada possui os componentes mínimos de um *crawler*: lista de URLs, mecanismo para baixar páginas, repositório local e extrator de URLs. Entretanto, podem ser feitas modificações de acordo com as necessidades de rastreamento, através da adição e combinação de componentes existentes que respeitem os requisitos estabelecidos (Baeza-Yates & Castillo, 2004).

2.1.2 Tipos de *Web Crawlers*

Segundo Kumar, Bhatia e Rattan (2017), na literatura não há um modelo de classificação de *Web crawlers* pré-definido. Por esse motivo, os autores criaram uma taxonomia reunindo os tipos convencionais. Essa taxonomia é representada pela Figura 2.2.

Figura 2.2 – Taxonomia de *Web Crawlers*



Fonte: Adaptado de Kumar, Bhatia e Rattan (2017)

Web crawlers universais (ou amplos) são responsáveis por buscar páginas sem estabelecer um limite para determinado tópico ou domínio. Ao encontrar uma página, continuam seguindo todos os URLs de cada uma das páginas obtidas. É o tipo mais simples e primitivo, não apresentando alta eficiência em tempo computacional, pois perde muito tempo visitando links inúteis (Bhatt, Vyas & Pandya, 2015).

A categoria de *Web crawlers* preferenciais busca páginas mediante tópicos. Eles não rastreiam todos os links encontrados, em vez disso, o usuário escolhe um tópico de interesse e o rastreador estima a relevância de cada página (Gaur & Sharma, 2014). Existem duas subcategorias de *crawlers* preferenciais: focados e para fóruns (Bhatt, Vyas & Pandya, 2015).

Os *crawlers* focados são agentes responsáveis por coletar páginas relacionadas a um tópico específico, priorizando a extração e exploração de URLs. O rastreamento pode ser feito através de temas, palavras-chave, diretórios da Web e índices de textos que são utilizados para calcular a probabilidade de uma página ser relevante antes de baixá-la. Nesse tipo de *crawler* pode-se aplicar algoritmos de busca, como *Breadth-First* e *Fish-search* durante o rastreamento (Yanuar-Firdaus & Suryani, 2012). O primeiro algoritmo organiza os URLs em uma árvore e começa a busca seguindo o primeiro URL (nó raiz) em direção aos nós vizinhos. Se ele corresponder à busca, então a página tem seu conteúdo baixado no repositório local, caso contrário, o processo se repete pelos demais níveis da árvore. Já o segundo verifica todos os

URLs contidos em cada página e um índice de relevância é calculado. Somente páginas com alto índice têm todos os URLs visitados, evitando assim que informação irrelevante seja capturada.

Crawlers para fóruns realizam o rastreamento a partir da classificação de URLs em diferentes grupos: URL de índice, URL de encadeamento e URL de indexação. São capazes de detectar o tipo de URL e obter todo o conteúdo da página, considerando o encadeamento imposto pela estrutura do fórum. Logo, um único URL pode extrair informações contidas em várias páginas, sem ser necessário visitar todas.

Crawlers incrementais são usados para atualizar, de forma seletiva, uma coleção de páginas armazenadas localmente. As atualizações visam substituir páginas antigas por páginas novas e mais importantes. Esse rastreador é capaz de estimar a frequência com que as páginas mudam, revisitando unicamente as que foram alteradas, ao invés de atualizar o conjunto inteiro (Cho & Garcia-Molina,1999).

Os *crawlers* inteligentes seguem duas abordagens temáticas distintas. A primeira visa escolher os melhores URLs por intermédio de uma lista de URLs, de maneira a descartar páginas irrelevantes. Essa abordagem é chamada de aprendizado supervisionado, e requer um conjunto de URLs como dados de treinamento, para que o rastreador aprenda a decidir quais são relevantes. Neste tipo de aprendizado, os URLs que possuem palavras-chave específicas são utilizados, de modo que o *crawler* seja capaz de encontrá-las em URLs de outras páginas. A segunda, denominada aprendizado não supervisionado, utiliza métricas para avaliar a importância de cada URL da lista, de modo a organizá-los em ordem decrescente, visitando os mais importantes primeiro (Sahu & Bharne, 2016).

2.2 Expressões regulares

O conteúdo obtido em uma busca por meio de técnicas de rastreamento Web não é inteligível, a priori, por conter diversas *tags* HTML, *scripts*, folhas de estilo, marcadores e outros recursos que constituem páginas Web. Sendo assim, é necessário aplicar estratégias para a obtenção de textos puros e passíveis de entendimento e compreensão.

Uma das estratégias empregadas é o uso de expressões regulares, isto é, padrões compostos por metacaracteres que, associados a caracteres literais, compõem uma expressão (Firoiu, Oates & Cohen, 2002). A expressão é analisada por um processador de expressões regulares e retorna verdadeiro, caso o texto atenda às suas condições.

As expressões regulares são criadas através de linguagem formal e podem ser associadas para formar padrões de caráter complexo, como expressões aritméticas, por exemplo. De maneira geral, pode haver mais de uma expressão regular para obter o mesmo texto, pois a sintaxe e os operadores diferem entre implementações (Fernau, 2005).

A sintaxe das expressões regulares envolve a manipulação de caracteres não alfabéticos, denominados metacaracteres. Os metacaracteres são classificados em: especificadores, quantificadores, âncoras e agrupadores (Fernau, 2005).

Os especificadores são responsáveis por descrever o conjunto de caracteres permitidos na expressão regular criada. A Tabela 2.1 contém uma lista com alguns especificadores e suas respectivas finalidades.

Tabela 2.1 – Especificadores

Metacaractere	Significado	Nome
.	Simboliza quaisquer caracteres, com exceção da quebra de linha.	Curinga
[ ]	Simboliza caracteres incluídos no conjunto.	Conjunto
[^]	Simboliza caracteres não incluídos no conjunto.	Negação do conjunto
\d	Apenas números são permitidos no conjunto.	Números
\D	Apenas números não são permitidos no conjunto.	Exceto números
\w	Caracteres numéricos e letras são permitidos.	Alfanuméricos

Fonte: Adaptação de Denis, Lemay & Terlutte (2004)

Os quantificadores têm a função de definir o número de caracteres permitidos no texto para encontrar uma correspondência. A Tabela 2.2 contém uma lista com quantificadores e suas respectivas finalidades.

Tabela 2.2 – Quantificadores

Metacaractere	Significado
*	Indica 0 ou mais correspondências.
+	Indica 1 ou mais correspondências.
?	Indica 0 ou 1 correspondência.
{n}	Indica *n* correspondências.
{x,y}	Significa no mínimo *x* correspondências e no máximo *y* correspondências.

Fonte: Adaptação de Denis, Lemay & Terlutte (2004)

Os metacaracteres âncoras estabelecem a posição exata em que uma correspondência ocorrerá, considerando o conjunto de caracteres existente. Quando usados, o processador de expressões regulares não percorre toda a cadeia de caracteres, buscando somente uma correspondência na posição definida. A Tabela 2.3 lista as âncoras mais usadas.

Tabela 2.3 – Âncoras

Metacaractere	Significado
^	Há correspondência exclusivamente no início da cadeia de caracteres.
$	Há correspondência exclusivamente no fim da cadeia de caracteres, antes de \n.
\A	A correspondência ocorre no início da cadeia de caracteres, não havendo opções multilinha.
\b	A correspondência ocorre no limite de uma palavra.
\G	A correspondência tem início na posição em que a correspondência anterior se encerrou.

Fonte: Adaptação de Denis, Lemay & Terlutte (2004)

Os metacaracteres de agrupamento definem grupos compostos por subexpressões regulares que podem ser testadas alternadamente. A Tabela 2.4 apresenta dois agrupadores convencionais.

Tabela 2.4 – Agrupamento

Metacaractere	Significado
|	Utilizado para definir alternativa, ou seja, tanto a expressão regular à esquerda quanto à direita.
()	Utilizado para definição de grupos em que os demais metacaracteres podem ser aplicados.

Fonte: Adaptação de Denis, Lemay & Terlutte (2004)

2.3 Técnicas de pré-processamento

A busca Web envolve a manipulação automática de conteúdo textual que é composto por palavras e recursos linguísticos. Essa manipulação depende do uso de técnicas de pré-processamento (Poteet & Kao, 2007).

O pré-processamento consiste na aplicação de tarefas que objetivam abstrair as estruturas do texto, de forma a obter apenas informações consideradas relevantes (Luengo, Guerrero & Herrera, 2014). Durante o processo, é feita a eliminação de conteúdo irrelevante e a padronização de termos, de modo que não haja interferências na análise das informações.

Pré-processar textos tende a ser um processo dispendioso, pois não existe somente uma técnica que possa ser aplicada para obter representações adequadas de determinado texto (Matsubara, Martins, & Monard, 2003). Além disso, em algumas situações, a aplicação incorreta das técnicas de pré-processamento pode provocar a perda de informações importantes presentes no texto.

Dentre as diversas técnicas que abrangem o pré-processamento, pode-se destacar as técnicas aplicadas neste trabalho: tokenização, remoção de *stop words* e sumarização.

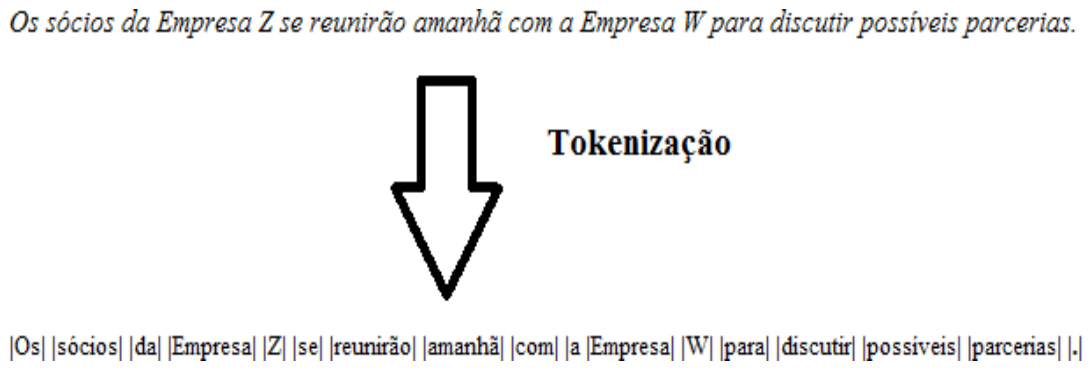
2.3.1 Tokenização

A tokenização (aportuguesamento da palavra *tokenization*) é um dos primeiros procedimentos do pré-processamento e sua aplicação tem por objetivo transformar o texto em apenas um conjunto de palavras (*tokens*), de acordo com estratégias previamente estabelecidas (Branco & Silva, 2006).

Uma das estratégias aplicadas é a divisão de um texto em seus delimitadores. Essa estratégia é largamente utilizada e demonstra ótimos resultados, se feita corretamente (Carvalho, et al., 2003). O delimitador mais conhecido e empregado é o espaço em branco, mas existem outros, como: () <>!-?.;'- "|.

A tokenização pode ser exemplificada pela Figura 2.3, que mostra a realização desse processo na frase "Os sócios da Empresa Z se reunirão amanhã com a Empresa W para discutir possíveis parcerias.". Neste exemplo, os tokens estão separados pelo caractere "|".

Figura 2.3 – Processo de tokenização



Fonte: Elaboração do autor

Neste trabalho, a tokenização será realizada por meio do uso do delimitador espaço em branco, e sua necessidade de aplicação refere-se à utilização posterior de outras técnicas de pré-processamento de textos, como a remoção de *stop words*.

2.3.2 Remoção de *stop words*

O pré-processamento também envolve uma etapa de identificação de conteúdo que pode ser desconsiderado, ou seja, termos que não possuem informação de maior relevância para o entendimento e compreensão geral do texto (Dias & Malheiros, 2004). Essa etapa é denominada remoção de palavras de parada (*stop words*, em língua inglesa).

Stop words são palavras que não possuem importância em um determinado contexto, sendo consideradas irrelevantes na análise textual, por não proverem informações significativas e apresentarem alta ocorrência. Essas palavras abrangem preposições, artigos, advérbios, pronomes, conjunções, consoantes e vogais (Guo et al., 2004).

Aproximadamente 50% das palavras de um texto são retiradas por meio de *stop lists* (Dias & Malheiros, 2004). *Stop list* corresponde a uma lista de palavras julgadas irrelevantes, podendo ser criada através de listas existentes, conforme a necessidade de aplicação e uso. A Tabela 2.5 apresenta um exemplo de *stop list* com 80 palavras em língua portuguesa.

Tabela 2.5 – Lista com *stop words*

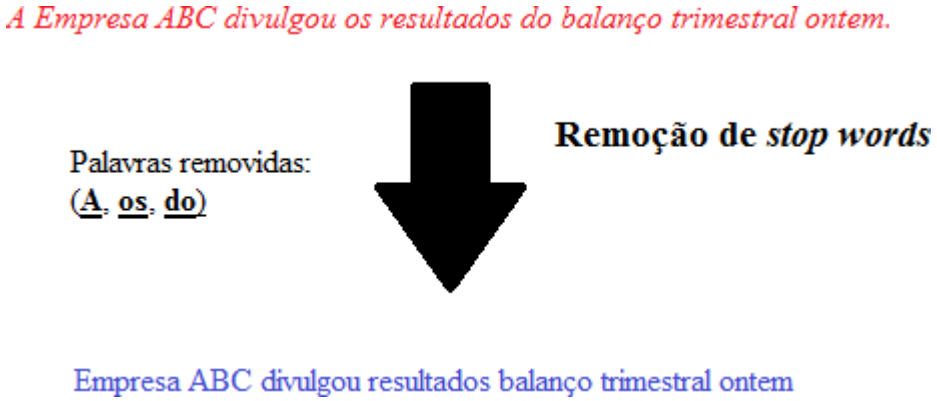
SÃO	DELA	SUAS	TODAVIA	DE
ERAM	ESTA	MEU	TAMPOUCO	A
SEJA	AQUILO	ÀS	ISSO	O
SERÁ	ISTO	MINHA	ATÉ	QUE
FOSSE	AQUELES	TÊM	AOS	E
TEM	AQUELAS	NUMA	SIDO	DO
TERÃO	ESTES	PELOS	CONTUDO	EM
TIVERAM	ESTAS	ELAS	DISSO	UM
HOUVE	AQUELA	HAVIA	DISTO	PARA
HÁ	AQUELE	SERIA	TEVE	É
NOSSAS	DELAS	QUAL	TIVE	COM
NOSSOS	QUANDO	NÓS	FUI	NÃO
NOSSA	MUITO	TENHO	HAJA	UMA
NOSSO	JÁ	LHE	ESTAVA	OS
VOCÊS	SER	DELES	TEU	NO
TUA	EU	ESSAS	DEVIA	SE
TEUS	DEPOIS	ESSES	DIZEM	NA
TUAS	SEM	PELAS	ESTEJA	POR
MINHAS	MESMO	ESTE	AO	MAIS
MEUS	NEM	DELE	NUNCA	AS

Fonte: Adaptação de Porter (2001)

O processo de remoção de *stop words* não possui grande complexidade, em termos de programação computacional, pois diversas linguagens oferecem os recursos necessários, tais como Python, PHP, Java e C++. A remoção de *stop words* pode ser exemplificada pela Figura 2.4, que mostra a realização desse processo na frase “A Empresa ABC divulgou os resultados do balanço trimestral ontem.”.

