Arildo Magno de Macedo

Protótipo de Aplicação para Detecção de Plágio

Formiga - MG

2022

Arildo Magno de Macedo

Protótipo de Aplicação para Detecção de Plágio

Projeto do trabalho de conclusão de curso apresentado ao Instituto Federal Minas Gerais - Campus Formiga.

Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais

Campus Formiga

Ciência da Computação

Formiga - MG

2022

Resumo

O plágio é uma forma de má conduta de pesquisa e uma grave violação das normas da ciência. É a deturpação das ideias ou palavras de outra pessoa como se fossem próprias, sem o devido reconhecimento da fonte original. (ANDERSON; STENECK, 2011). Neste artigo, é utilizado algoritmos de similaridade entre palavras, para detectar a probabilidade de plágio entre documentos. Composto