A remoção de *stop words* será empregada neste trabalho em dois momentos distintos. Primeiro, nas palavras-chave inseridas pelo usuário, de modo a evitar que páginas irrelevantes sejam retornadas na busca. Segundo, durante a criação da base textual, quando o atributo principal definido for “palavra”, conforme é explicado no Capítulo 3 (Materiais e métodos).

Figura 2.4 – Remoção de *stop words*



Fonte: Elaboração do autor

2.3.3 Sumarização

Sumarizar é o processo de seleção das informações mais importantes de um texto e é realizada a fim de resumir e enfatizar as ideias nele contidas. A Sumarização Automática é uma subárea do Processamento de Língua Natural que objetiva à produção automática de sumários a partir de um ou vários textos (Teufel & Moens, 2002). Os sumários, popularmente conhecidos como resumos, podem ser criados a partir de várias estratégias e pelo uso de conhecimentos de naturezas diversas.

A sumarização automática surgiu para tornar o processo de resumo mais eficiente e prático. Com o advento da Web, o repertório de textos produzidos tornou-se ainda mais amplo, incorrendo em uma quantidade maior de dados informativos a serem pesquisados.

Os sumários podem ser classificados de acordo com diferentes aspectos, sendo um deles o modo como são obtidos. Sparck Jones (1993) classifica-os como *extracts* ou *abstracts*, sendo que ambos os termos referem-se à Sumarização Extrativa e Sumarização Abstrativa, respectivamente.

A sumarização extrativa gera um sumário através da seleção de sentenças representativas do texto original, sem modificá-las. A seleção é feita através de um mecanismo de ranqueamento que visa obter as sentenças com as melhores pontuações, ou seja, as mais importantes (Pardo, 2008).

Por outro lado, a sumarização abstrativa analisa os documentos e automaticamente gera novas sentenças. Esta abordagem tenta produzir novos textos a partir dos fragmentos originais identificados como importantes (Cardoso, Pardo & Nunes, 2001).

A sumarização extrativa será aplicada neste trabalho, quando o usuário desejar que o conteúdo retornado contenha apenas as informações relevantes.

2.4 Bases textuais

As bases textuais podem ser definidas como coleções de dados textuais tabulados. Cada coluna da tabela representa uma variável em particular, e as linhas correspondem a um determinado membro do conjunto de dados textuais em questão (Bordes et. al., 2012).

O uso de bases textuais, em geral, está relacionado à aplicação de tarefas de mineração de textos ou aprendizado de máquina, como: classificação, agrupamento, regras de associação, dentre outras. Para a aplicação de tais tarefas, são utilizados algoritmos como: árvores de decisão, Naïve Bayes, dentre outros.

Por essa razão, para que uma base textual possa ser considerada consistente para a aplicação de um algoritmo, é necessário que se utilize atributos (colunas) adequados à análise realizada. Além disso, é preciso que haja muitas informações contidas, ou seja, um número considerável de linhas na base (Imamura, 2000).

A Tabela 2.6 apresenta um exemplo de base textual relacionada a livros. Essa base poderia ser utilizada, por exemplo, para treinar um algoritmo de aprendizado de máquina cujo objetivo seja classificar trechos textuais em literários e não-literários.

Tabela 2.6 – Exemplo de base de dados textual

Id	Livro	Trecho	Autor	Ano	Classe
1	Helena	O Dr. Camargo, médico e velho amigo da casa, logo que regressou do enterro, foi ter com Estácio, a quem encontrou no gabinete particular do finado, em companhia de D. Úrsula.	Machado de Assis	1876	Literário
2	O grande livro da Palmirinha	Modo de Preparo No liquidificador, bata os ovos, o óleo, o leite, e a mistura de morango até ficar cremoso. Transfira para uma tigela, adicione a farinha, o fermento e misture com uma colher. Despeje em uma fôrma de 26 cm de diâmetro untada e enfarinhada. Leve ao forno médio, preaquecido, por 25 minutos ou até assar e dourar levemente. Deixe amornar e desenforme. Em uma panela, leve ao fogo baixo os ingredientes da cobertura até derreter. Regue o bolo com a cobertura e sirva em seguida.	Palmira Onofre	2014	Não-Literário
3	Macário	Satan: Onde vais?	Álvares	1855	Literário

		Macário: Sempre tu, maldito! Satan: Onde vais? Sabes de Penseroso? Macário: Vou ter com ele. Satan: Vai, doido, vai! que chegarás tarde! Penseroso morreu.	de Azevedo		
4	Os Lusíadas	Canto IV Despois de procelosa tempestade, Nocturna sombra e sibilante vento, Traz a manhã serena claridade, Esperança de porto e salvamento; Aparta o Sol a negra escuridade, Removendo o temor ao pensamento: Assi no Reino forte aconteceu Despois que o Rei Fernando faleceu	Luis de Camões	1572	Literário
5	O golpe de 64	Em 31 de março de 1964, as forças militares brasileiras deflagraram um golpe de Estado contra o presidente João Goulart e instalaram no país uma ditadura que durou duas décadas. A ruptura institucional não resultou de uma ação intempestiva. Foi antes o apogeu de um longo e cada vez mais acirrado confronto entre pessoas, partidos e movimentos com concepções divergentes sobre o futuro político e social do Brasil.	Oscar Pilagallo	2014	Não-Literário
6	História da Guerra Civil Americana	O conflito teve sua origem na controversa questão da escravidão, especialmente nos territórios ocidentais. As potências estrangeiras não intervieram na época. Após quatro anos de sangrentos combates que deixaram mais de 600 mil americanos mortos e destruíram grande parte da infraestrutura do sul do país, a Confederação entrou em colapso, a escravidão foi abolida, um complexo processo de reconstrução começou, a unidade nacional retornou e a garantia de direitos civis aos escravos libertos teve início.	John D. Wright	2008	Não-Literário
7	Negrinha	Nascera na senzala, de mãe escrava, e seus primeiros anos vivera- os pelos cantos escuros da cozinha, sobre velha esteira e trapos imundos. Sempre escondida, que a patroa não gostava de crianças.	Monteiro Lobato	1920	Literário
...	...	...	...	...	...
1200	Eu	Eu, filho do carbono e do amoníaco Monstro de escuridão e rutilância Sofro, desde a epigênese da infância A influência má dos signos do zodíaco."	Augusto dos Anjos	1912	Literário

Fonte: Elaboração do autor

3 MATERIAIS E MÉTODOS

A consecução dos objetivos deste trabalho envolve a proposta de uma ferramenta para a busca automática de informações textuais na Web e a criação de uma base textual. Dentre as atividades necessárias ao desenvolvimento desta ferramenta, destacam-se a modelagem, implementação, testes e validação. Essas atividades são apresentadas nos subcapítulos 3.1, 3.2 e 3.3, respectivamente.

3.1 Modelagem

A documentação e a modelagem do SisBaseText foram feitas a partir da *Unified Modeling Language* (UML), que pode ser entendida como uma linguagem que define vários artefatos gráficos para a elaboração padronizada de modelos de software (Booch, et al., 2005). Essa linguagem tem como objetivo a especificação, documentação e estruturação da parte lógica, tanto superficial quanto específica, sendo que as representações gráficas são feitas por meio de diagramas.

A UML dispõe de duas categorias de diagramas: comportamentais e estruturais. Os diagramas comportamentais têm por objetivo visualizar, especificar, construir e documentar aspectos dinâmicos de um sistema. Por outro lado, os diagramas estruturais estão relacionados aos aspectos estáticos do sistema.

Dentre os sete diagramas da categoria comportamental, foi elaborado apenas o diagrama de caso de uso. O diagrama de caso de uso representa o comportamento de uma ferramenta, por meio da apresentação de suas funcionalidades e dos atores que interagem com elas.

Na categoria estrutural foi elaborado o diagrama de classes. Este diagrama representa a estrutura de uma ferramenta por meio da apresentação dos componentes de software que a compõem. Esses componentes são chamados de classes e definem os atributos e operações executadas pela ferramenta.

3.2 Implementação

O desenvolvimento da ferramenta foi realizado através da codificação nas linguagens *Hypertext Preprocessor* (PHP) e Python. O PHP é a linguagem de código aberto mais utilizada no desenvolvimento Web (The PHP Group, 2019), enquanto o Python é uma linguagem de programação de alto nível, funcional, de tipagem dinâmica e forte, tendo como vantagens a facilidade de leitura e pouca escrita de código-fonte (Lutz, 2013).

A ferramenta é composta por seis módulos: *Busca Web*, *Gerenciamento de fontes*, *Definição do atributo principal*, *Seleção de conteúdo*, *Criação e edição da base* e *Exportação da base*.

O módulo *Busca Web* envolve a aplicação de dois tipos de *Web crawler*: focado (com a implementação do algoritmo *Fish Search*) e inteligente (aprendizado não-supervisionado). Ambos são apresentados no Capítulo 2. Foram desenvolvidos três tipos de busca: busca por URL, busca com ou sem fonte e busca em rede social. Cada uma delas com peculiaridades que contribuem para que a informação desejada possa ser encontrada.

A busca por URL deverá ser utilizada quando o usuário desejar que as informações de uma página específica componham a base textual. Para tanto, ele terá de preencher um campo com o URL de interesse e selecionar se deseja ou não que o conteúdo obtido seja sumarizado.

A busca com ou sem fonte visa procurar informações a partir de palavras-chave inseridas pelo usuário, considerando ou não uma fonte especifica de informação. Nessa busca, o usuário deve inserir as palavras-chave, selecionar um buscador (Google, Google News, Bing ou Bing Notícias), escolher uma data, selecionar uma opção relativa à sumarização e, opcionalmente, selecionar uma fonte.

A busca em rede social tem por objetivo procurar informações textuais especificamente em redes sociais, sendo *Twitter e Instagram* as disponíveis. Assim, o usuário deve inserir o perfil em que deseja buscar *posts*, selecionar uma das redes sociais disponíveis e escolher se deseja ou não sumarizar os resultados. Neste caso apenas os textos são salvos, as imagens são desconsideradas.

O segundo módulo envolve o gerenciamento de fontes de pesquisa, possibilitando o cadastro, consulta e exclusão das fontes que poderão ser usadas na busca Web. O nome da fonte e o URL são os parâmetros dessas operações.

Após ser realizada a busca Web, o usuário deverá escolher qual será o tipo do atributo principal da base. As opções disponibilizadas são: base de textos, base de sentenças e base de palavras. Caso a opção escolhida seja "base de textos", o conteúdo retornado pela busca será dividido em textos, de acordo com a respectiva fonte de cada página. Se a opção "base de sentenças" for selecionada, cada página terá seu conteúdo dividido em sentenças através do caractere ponto. Se o usuário escolher "base de palavras" o conteúdo será dividido em palavras e a remoção de *stop words* será aplicada.

No módulo “Seleção de conteúdo” o usuário será capaz de visualizar o conteúdo retornado pela busca. Esse conteúdo estará separado de acordo com a escolha feita no módulo anterior (textos, sentenças ou palavras). O objetivo deste módulo é que o usuário analise e selecione o conteúdo que julgar pertinente para compor a base textual.

O módulo de "criação e edição da base" é responsável por apresentar ao usuário a base criada, contendo as informações selecionadas no módulo anterior. Além do atributo principal "Conteúdo", os atributos ID, Data da Coleta, Sumarização, Tipo de base, Tipo de busca e URL são previamente fornecidos. O usuário é capaz de adicionar, remover ou editar atributos.

O módulo de exportação da base possibilita ao usuário fazer o *download* da base nos seguintes formatos de arquivo: “.txt”, “.csv” e “.xls”.

3.3 Testes e Validação

O sistema foi avaliado por meio de testes funcionais envolvendo algumas métricas de recuperação da informação. Testes funcionais são aqueles que verificam o comportamento das entradas e saídas de um sistema de informação, considerando os requisitos e ignorando a estrutura interna da ferramenta. As métricas de recuperação de informação utilizadas para avaliar o SisBaseText foram precisão e revocação.

A precisão é responsável por medir o número de documentos relevantes recuperados considerando-se o total de documentos recuperados, conforme a Equação 1. A precisão mede a eficácia do sistema em excluir documentos que não sejam

relevantes de acordo com a necessidade de informação. Uma precisão alta é desejável, pois indica que grande parte dos documentos recuperados pelo sistema são relevantes (Souza, 2006).

$$Precisão = \frac{documentos\ relevantes \cap documentos\ recuperados}{documentos\ recuperados} \quad (1)$$

Revocação pode ser definida como o número de itens relevantes recuperados considerando-se todos os itens relevantes de um determinado conjunto. Essa métrica indica a eficácia do sistema em recuperar documentos pertinentes. Uma revocação total (100%) ocorre quando todos os documentos recuperados são relevantes.

$$Revocação = \frac{documentos\ relevantes \cap documentos\ recuperados}{documentos\ relevantes} \quad (2)$$

4 RESULTADOS

Este capítulo tem o propósito de apresentar os resultados deste trabalho e foi organizada em três subcapítulos. O primeira se dedica à exposição dos resultados oriundos da atividade de modelagem do SisBaseText, enquanto o segunda se refere aos resultados pertinentes à atividade de implementação e o terceira diz respeito à validação do sistema.

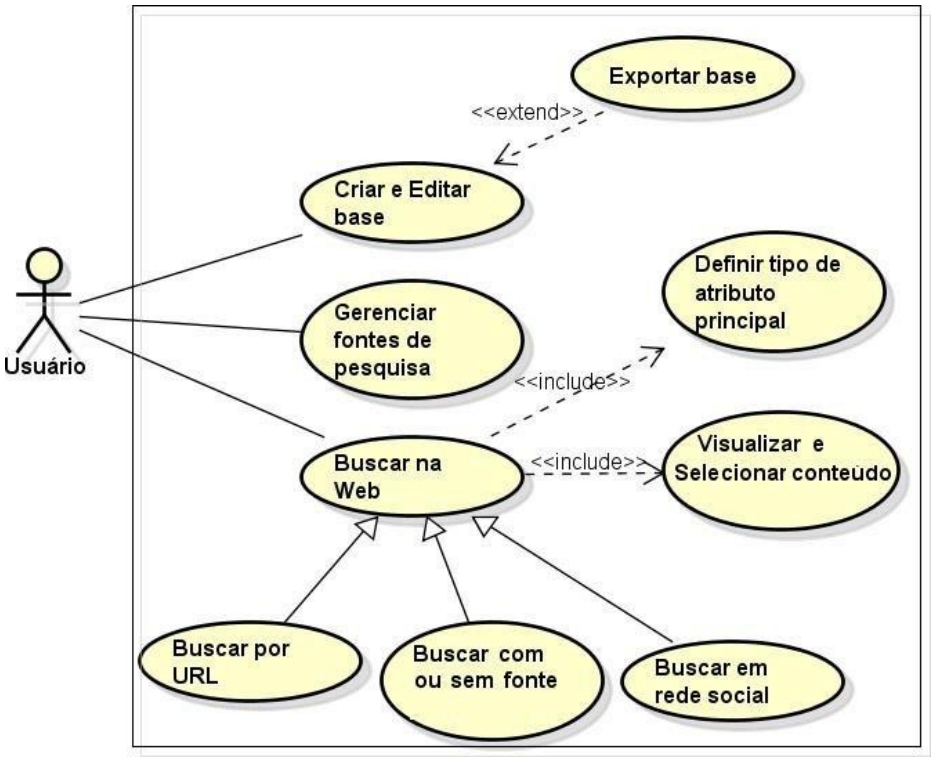
4.1 Resultados da modelagem

A etapa de modelagem consiste da análise e projeto do SisBaseText que ocorreu mediante a elaboração de diagramas UML. O Subcapítulo 4.1.1 apresenta o diagrama de caso de uso, enquanto o Subcapítulo 4.1.2 mostra o diagrama de classes.

4.1.1 Diagrama de Caso de Uso

O diagrama de caso de uso permite representar graficamente as funcionalidades que serão proporcionadas pela ferramenta, conforme apresentado na Figura 4.1. O SisBaseText possibilita ao usuário gerenciar fontes de pesquisa, fazer três tipos de busca na Web, definir o atributo principal da base textual, visualizar e selecionar conteúdo, criar, editar e exportar a base.

Figura 4.1 – Diagrama de Casos de Uso



Fonte: Elaboração do autor

4.1.2 Diagrama de Classes

Também foi elaborado um diagrama de classes que permite representar graficamente a estrutura do SisBaseText. A Figura 4.2 apresenta esse diagrama, que contém os componentes descritos a seguir.

SourceManagementGUI: Interface gráfica do gerenciamento de fontes de pesquisa. Apresenta os campos e botões necessários para as operações de cadastro, consulta e exclusão.

SourceManagementWindow: Classe Controle responsável por atualizar a lista de fontes a partir do arquivo de texto em que está armazenada.

SourceManagement: Classe Entidade cuja responsabilidade é o cadastro e exclusão de fontes de pesquisa.

SearchGUI: Interface gráfica da busca Web. Disponibiliza os campos e botões para cada um dos tipos de busca existentes.

SearchWindow: Classe Controle que verifica o tipo de busca escolhido pelo usuário e exibe mensagens quando há problemas nos campos inseridos. Além disso, carrega as fontes de pesquisa a partir do módulo de gerenciamento de fontes.

Search: Classe abstrata de busca Web. Possui métodos utilizados na busca que são herdados por suas classes derivadas.

URLSearch: Classe derivada da classe **Search** que realiza a busca de informações através de um URL especificado pelo usuário.

SpecificSearch: Busca informações por meio de palavras-chave, podendo ou não usar uma fonte de pesquisa, de acordo com a escolha do usuário. É uma classe derivada de **Search**.

SocialNetworkSearch: É uma classe derivada de **Search** que busca postagens no *Twitter* ou *Instagram*, a partir de um perfil inserido pelo usuário.

SearchEngine: Classe entidade que armazena os motores de busca disponíveis (Google, Google News, Bing e Bing Notícias). Esta classe está associada apenas à **SpecificSearch**.

SetDatasetType: Classe Entidade cuja responsabilidade é determinar o tipo de atributo principal da base, ou seja, base de textos, sentenças ou palavras.

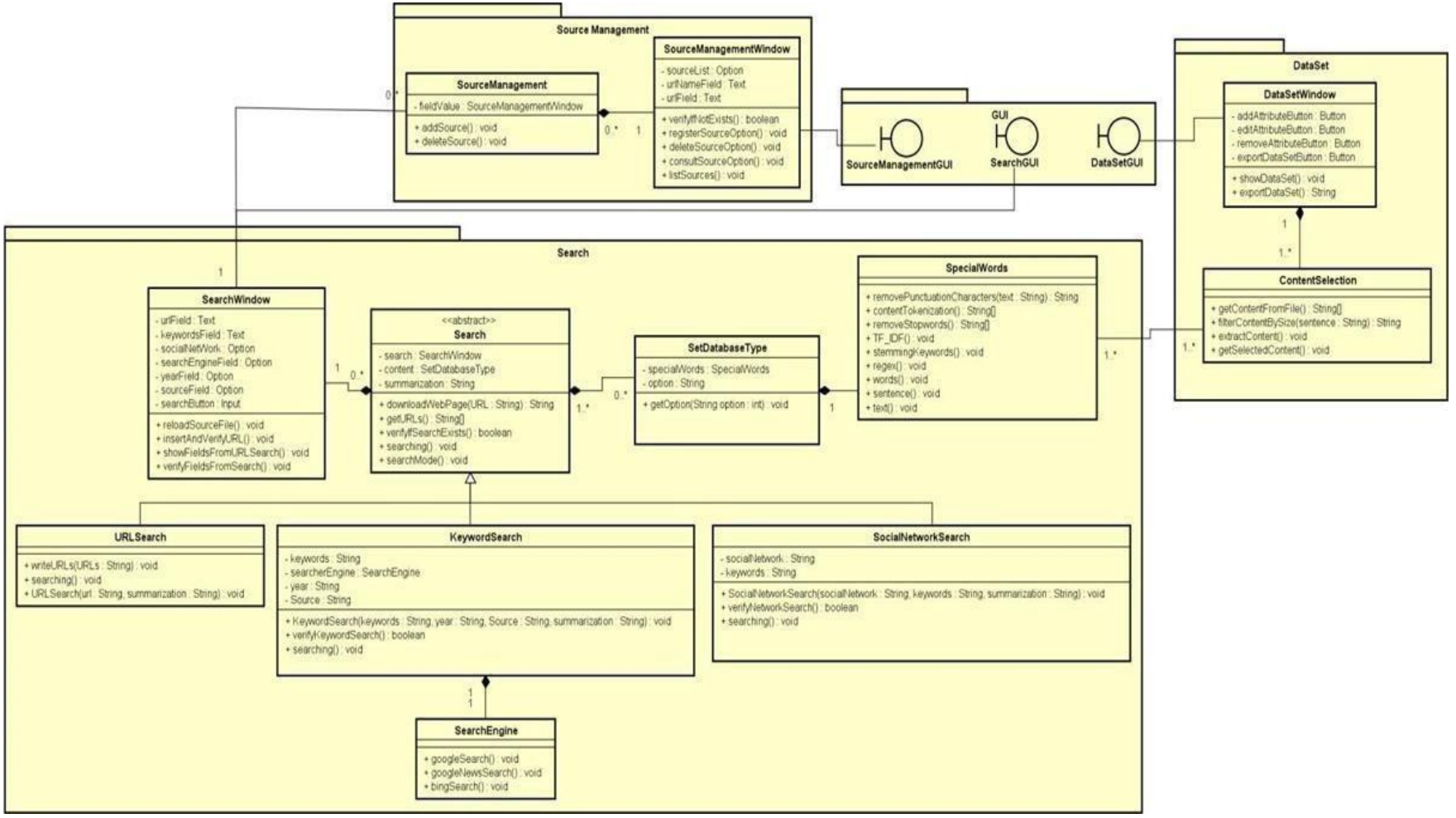
SpecialWords: Essa classe, do tipo entidade, é responsável por aplicar expressões regulares no texto obtido na busca Web e também realizar o pré-processamento, através das técnicas apresentadas no Subcapítulo 2.3.

DatasetGUI: Interface gráfica que exibe a base textual criada. Possibilita também que o usuário adicione, modifique ou exclua atributos.

Dataset: Classe do tipo entidade cuja responsabilidade é criar a base textual. Esta classe também faz a exportação da base para o formato de arquivo selecionado pelo usuário.

ContentSelection: Classe entidade que apresenta ao usuário uma lista com todo o conteúdo obtido na busca Web. Além disso, possibilita que o mesmo selecione o conteúdo de interesse que irá compor a base textual.

Figura 4.2 – Diagrama de classes



Fonte: Elaboração do autor

4.2 Resultados da implementação

A etapa de implementação consistiu no desenvolvimento de uma versão beta do SisBaseText. A versão beta de um sistema corresponde a uma versão ainda em estágio de desenvolvimento, mas que apresenta as principais funcionalidades implementadas.

Os resultados desta etapa estão dispostos em três subcapítulos: o Subcapítulo 4.2.1 apresenta a expressão regular desenvolvida, enquanto o Subcapítulo 4.2.2 mostra o gerenciamento de fontes de pesquisa e o Subcapítulo 4.2.3 discorre sobre a busca Web e a criação da base textual, apresentando um exemplo de cada busca.

4.2.1 Expressão regular desenvolvida

As páginas HTML são compostas por *tags* que têm a função de representar formulários, imagens, vídeos, tabelas, links e formatações textuais. Quando acessadas diretamente pelo navegador, realiza-se um processo automático denominado renderização[1] e todas as *tags* são interpretadas. Assim, o conteúdo é exibido aos usuários de forma compreensível (Karger, Ostler & Lee, 2009).

As técnicas de rastreamento não contemplam a renderização, pois mecanismos automáticos de download são utilizados, ao invés de acessos diretos mediante navegador. Em vista disso, deve-se empregar métodos para viabilizar a extração de conteúdo inteligível das páginas rastreadas. O uso de expressões regulares é uma boa estratégia, pois a remoção dos elementos HTML é realizada através da combinação de metacaracteres. O subcapítulo 2.2 discutiu as expressões regulares.

Entretanto, não há expressões regulares previamente desenvolvidas que eliminem *tags* sem afetar o texto contido na página. As poucas opções existentes desconsideram a remoção de anúncios, folhas de estilo CSS e *scripts* (Cohen, Ding & Bagherjeiran, 2015).

Por esse motivo, foi necessário desenvolver uma expressão regular capaz de remover todos os elementos citados no parágrafo acima, de maneira a preservar a integridade do conteúdo. O desenvolvimento da expressão regular demandou estudos sobre

[1] É um mecanismo que converte a linguagem de marcação HTML e as folhas de estilo CSS em um conteúdo formatado que pode ser entendido pelo usuário (Butkiewicz, Madhyastha & Sekar, 2011).

elementos presentes nas páginas HTML e recursos disponibilizados pela linguagem PHP para eliminá- los.

A expressão regular desenvolvida é composta por quatro padrões de metacaracteres que abrangem a remoção individual de todas as *tags* HTML, incluindo *scripts*, folhas de estilo e anúncios. A Tabela 4.1 apresenta os padrões desenvolvidos e seus respectivos significados.

Tabela 4.1 – Padrões desenvolvidos

Padrão	**Significado**	**Parâmetro**	***Tags* envolvidas**
#<\s*?$**cssTag**\b[^>]*>(*?)</$**cssTag**\b[^>]*>#s	Remove folhas de estilo CSS.	**cssTag:** indica a tag de estilos.	link e style
#<**code**(.*?)>(.*?)</**code**>#is	Remove códigos-fonte	**code:** indica a tag com o código-fonte	script, code, applet, noscript
/<**strucTag**[^>]*>([\s\S]*?)<\/ **strucTag** [^>]*>/	Remove a estrutura HTML.	**strucTag**: indica a tag de estrutura	!--, !DOCTYPE, a, abbr, acronym, area, article, aside, audio, b, base, basefont, bdi, bdo, big, blockquote, body, br, button, canvas, caption, center, cite, col, colgroup, command, data, datalist, del, dfn, dialog, dir, div, dl, dt, em, fieldset, figcaption, figure, font, form, h1-h6, head, header, hgroup, hr, html, i, img, input, ins, kbd, label, legend, li, main, map, mark, meta, meter, menu, nav, noembed, object, ol, optgroup, option, output, p, param, picture, pre, progress, q, rp, rt, ruby, s, samp, section, select, small, source, span, strike, strong, sub, sup, svg, table,tbody,td, textarea, tfoot, th, thead, time, title, tr, track, tt, u, ul, var, video, wbr
#<**adTag** class="[ad]\w*\b/g">(.*?)</**adtag**>#	Remove anúncios.	**adTag:** indica a tag de anúncios.	iframe, div, footer, address, details, embed, frame, frameset, noframes, summary, template,

Fonte: Elaboração do autor

A remoção das *tags* respeita a ordem apresentada na tabela acima, isto é, primeiro são eliminadas folhas de estilo, em seguida códigos-fonte, a estrutura HTML e,

por fim, os anúncios. É necessário seguir esta ordem para evitar que parte do conteúdo textual seja acidentalmente removido.

Desse modo, a expressão regular desenvolvida é capaz de extrair texto de quaisquer páginas, independentemente do tipo de estrutura. A Figura 4.3 apresenta o conteúdo textual obtido pela aplicação desta expressão regular em uma página HTML.

Figura 4.3 - Conteúdo extraído

```
Durante a sua viagem ao Japão, o presidente Jair Bolsonaro afirmou que conversou
com Fernando Azevedo e Silva, ministro da Defesa, e que o governo brasileiro
monitora eventuais manifestações no Brasil motivadas pela situação do Chile.
“Nós nos preparamos. Ontem conversei com o ministro da Defesa sobre a
possibilidade de termos movimentos como já tivemos no passado parecidos com o
que estão acontecendo no Chile e como devemos nos preparar para isso. […]
A gente se prepara para usar o artigo 142, que é para manutenção da lei e da
ordem, caso [os militares] venham a ser convocados”, declarou, ao ser questionado
sobre algumas análises de que o próximo País a ter agitação social seria o Brasil.
Bolsonaro ainda acusou Humberto Costa, senador pelo PT, de agitar as massas para
confronto no País.

Protestos
Os protestos no Chile começaram na sexta-feira (18) depois do aumento do preço das
passagens do metrô de Santiago – medida já suspensa pelo governo. Desde então,
a população incluiu outras demandas sociais nas manifestações. Até o momento, 15
pessoas morreram. Por quatro noites consecutivas, as Forças Armadas decretaram toque
de recolher.
```

Fonte: Elaboração do autor

4.2.2 Gerenciamento de fontes de pesquisa

O gerenciamento de fontes de pesquisa foi desenvolvido com objetivo de possibilitar o armazenamento das fontes de informação que o usuário julgar importantes para realizar a busca Web. Vale ressaltar que este módulo está acoplado apenas à busca com fonte, conforme apresentado no Subcapítulo 3.2. As figuras 4.4, 4.5 e 4.6 exemplificam as operações de cadastro, consulta e exclusão de uma fonte de pesquisa, enquanto a Figura 4.7 apresenta uma lista com fontes pré-cadastradas.

Figura 4.4 – Cadastro de fonte

Fonte: Elaboração do autor

Figura 4.5 - Consulta de fonte

Fonte: Elaboração do autor

Figura 4.6 – Exclusão de fonte

Fonte: Elaboração do autor

Figura 4.7 – Lista de fontes

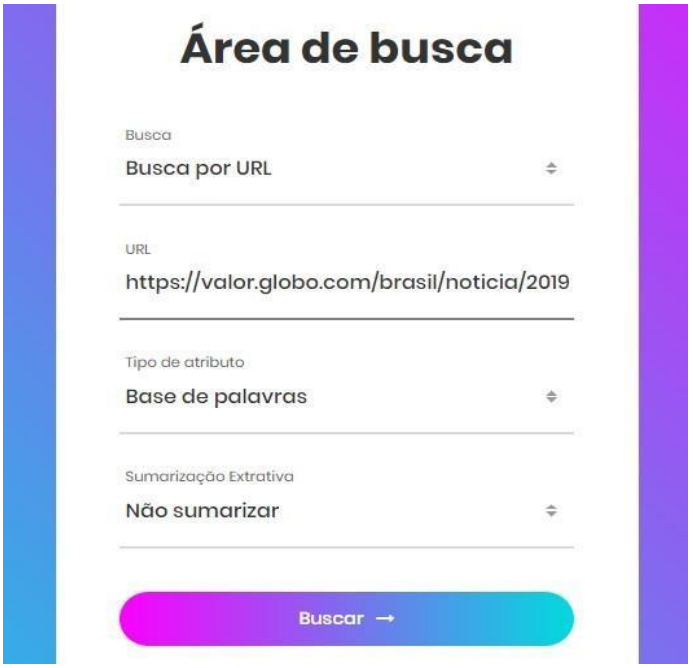
Fonte: Elaboração do autor

4.2.3 Busca Web e Criação da base textual

O módulo de busca Web possibilita ao usuário três formas distintas de busca, sendo que cada uma delas deve ser aplicada de acordo com a necessidade de informação, conforme descrito no Subcapítulo 3.2.

A Figura 4.8 apresenta a interface gráfica referente à tela inicial do sistema, com a opção busca por URL selecionada. Neste exemplo o URL inserido foi "https://valor.globo.com/brasil/noticia/2019/10/22/presidente- da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml". Este URL refere-se à revista de divulgação econômica "Valor Econômico". O tipo de atributo principal escolhido foi "Base de palavras", ou seja, o conteúdo retornado pela busca será dividido em palavras. Em relação à sumarização, foi selecionada a opção "Não sumarizar".

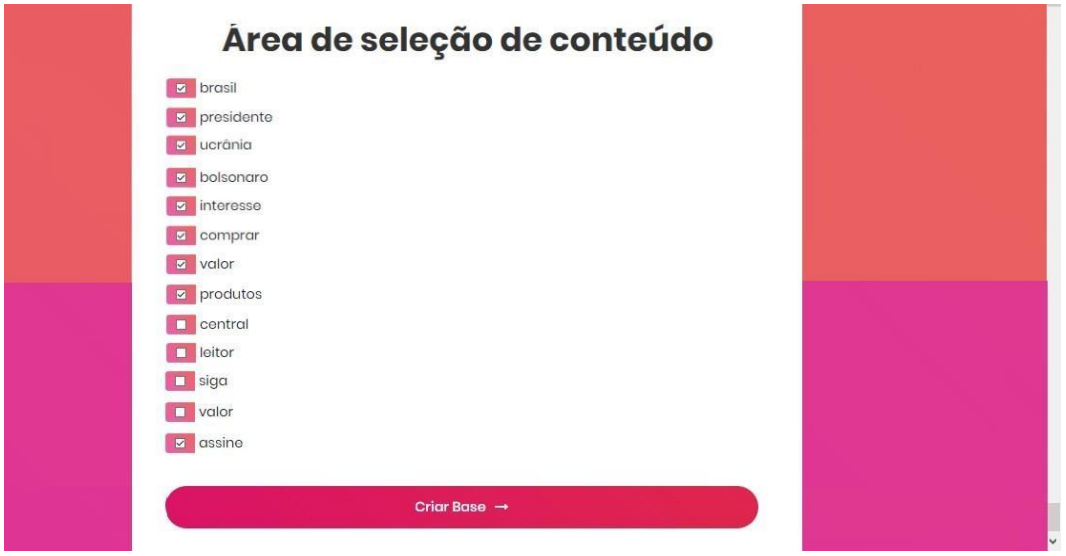
Figura 4.8 – Busca por URL



Fonte: Elaboração do autor

A Figura 4.9 contém parte dos resultados da busca mostrada na figura anterior. O conteúdo foi dividido em uma lista com 504 palavras, sendo que 53 foram selecionadas para compor a base textual.

Figura 4.9 – Visualização e seleção de conteúdo (Base de palavras)



Fonte: Elaboração do autor

A interface gráfica contendo a base textual criada é mostrada na Figura 4.10. Neste exemplo, não foram adicionados ou removidos quaisquer atributos. Foi escolhida a opção de exportação para um arquivo com a extensão .xls através do botão "*Export to xls*". A base exportada pode ser visualizada na Figura 4.11.

Figura 4.10 – Base textual

ID	Data de coleta	Conteúdo	Sumarização	Tipo de base	Tipo de busca	URL
0	23-10-2019 16:22:42	brasil	não	palavras	Busca por URL	https://valor.globo.com/bra da-ucrnia-diz-a-bolsonaro- ter-interesse-em-comprar- super-tucano-e-kc- 390.ghtml
1	23-10-2019 16:22:42	presidente	não	palavras	Busca por URL	https://valor.globo.com/bra da-ucrnia-diz-a-bolsonaro- ter-interesse-em-comprar- super-tucano-e-kc- 390.ghtml
2	23-10-2019 16:22:42	ucrânia	não	palavras	Busca por URL	https://valor.globo.com/bra da-ucrnia-diz-a-bolsonaro- ter-interesse-em-comprar- super-tucano-e-kc- 390.ghtml
3	23-10-2019 16:22:42	bolsonaro	não	palavras	Busca por URL	https://valor.globo.com/bra da-ucrnia-diz-a-bolsonaro- ter-interesse-em-comprar- super-tucano-e-kc- 390.ghtml
47	23-10-2019 16:22:42	mudança	não	palavras	Busca por URL	https://valor.globo.com/bra da-ucrnia-diz-a-bolsonaro- ter-interesse-em-comprar- super-tucano-e-kc- 390.ghtml
48	23-10-2019 16:22:42	redacional	não	palavras	Busca por URL	https://valor.globo.com/bra da-ucrnia-diz-a-bolsonaro- ter-interesse-em-comprar- super-tucano-e-kc- 390.ghtml
49	23-10-2019 16:22:42	pec	não	palavras	Busca por URL	https://valor.globo.com/bra da-ucrnia-diz-a-bolsonaro- ter-interesse-em-comprar- super-tucano-e-kc- 390.ghtml
50	23-10-2019 16:22:42	valor	não	palavras	Busca por URL	https://valor.globo.com/bra da-ucrnia-diz-a-bolsonaro- ter-interesse-em-comprar- super-tucano-e-kc- 390.ghtml
51	23-10-2019 16:22:42	produtos	não	palavras	Busca por URL	https://valor.globo.com/bra da-ucrnia-diz-a-bolsonaro- ter-interesse-em-comprar- super-tucano-e-kc- 390.ghtml
52	23-10-2019 16:22:42	assine	não	palavras	Busca por URL	https://valor.globo.com/bra da-ucrnia-diz-a-bolsonaro- ter-interesse-em-comprar- super-tucano-e-kc- 390.ghtml

Fonte: Elaboração do autor

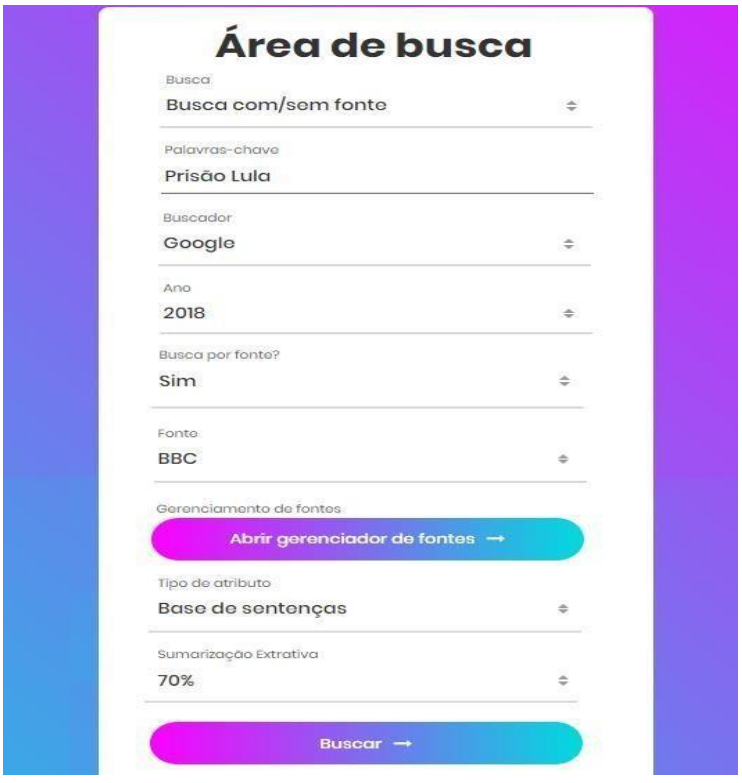
Figura 4.11 –Arquivo .xls contendo a base

	A	B	C	D	E	F	G
1							
2	ID	Data de coleta	Conteúdo	Sumarização	Tipo de base	Tipo de busca	URL
3	0	23-10-2019 16:22:42	brasil	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
4	1	23-10-2019 16:22:42	presidente	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
5	2	23-10-2019 16:22:42	ucrânia	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
6	3	23-10-2019 16:22:42	bolsonaro	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
7	4	23-10-2019 16:22:42	interesse	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
8	5	23-10-2019 16:22:42	comprar	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
9	6	23-10-2019 16:22:42	tucano	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
10	7	23-10-2019 16:22:42	volodymyr	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
11	8	23-10-2019 16:22:42	zelenskiy	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
12	9	23-10-2019 16:22:42	aviões	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
13	10	23-10-2019 16:22:42	super	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
14	11	23-10-2019 16:22:42	missões	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
15	12	23-10-2019 16:22:42	ataque	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
16	13	23-10-2019 16:22:42	examinar	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
17	14	23-10-2019 16:22:42	cargueiro	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
18	15	23-10-2019 16:22:42	militar	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
19	16	23-10-2019 16:22:42	fabricado	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
20	17	23-10-2019 16:22:42	brasil	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
21	18	23-10-2019 16:22:42	relato	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
22	19	23-10-2019 16:22:42	ministro	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
23	20	23-10-2019 16:22:42	relações	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
24	21	23-10-2019 16:22:42	exteriores	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
25	22	23-10-2019 16:22:42	ernesto	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
26	23	23-10-2019 16:22:42	araujo	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
27	24	23-10-2019 16:22:42	bolsa	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
28	25	23-10-2019 16:22:42	novo	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
29	26	23-10-2019 16:22:42	recorde	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
30	27	23-10-2019 16:22:42	previdência	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
31	28	23-10-2019 16:22:42	dólar	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
32	29	23-10-2019 16:22:42	reforma	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
33	30	23-10-2019 16:22:42	repercute	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml
34	31	23-10-2019 16:22:42	mercados	não	palavras	Busca por URL	https://valor.globo.com/brasil/noticia/2019/10/22/presidente-da-ucrnia-diz-a-bolsonaro-ter-interesse-em-comprar-super-tucano-e-kc-390.ghtml

Fonte: Elaboração do autor

A busca com fonte é exemplificada pela Figura 4.12. Nesta tela o campo de palavras-chave contém a busca: “Prisão Lula”, enquanto o buscador selecionado foi o “GOOGLE”, o ano “2018” e a fonte “BBC”. Além disso, o tipo de atributo selecionado foi “Base de sentenças” e a Sumarização aplicada foi de 70%, sendo esta a quantidade de texto que será mantida em detrimento ao texto original.

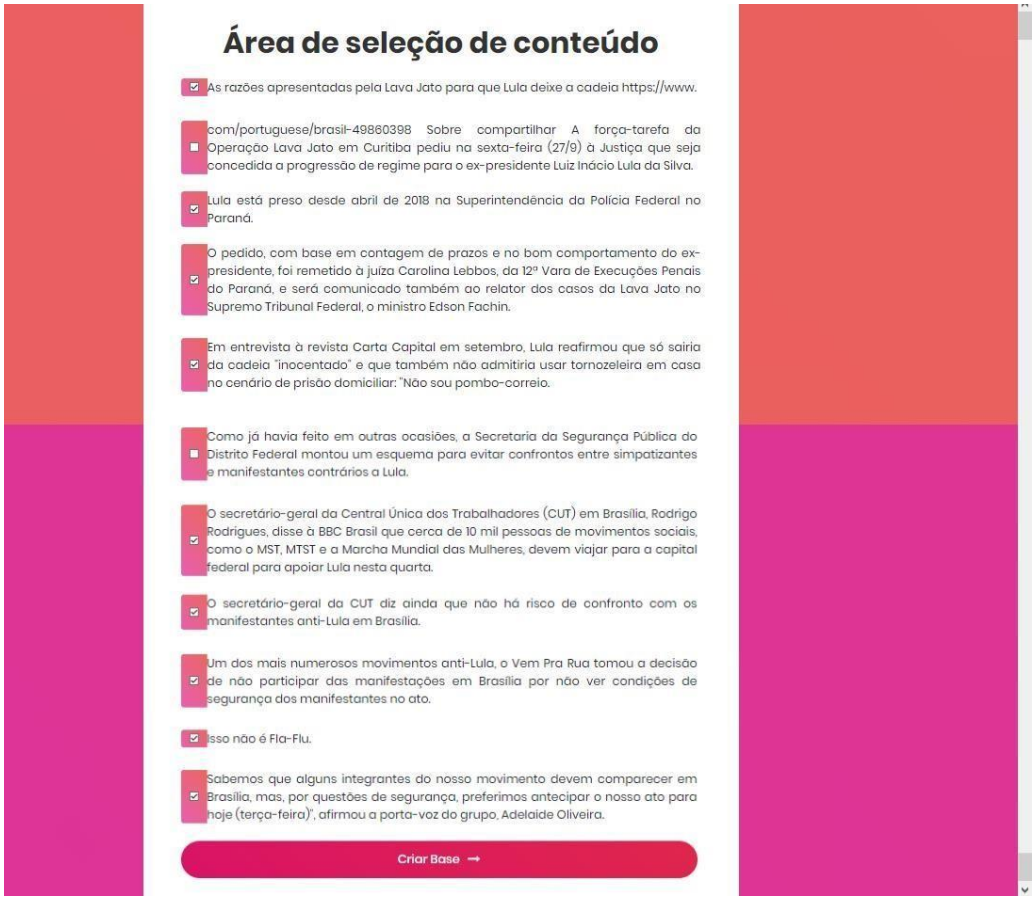
Figura 4.12 – Busca com fonte



Fonte: Elaboração do autor

O conteúdo retornado pela busca com fonte foi separado em 413 sentenças, sendo que 100 sentenças foram selecionadas para a criação da base textual. Esta tela é mostrada na Figura 4.13.

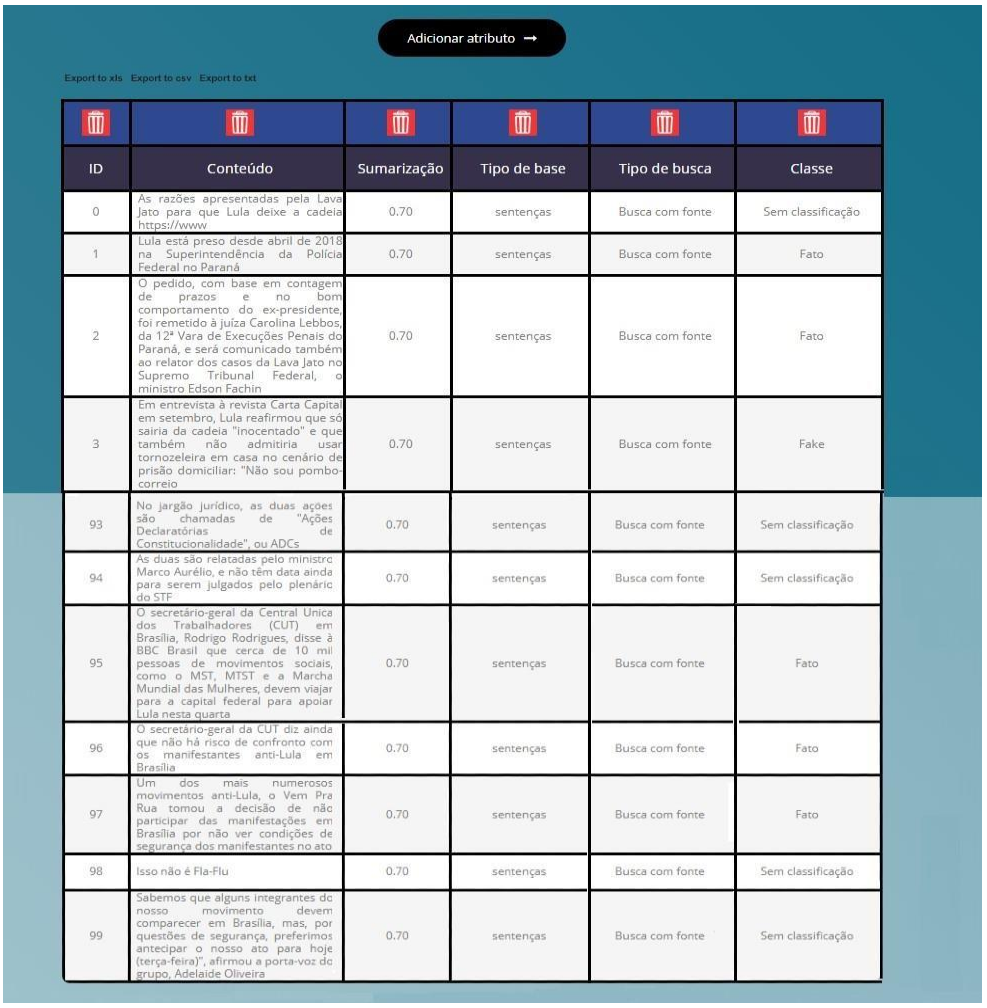
Figura 4.13 – Visualização e seleção de conteúdo (Base de sentenças)



Fonte: Elaboração do autor

A base textual criada é mostrada na Figura 4.14. Neste exemplo, foram removidos os atributos Data de Coleta e URL. Por outro lado, foi inserido o atributo Classe, que pode assumir os valores de “Fato”, “Fake”, ou “Sem classificação”. Uma base textual como esta poderia, posteriormente, ser usada para treinar um algoritmo de aprendizado de máquina, conforme foi discorrido no Subcapítulo 2.4.

Figura 4.14 – Base textual com atributo classificador



ID	Conteúdo	Sumarização	Tipo de base	Tipo de busca	Classe
0	As razões apresentadas pela Lava Jato para que Lula deixe a cadeia https://www	0.70	sentenças	Busca com fonte	Sem classificação
1	Lula está preso desde abril de 2018 na Superintendência da Polícia Federal no Paraná	0.70	sentenças	Busca com fonte	Fato
2	O pedido, com base em contagem de prazos e no bom comportamento do ex-presidente, foi remetido à juíza Carolina Lebbos, da 12ª Vara de Execuções Penais do Paraná, e será comunicado também ao relator dos casos da Lava Jato no Supremo Tribunal Federal, o ministro Edson Fachin	0.70	sentenças	Busca com fonte	Fato
3	Em entrevista à revista Carta Capital em setembro, Lula reafirmou que só sairia da cadeia "inocentado" e que também não admitiria usar tornozeleira em casa no cenário de prisão domiciliar: "Não sou pombo-correio	0.70	sentenças	Busca com fonte	Fake
93	No jargão jurídico, as duas ações são chamadas de "Ações Declaratórias de Constitucionalidade", ou ADCs	0.70	sentenças	Busca com fonte	Sem classificação
94	As duas são relatadas pelo ministro Marco Aurélio, e não têm data ainda para serem julgados pelo plenário do STF	0.70	sentenças	Busca com fonte	Sem classificação
95	O secretário-geral da Central Única dos Trabalhadores (CUT) em Brasília, Rodrigo Rodrigues, disse à BBC Brasil que cerca de 10 mil pessoas de movimentos sociais, como o MST, MTST e a Marcha Mundial das Mulheres, devem viajar para a capital federal para apoiar Lula nesta quarta	0.70	sentenças	Busca com fonte	Fato
96	O secretário-geral da CUT diz ainda que não há risco de confronto com os manifestantes anti-Lula em Brasília	0.70	sentenças	Busca com fonte	Fato
97	Um dos mais numerosos movimentos anti-Lula, o Vem Pra Rua tomou a decisão de não participar das manifestações em Brasília por não ver condições de segurança dos manifestantes no ato	0.70	sentenças	Busca com fonte	Fato
98	Isso não é Fla-Flu	0.70	sentenças	Busca com fonte	Sem classificação
99	Sabemos que alguns integrantes do nosso movimento devem comparecer em Brasília, mas, por questões de segurança, preferimos antecipar o nosso ato para hoje (terça-feira)", afirmou a porta-voz do grupo, Adelaide Oliveira	0.70	sentenças	Busca com fonte	Sem classificação

Fonte: Elaboração do autor

Foi escolhida a opção de exportação para um arquivo com a extensão .csv. A base exportada pode ser visualizada na Figura 4.15.

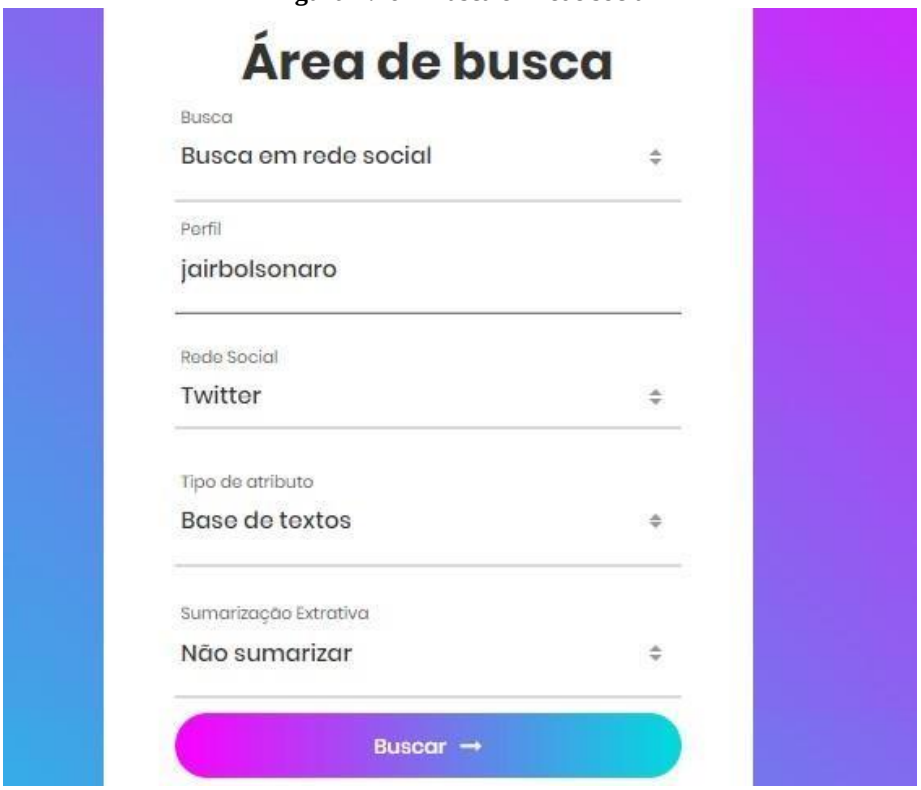
Figura 4.15 –Arquivo .csv contendo a base

0","23-10-2019 18:39:51","As razões apresentadas pela Lava Jato para que Lula deixe a cadeia https://www","0.70","sentenças","Busca com fonte","Sem classificação
"1","23-10-2019 18:39:51","Lula está preso desde abril de 2018 na Superintendência da Polícia Federal no Paraná","0.70","sentenças","Busca com fonte","Fato
"2","23-10-2019 18:39:51","O pedido, com base em contagem de prazos e no bom comportamento do ex-presidente, foi remetido à juíza Carolina Lebbos, da 12ª Vara de Execuções Penais do Paraná, e será comunicado também ao relator dos casos da Lava Jato no Supremo Tribunal Federal, o ministro Edson Fachin","0.70","sentenças","Busca com fonte","Fato
"3","23-10-2019 18:39:51","Em entrevista à revista Carta Capital em setembro, Lula reafirmou que só sairia da cadeia "inocentado" e que também não admitiria usar tornozeleira em casa no cenário de prisão domiciliar: "Não sou pombo-correio","0.70","sentenças","Busca com fonte","Fake
"4","23-10-2019 18:39:51","Se quiserem colocar uma corrente, coloquem no pescoço do Moro (Sergio Moro, atual ministro da Justiça que condenou o ex-presidente quando era juiz federal), não na minha canela","0.70","sentenças","Busca com fonte","Fake
"5","23-10-2019 18:39:51","Lula pode ir para prisão domiciliar? No pedido de sexta-feira, os procuradores lembram que o ex-presidente está "na iminência de atender ao critério temporal" determinado pelo Artigo 112 da Lei de Execuções Penais","0.70","sentenças","Busca com fonte","Fato
"6","23-10-2019 18:39:51","Pela legislação, o preso que cumpriu um sexto da pena a que foi condenado e apresentou bom comportamento, entre outros critérios, tem direito a progredir para um regime menos rigoroso de prisão - no caso de Lula, deixar o regime fechado e ir para o semiaberto","0.70","sentenças","Busca com fonte","Fato
"7","23-10-2019 18:39:51","No texto, o Ministério Público Federal pede que "seja deferida a Luiz Inácio Lula da Silva a progressão ao regime semiaberto, na forma dos arts","0.70","sentenças","Busca com fonte","Fato
"8","23-10-2019 18:39:51","91 e seguinte da LEP, devendo ser observado pelo juízo o disposto na Súmula Vinculante nº 56"","0.70","sentenças","Busca com fonte","Sem classificação
"9","23-10-2019 18:39:51","A súmula 56 é uma decisão do STF determinando, entre outras coisas, que o preso não pode ser mantido em um regime "mais gravoso" apenas pela falta de vagas em estabelecimentos onde cumpriria o regime mais brando","0.70","sentenças","Busca com fonte","Sem classificação
"10","23-10-2019 18:39:51","No caso de falta de vagas para cumprimento do semiaberto - regime em que o preso trabalha fora durante o dia e passa as noites recolhido na prisão - a saída pode ser a prisão domiciliar","0.70","sentenças","Busca com fonte","Fake
"11","23-10-2019 18:39:51","Até que sejam estruturadas as medidas alternativas propostas, poderá ser deferida a prisão domiciliar ao sentenciado"","0.70","sentenças","Busca com fonte","Fake
"12","23-10-2019 18:39:51","No Estado de São Paulo, por exemplo, há um déficit crônico de vagas para cumprimento do semiaberto","0.70","sentenças","Busca com fonte","Fake
"13","23-10-2019 18:39:51","O petista foi sentenciado, inicialmente, a nove anos e seis meses de prisão, em julgamento na 13ª Vara Federal de Curitiba — onde atuava o então juiz e atual ministro da Justiça e Segurança Pública Sergio Moro","0.70","sentenças","Busca com fonte","Fato
"14","23-10-2019 18:39:51","Em segunda instância, o ex-presidente foi julgado no Tribunal Regional Federal da 4ª Região (TRF-4), que aumentou a pena para 12 anos e um mês","0.70","sentenças","Busca com fonte","Fato
"15","23-10-2019 18:39:51","É sobre esta base que é feito o cálculo temporal para determinar que ele tem direito a progredir para o semiaberto","0.70","sentenças","Busca com fonte","Sem classificação
"16","23-10-2019 18:39:51","Seu julgamento está interrompido por pedido de vista de Gilmar Mendes e ainda não há data para sua retomada","0.70","sentenças","Busca com fonte","Fato
"17","23-10-2019 18:39:51","Caso o recurso seja aceito pelo STF, ele pode levar à anulação de todos os atos processuais de Moro quando era juiz, em processos e inquéritos contra Lula","0.70","sentenças","Busca com fonte","Fake
"18","23-10-2019 18:39:51","Isso cancelaria a condenação de Lula no caso Triplex do Guarujá por corrupção passiva e lavagem de dinheiro, mesmo que a sentença já tenha sido confirma pelo STJ","0.70","sentenças","Busca com fonte","Fato
"19","23-10-2019 18:39:51","Também anularia a condenação de Lula no caso do Sítio de Atibaia pela juíza Gabriela Hardt, já que ela assumiu o caso em sua etapa final","0.70","sentenças","Busca com fonte","Fato
"20","23-10-2019 18:39:51","O petista acabou barrado da eleição presidencial do ano passado pela lei da Ficha Limpa, após o TRF-4 ter confirmado a condenação do petista por Moro","0.70","sentenças","Busca com fonte","Sem classificação
"21","23-10-2019 18:39:51","STJ reduz pena de Lula: o que acontece agora com o ex-presidente? https://www","0.70","sentenças","Busca com fonte","Sem classificação
"22","23-10-2019 18:39:51","Em nota, a defesa do ex-presidente Lula criticou a decisão dos ministros","0.70","sentenças","Busca com fonte","Sem classificação
"23","23-10-2019 18:39:51","Para a defesa do ex-presidente, "o único desfecho possível é a absolvição do ex-presidente Lula, porque ele não praticou qualquer crime"","0.70","sentenças","Busca com fonte","Fake
"24","23-10-2019 18:39:51","Pela primeira vez um Tribunal reconheceu que as penas aplicadas pelo ex-juiz Sérgio Moro e pelo TRF4 foram abusivas","0.70","sentenças","Busca com fonte","Sem classificação
"25","23-10-2019 18:39:51","Segundo Castelo Branco, podem ser apresentados tanto pela defesa de Lula quanto pelo Ministério Público Federal (MPF)","0.70","sentenças","Busca com fonte","Fake
"26","23-10-2019 18:39:51","Em seguida, tanto o MPF quanto a defesa podem contestar a decisão da 5ª Turma do STJ no Supremo, através do chamado Recurso Extraordinário","0.70","sentenças","Busca com fonte","Fake
"27","23-10-2019 18:39:51","Participaram do julgamento de hoje os ministros Felix Fischer (relator dos casos da Lava Jato no STJ), Jorge Mussi, Reynaldo Soares da Fonseca e Ribeiro Dantas","0.70","sentenças","Busca com fonte","Fato
"28","23-10-2019 18:39:51","Os quatro ministros que participaram da sessão concordaram de forma unânime com a redução de pena","0.70","sentenças","Busca com fonte","Fato

Fonte: Elaboração do autor

O terceiro tipo de busca implementado foi a busca em rede social. A Figura 4.16 contém um exemplo, cujo perfil escolhido foi “jairbolsonaro”, a rede social selecionada foi “*Twitter*”, o atributo principal selecionado foi “Base de Textos” e não houve uso da sumarização.

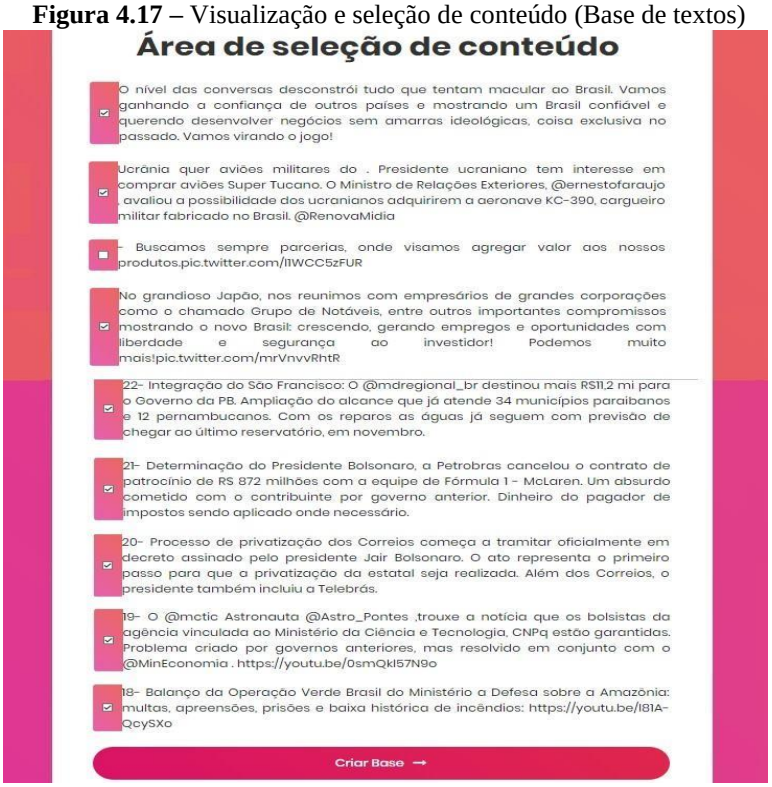
Figura 4.16 – Busca em rede social



Fonte: Elaboração do autor

O conteúdo obtido pela busca representada na Figura anterior consiste em 40 *posts*. Esse número representa o limite de retorno permitido pelo *Twitter* em cada busca realizada. Foram selecionados 37 posts para a base textual, conforme mostrado parcialmente pela Figura 4.17.

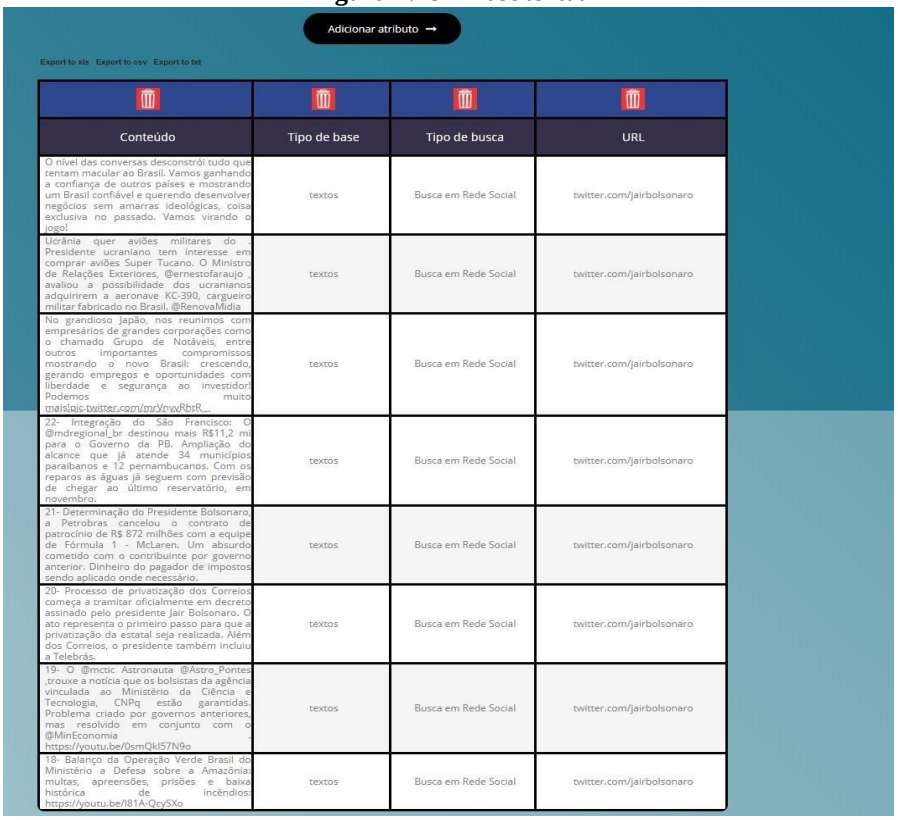
Figura 4.17 – Visualização e seleção de conteúdo (Base de textos)



Fonte: Elaboração do autor

A base textual criada é mostrada na Figura 4.18. Neste exemplo, foram removidos os atributos ID, Data de Coleta e Sumarização. Nenhum atributo foi adicionado.

Figura 4.18 – Base textual

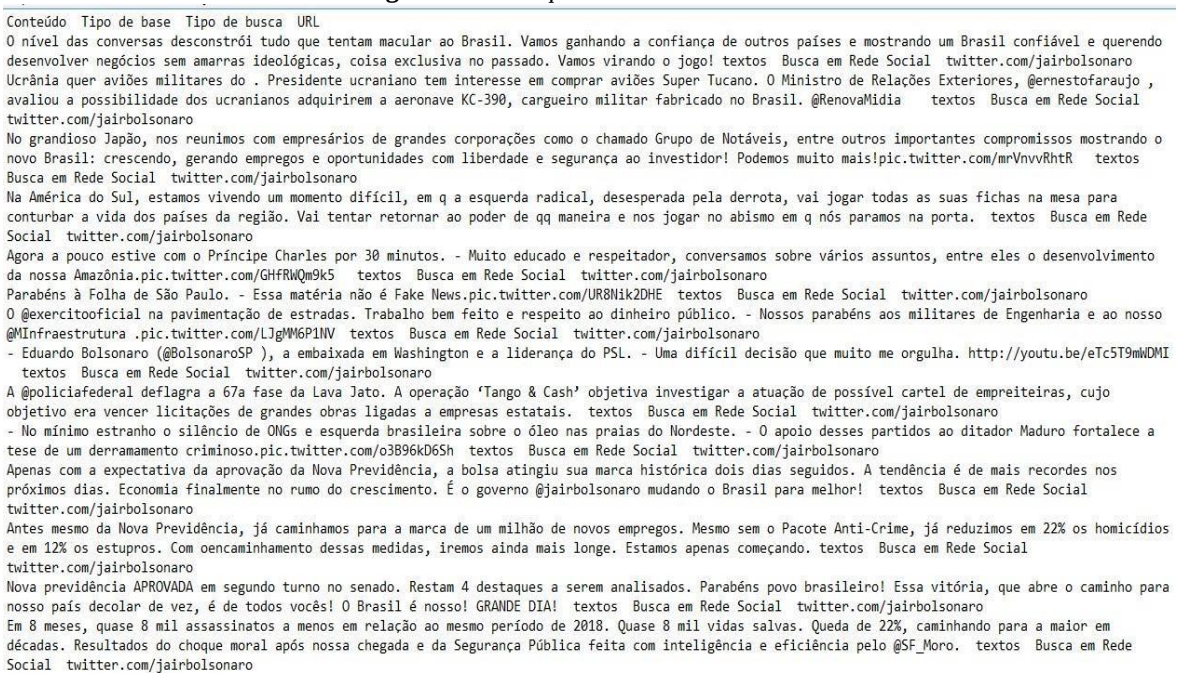


Conteúdo	Tipo de base	Tipo de busca	URL
O nível das conversas desconstrói tudo que tentam macular ao Brasil. Vamos ganhando a confiança de outros países e mostrando um Brasil confiável e querendo desenvolver negócios sem amarras ideológicas, coisa exclusiva no passado. Vamos virando o jogo!	textos	Busca em Rede Social	twitter.com/jairbolsonaro
Ucrânia quer aviões militares do Presidente ucraniano tem interesse em comprar aviões Super Tucano. O Ministro de Relações Exteriores, @ernestofaraujo , avaliou a possibilidade dos ucranianos adquirirem a aeronave KC-390, cargueiro militar fabricado no Brasil. @RenovaMidia	textos	Busca em Rede Social	twitter.com/jairbolsonaro
No grandioso Japão, nos reunimos com empresários de grandes corporações como o chamado Grupo de Notáveis, entre outros importantes compromissos mostrando o novo Brasil: crescendo, gerando empregos e oportunidades com liberdade e segurança ao investidor! Podemos muito mais!pic.twitter.com/mrVnvvRhtR...	textos	Busca em Rede Social	twitter.com/jairbolsonaro
22- Integração do São Francisco: O @mdregional_br destinou mais R$11,2 mi para o Governo da PB. Ampliação do alcance que já atende 34 municípios paraibanos e 12 pernambucanos. Com os reparos as águas já seguem com previsão de chegar ao último reservatório, em novembro.	textos	Busca em Rede Social	twitter.com/jairbolsonaro
21- Determinação do Presidente Bolsonaro, a Petrobras cancelou o contrato de patrocínio de R$ 872 milhões com a equipe de Fórmula 1 - McLaren. Um absurdo cometido com o contribuinte por governo anterior. Dinheiro do pagador de impostos sendo aplicado onde necessário.	textos	Busca em Rede Social	twitter.com/jairbolsonaro
20- Processo de privatização dos Correios começa a tramitar oficialmente em decreto assinado pelo presidente Jair Bolsonaro. O ato representa o primeiro passo para que a privatização da estatal seja realizada. Além dos Correios, o presidente também incluiu a Telebrás.	textos	Busca em Rede Social	twitter.com/jairbolsonaro
19- O @mctic Astronauta @Astro_Pontes ,trouxe a notícia que os bolsistas da agência vinculada ao Ministério da Ciência e Tecnologia, CNPq estão garantidas. Problema criado por governos anteriores, mas resolvido em conjunto com o @MinEconomia . https://youtu.be/0smQkI57N9o	textos	Busca em Rede Social	twitter.com/jairbolsonaro
18- Balanço da Operação Verde Brasil do Ministério a Defesa sobre a Amazônia: multas, apreensões, prisões e baixa histórica de incêndios: https://youtu.be/I81A-QcySXo	textos	Busca em Rede Social	twitter.com/jairbolsonaro

Fonte: Elaboração do autor

A exportação da base textual criada foi feita em um arquivo .txt, conforme a Figura 4.19.

Figura 4.19 –Arquivo .txt contendo a base



Conteúdo Tipo de base Tipo de busca URL
O nível das conversas desconstrói tudo que tentam macular ao Brasil. Vamos ganhando a confiança de outros países e mostrando um Brasil confiável e querendo desenvolver negócios sem amarras ideológicas, coisa exclusiva no passado. Vamos virando o jogo! textos Busca em Rede Social twitter.com/jairbolsonaro
Ucrânia quer aviões militares do . Presidente ucraniano tem interesse em comprar aviões Super Tucano. O Ministro de Relações Exteriores, @ernestofaraujo , avaliou a possibilidade dos ucranianos adquirirem a aeronave KC-390, cargueiro militar fabricado no Brasil. @RenovaMidia textos Busca em Rede Social twitter.com/jairbolsonaro
No grandioso Japão, nos reunimos com empresários de grandes corporações como o chamado Grupo de Notáveis, entre outros importantes compromissos mostrando o novo Brasil: crescendo, gerando empregos e oportunidades com liberdade e segurança ao investidor! Podemos muito mais!pic.twitter.com/mrVnvvRhtR textos Busca em Rede Social twitter.com/jairbolsonaro
Na América do Sul, estamos vivendo um momento difícil, em q a esquerda radical, desesperada pela derrota, vai jogar todas as suas fichas na mesa para conturbar a vida dos países da região. Vai tentar retornar ao poder de qq maneira e nos jogar no abismo em q nós paramos na porta. textos Busca em Rede Social twitter.com/jairbolsonaro
Agora a pouco estive com o Príncipe Charles por 30 minutos. - Muito educado e respeitador, conversamos sobre vários assuntos, entre eles o desenvolvimento da nossa Amazônia.pic.twitter.com/GHfRWQm9k5 textos Busca em Rede Social twitter.com/jairbolsonaro
Parabéns à Folha de São Paulo. - Essa matéria não é Fake News.pic.twitter.com/UR8Nik2DHE textos Busca em Rede Social twitter.com/jairbolsonaro
O @exercitooficial na pavimentação de estradas. Trabalho bem feito e respeito ao dinheiro público. - Nossos parabéns aos militares de Engenharia e ao nosso @MInfraestrutura .pic.twitter.com/LJgMM6P1NV textos Busca em Rede Social twitter.com/jairbolsonaro
- Eduardo Bolsonaro (@BolsonaroSP), a embaixada em Washington e a liderança do PSL. - Uma difícil decisão que muito me orgulha. http://youtu.be/eTc5T9mWDMI textos Busca em Rede Social twitter.com/jairbolsonaro
A @policiafederal deflagra a 67a fase da Lava Jato. A operação 'Tango & Cash' objetiva investigar a atuação de possível cartel de empreiteiras, cujo objetivo era vencer licitações de grandes obras ligadas a empresas estatais. textos Busca em Rede Social twitter.com/jairbolsonaro
- No mínimo estranho o silêncio de ONGs e esquerda brasileira sobre o óleo nas praias do Nordeste. - O apoio desses partidos ao ditador Maduro fortalece a tese de um derramamento criminoso.pic.twitter.com/o3B96kD6Sh textos Busca em Rede Social twitter.com/jairbolsonaro
Apenas com a expectativa da aprovação da Nova Previdência, a bolsa atingiu sua marca histórica dois dias seguidos. A tendência é de mais recordes nos próximos dias. Economia finalmente no rumo do crescimento. É o governo @jairbolsonaro mudando o Brasil para melhor! textos Busca em Rede Social twitter.com/jairbolsonaro
Antes mesmo da Nova Previdência, já caminhamos para a marca de um milhão de novos empregos. Mesmo sem o Pacote Anti-Crime, já reduzimos em 22% os homicídios e em 12% os estupros. Com oencaminhamento dessas medidas, iremos ainda mais longe. Estamos apenas começando. textos Busca em Rede Social twitter.com/jairbolsonaro
Nova previdência APROVADA em segundo turno no senado. Restam 4 destaques a serem analisados. Parabéns povo brasileiro! Essa vitória, que abre o caminho para nosso país decolar de vez, é de todos vocês! O Brasil é nosso! GRANDE DIA! textos Busca em Rede Social twitter.com/jairbolsonaro
Em 8 meses, quase 8 mil assassinatos a menos em relação ao mesmo período de 2018. Quase 8 mil vidas salvas. Queda de 22%, caminhando para a maior em décadas. Resultados do choque moral após nossa chegada e da Segurança Pública feita com inteligência e eficiência pelo @SF_Moro. textos Busca em Rede Social twitter.com/jairbolsonaro

Fonte: Elaboração do autor

4.3 Validação do sistema

O SisBaseText foi avaliado por meio de testes funcionais envolvendo as métricas de precisão e revocação, conforme citadas no Subcapítulo 3.3. Os testes realizados envolveram apenas a busca com fonte para verificar a capacidade do sistema em retornar a informação desejada. A busca por URL e a busca em rede social não foram consideradas, pois sempre retornarão as páginas de interesse. Não foi avaliada a qualidade da base textual, pois esse processo é semiautomático e as bases geradas dependem fortemente da finalidade para qual o usuário irá aplicá-las.

Os testes envolveram a escolha de fontes de pesquisa, escolha de assuntos para as buscas, seleção manual de páginas, cálculo de precisão e revocação por fonte, cálculo de precisão e revocação por assunto e comparação de resultados obtidos.

Foram escolhidas cinco fontes de maneira aleatória, sendo elas: Valor Econômico, FAPESP, O Antagonista, Folha de SP e Amazon. A escolha de assuntos para as buscas foi feita a partir da priorização de temas recentes e considerados de alto impacto pela mídia, por conta da grande repercussão. Logo, foi decidido que a busca focaria no

seguintes assuntos: “Escândalo da Petrobrás”, “Tragédia em Brumadinho” e “Eleições Bolsonaro”.

Após a definição das fontes de pesquisa e dos assuntos de interesse foi necessário iniciar uma busca e seleção manual de URLs. Para cada assunto foram selecionados 90 URLs que estavam divididos igualmente entre as cinco fontes de pesquisa. O mecanismo de busca utilizado foi “Google”. Todos os URLs receberam um rótulo “relevante” ou “irrelevante”. A Tabela 4.2 apresenta parcialmente este procedimento considerando o assunto “Eleições Bolsonaro” e a fonte “Folha de SP”.

Tabela 4.2 – Resultados parciais de uma busca manual

Eleições Bolsonaro		
Fonte	**URL escolhida manualmente**	**Relevância**
Folha de SP	https://www1.folha.uol.com.br/poder/2018/10/28	Relevante
	https://www1.folha.uol.com.br/poder/2018/11/27	Relevante
	https://www1.folha.uol.com.br/poder/2018/11/25	Relevante
	https://www1.folha.uol.com.br/poder/2018/09/17	Irrelevante
	https://www1.folha.uol.com.br/poder/2018/09/19	Irrelevante
	https://www1.folha.uol.com.br/poder/2018/07/29	Relevante
	https://www1.folha.uol.com.br/poder/2018/08/30	Relevante
	https://www1.folha.uol.com.br/poder/2018/05/02	Relevante
	https://www1.folha.uol.com.br/poder/2018/03/03	Relevante
	https://www1.folha.uol.com.br/poder/2018/10/20	Irrelevante
	https://www1.folha.uol.com.br/poder/2018/09/14	Relevante
	https://www1.folha.uol.com.br/poder/2018/10/19	Relevante
	https://www1.folha.uol.com.br/poder/2018/10/07	Irrelevante
	https://www1.folha.uol.com.br/poder/2018/07/08	Relevante
	https://www1.folha.uol.com.br/poder/2018/09/06	Irrelevante
	https://www1.folha.uol.com.br/poder/2018/05/10	Relevante
	https://www1.folha.uol.com.br/poder/2018/08/23	Irrelevante
	https://www1.folha.uol.com.br/poder/2018/10/23	Relevante

Fonte: Elaboração do autor

Um procedimento semelhante ao citado no parágrafo anterior foi feito através do SisBaseText, ou seja, também foram buscados os mesmos assuntos nas fontes de pesquisa definidas e 90 URLs foram retornados através do mecanismo de busca “Google”. O objetivo era comparar os URLs selecionados manualmente com os URLs encontrados pelo SisBaseText, de modo que fosse possível medir a capacidade de retorno de páginas consideradas relevantes em detrimento às irrelevantes. Essa comparação pode ser exemplificada pela Tabela 4.3.

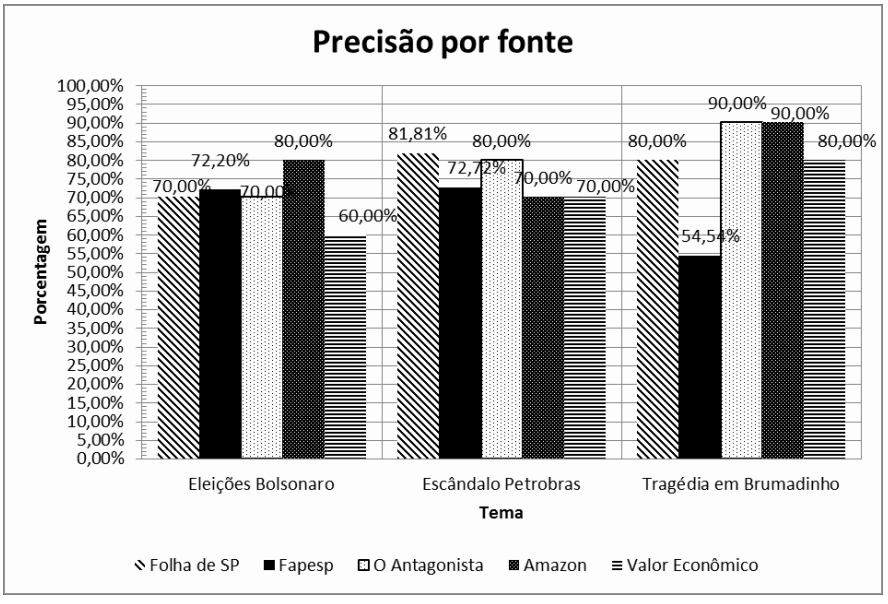
Tabela 4.3 – Comparação entre URLs

Eleições Bolsonaro				
Fonte	**Manual**	**Relevância**	**Rastreada**	**URLs Correspodem?**
Folha de SP	URL1	Relevante	URL1	Sim
	URL2	Relevante	URL2	Sim
	URL3	Relevante	URL3	Não
	URL4	Irrelevante	URL4	Sim
	URL5	Irrelevante	URL5	Não
	URL6	Relevante	URL6	Não
	URL7	Relevante	URL7	Sim
	URL8	Relevante	URL8	Sim
	URL9	Relevante	URL9	Sim
	URL10	Irrelevante	URL10	Não
	URL11	Relevante	URL11	Não
	URL12	Relevante	URL12	Não
	URL13	Irrelevante	URL13	Não
	URL14	Relevante	URL14	Sim
	URL15	Irrelevante	URL15	Sim
	URL16	Relevante	URL16	Não
	URL17	Irrelevante	URL17	Sim
	URL18	Relevante	URL18	Sim

Fonte: Elaboração do autor

A partir das comparações feitas os índices de precisão e revocação foram calculados para cada uma das fontes. O Gráfico 4.1 apresenta a precisão por fonte.

Gráfico 4.1 – Precisão por fonte

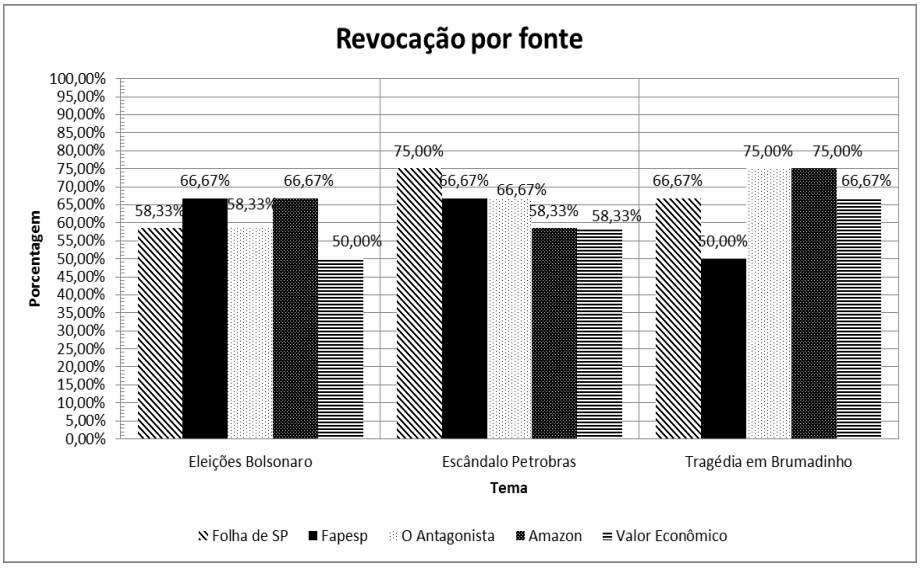


Fonte: Elaboração do autor

Através do Gráfico 4.1 pode-se perceber que a precisão tem uma variação consideravelmente alta de acordo com a fonte utilizada, pois os valores encontram-se no intervalo de 54,54% e 90,00%. Isso ocorre porque em algumas fontes o SisBaseText foi capaz de encontrar um número maior de páginas relevantes do que em outras.

Por outro lado, o Gráfico 4.2 demonstra o índice de revocação com uma variação menor, apresentando valores no intervalo de 50,00% a 75,00%. Isso indica que o sistema conseguiu recuperar mais de 50% dos URLs rotulados como “relevantes”, considerando-se o total de URLs do conjunto de testes.

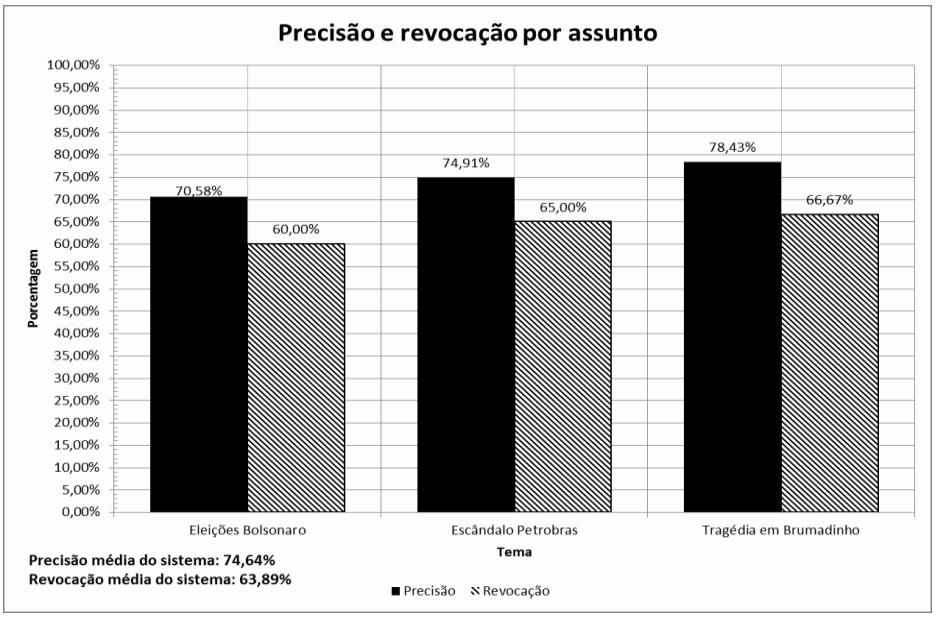
Gráfico 4.2 – Revocação por fonte



Fonte: Elaboração do autor

Os índices também foram calculados considerando-se os três assuntos buscados, conforme representado pelo Gráfico 4.3. A variação da precisão e revocação por assunto é baixa, sendo 7,85% e 6,67%, respectivamente. Isso demonstra que o SisBaseText obteve desempenho semelhante ao buscar páginas sobre cada um dos temas. A partir dos valores obtidos no gráfico foi possível calcular a precisão e revocação média do sistema. A precisão média do sistema é de 74,64% e a revocação média 63,89%.

Gráfico 4. 3 – Precisão e revocação por assunto



Fonte: Elaboração do autor

A última etapa dos testes funcionais foi verificar se os valores de precisão média e revocação média eram condizentes com outros trabalhos envolvendo sistemas que realizam rastreamento Web. Foram analisados os trabalhos de Dong et.al. (2004), Pal et.al. (2009) e Stamatakis et.al. (2014). A Tabela 4.4 apresenta os valores obtidos por esses autores.

Tabela 4.4 – Precisão e Revocação de trabalhos correlatos

Autor	Precisão média	Revocação média
Dong	77,26%	52,65%
Pal	75,25%	25,25%
Stamatakis	45,20%	92,10%
Pedro Artico (autor deste trabalho)	74,64%	63,89%

Fonte: Elaboração do autor

Os valores mostram que o SisBaseText apresenta precisão e revocação médias coerentes com os trabalhos correlatos considerados na comparação. Sua precisão é ligeiramente inferior àquela que foi obtida por Dong et. al. (2004) e Pal et al. (2009) e muito superior ao resultado de Stamatakis et al. (2003). A revocação, por sua vez, é superior àquela que foi obtida por Dong et al. (2003) e muito inferior ao resultado de Stamatakis et al. (2003). Neste caso, deve-se considerar que a alta revocação desse trabalho correlato foi obtida em detrimento de sua precisão, já que esta foi muito inferior àquela obtida pela ferramenta apresentada neste trabalho.

5 CONCLUSÃO

O SisBaseText, conforme explicado ao longo deste trabalho, tem como principal objetivo aplicar a busca Web para possibilitar que uma base textual seja criada. As diversas opções de busca disponibilizadas e as opções de remoção, adição e edição de atributos da base textual revelam o caráter genérico do sistema, que visa atender usuários com diferentes objetivos.

Os testes de precisão e revocação realizados mostraram-se condizentes com trabalhos correlatos, revelando a eficiência do sistema em buscar a informação requerida pelo usuário. Além disso, a quantidade de informação retornada por cada busca mostrou-se suficiente para criar bases que possam ser consideradas passíveis de utilização.

A versão beta, implementada nas linguagens Python e PHP, possui flexibilidade quanto à inserção de novos módulos e a implementação de novas funcionalidades. Trabalhos futuros envolvem a melhoria da ferramenta, trazendo a possibilidade de combinação dos tipos de busca e a importação de bases textuais, para que seja possível criar bases integradas. Além disso, a hospedagem da ferramenta em um servidor Web será importante para que a mesma possa ser disponibilizada e amplamente utilizada.

REFERÊNCIAS

BAEZA-YATES, R., CASTILLO, C. Crawling the infinite Web: five levels are enough. Proceedings of the third Workshop on Web Graphs (WAW), Rome, Italy, Lecture Notes in Computer Science, Springer, vol. 3243, pp. 156-167, 2004.

BAEZA-YATES, R., CASTILLO, C., MARÍN, M., RODRIGUEZ, A. Crawling a country: better strategies than breadth-first for web page ordering. Proceedings of the 14th international conference on World Wide Web (WWW 2005), Chiba, Japan, pp. 864-872, 2005.

BHATT, D., VYAS, D. A., PANDYA, S. Focused Rastreador Web. Institute of Technology, Nirma University. Advances in Computer Science and Information Technology, 2015.

BOOCH, G., RUMBAUGH, J., JACOBSON, I. Unified Modeling Language User Guide (2 ed.). Addison-Wesley Professional, 2005.

BORDES, A., GLOROT, X., WESTON, J., BENGIO, Y. Joint Learning of Words and Meaning Representations for Open-Text Semantic Parsing. AISTATS, 2012.

BRANCO, A., SILVA, J. R. A Suite of Shallow Processing Tools for Portuguese: LX- Suite. University of Lisbon, Department of Informatics, 2006.

BUTKIEWICZ, M., MADHYASHTA, H. V., SEKAR, V. Understanding website complexity: measurements, metrics, and implications. In Proc. of the SIGCOMM conference on Internet Measurement Conference (IMC), 2011.

CARDOSO, P.; PRADO, T.; NUNES, M. Métodos para Sumarização Automática Multidocumento Usando Modelos Semântico-Discursivos. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP), 2001.

CARVALHO, D. R. et al. Ferramenta de pré e pós-processamento para Data Mining. In: SEMINÁRIO DE COMPUTAÇÃO, VII, Blumenau, Brasil, 2003. Anais... Blumenau: FURB.

CHO, J., GARCIA-MOLINA, H. The Evolution of the Web and Implications for an Incremental Rastreador. Department of Computer Science Stanford, CA, 1999.

CHEN, B., BUTTE, A.J. Leveraging big data to transform target selec-tion and drug discovery. Clin. Pharmacol. Ther.99, 285–297, 2016.

COHEN, J. P., DING, W.; Bagherjeiran, A. Semi-Supervised Web Wrapper Repair via Recursive Tree Matching, 2015.

DENIS, F., LEMAY, A., TERLUTTE, A. Learning regular languages using RFSAs. Theoretical Computer Science, 2004.

DIAS, M. A. L., MALHEIROS, M. G. Estudo de Técnicas de Radicalização para a Língua Portuguesa. Lajeado, Rio Grande do Sul, Brasil, 2004.

DONG, P., WONG L., Ng, S., LOH M., MONDRY, A. Quantitative evaluation of recall and precision of CAT Crawler, a search engine specialized on retrieval of Critically Appraised Topics. BMC Med Inform Decis Mak, 2004.

FERNAU, H. Algorithms for learning regular expressions from positive data. Universität Trier, FB 4, Abt. Informatik, Germany, 2005.

FIROIU, L., OATES, T., COHEN, P. Learning regular languages from positive evidence. Computer Science Department, LGRC, University of Massachusetts, Amherst, 2002.

GAUR, R., SHARMA, D. K. Focused Crawling with Ontology using Semi- Automatic Tagging for Relevancy, 2014.

GUO, G., WANG, H., BELL, D., BI, Y., GREER, K. (2004). An kNN Model based Approach and Its Application in Text Categorization. School of Computing and Mathematics, University of Ulster Newtownabbey, Northern Ireland, UK, Vol. 2945/2004.

IMAMURA, C. Y. Pré-processamento para extração de conhecimento de bases textuais. ICMC- USP, 2000.

KARGER, D., OSTLER, S., LEE, R. The web page as a WYSIWYG end-user customizable databasebacked information management application, In Proceedings of the 22nd annual ACM Symposium on User Interface Software and Technology (UIST ‘09), Victoria, BC, Canadá, 2009.

KUMAR, M., BATHIA, R., RATTAN, D. A survey of Rastreador Webs for information retrieval. WIREs Data Mining Knowl Discov, 2017.

LUENGO, J., GUERRERO, S., HERRERA, F. Data Preprocessing in Data Mining. Springer, 2014.

LUTZ, M. Learning Python: Powerful Object-Oriented Programming (English Edition) (Vol. 5). California: O'Reilly Media, 2013.

MATSUBARA, E. T., MARTINS, C. A., MONARD, M. C. PreTexT — uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. Relatório Técnico 209, ICMC- USP, 2003.

MILLER, R. C., BHARAT, K. SPHINX: a framework for creating personal, site-specific Rastreador Webs. Comput Networks ISDN Syst,119–130, 1998.

PAL, A., TOMAR, D. S., SHRIVASTAVA, S.C. Effective Focused Crawling Based on Content and Link Structure Analysis. Department of Computer Science & Engineering. Maulana Azad National Institute of Technology Bhopal, India. 2009.

PARDO, T. Sumarização Automática: Principais Conceitos e Sistemas para o Português Brasileiro. Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - ICMC-USP, 2008.

POPESCU, A. – Information Extraction from Unstructured Web Text - A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy 2007.

PORTER, M. F. Portuguese stemming algorithm. disponível em: http://snowball.tartarus.org/algorithms/portuguese/stemmer.html. Acesso em 15 de outubro de 2019.

POTEET, S., KAO, A. Natural Language Processing and Text Mining (1ª ed.). Springer Science & Business Media, 2007.

ROUILLARD, A. D., GUNDERSEN, G. W., FERNANDEZ, N. F., WANG, Z., MONTEIRO, C. D., MCDERMOTT, M. G., MA'AYAN, A. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database: The Journal of Biological Databases and Curation, 2016.

SAHU, M. B., BHARNE, S. A Survey On Various Kinds Of Rastreador Webs And Intelligent Rastreador. International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, 2016.

SAINI, C., ARORA, V. Information retrieval in rastreamento Web: A survey. International Conference on Advances in Computing Communications and Informatics. (pp. 2635- 2643). Jaipur: ICACCI, 2016.

SOUZA, R. R. Sistemas de Recuperação de Informações e Mecanismos de Busca na web: panorama atual e tendências. Perspect. ciênc. inf., Belo Horizonte, v.11 n.2, p. 161 -173, 2006.

SPARCK JONES, K. What might be in a summary? In Krause Knorz and Womser-Hacker (eds.), Information Retrieval 93, pp. 9-26. Universitatsverlag Konstanz. June, 1993.

STAMATAKIS, K., KARKALETSIS, V., PALIOURAS, G., HORLOCK, J., GROVER, C., CURRAN, J. R., DINGARE, S. Domain-Specific Web Site Identification: The CROSSMARC Focused Web Crawler. ResearchGate, 2014.

TEUFEL, S.; MOENS, M. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. Computational Linguistics, Vol. 28, N. 4, pp. 409-445, 2002.

THE PHP GROUP. O que é PHP? Disponível em Manual do PHP: https://secure.php.net/manual/pt_BR/intro-whatis.php. Acesso em: 10 set. 2019.

YANUAR-FIRDAUS, A. W., SURYANI, A. "Analisis dan Implementasi Focused Crawler Menggunakan Algoritma Fish Pada Web Kesehatan," Universitas Telkom, Bandung, 2012.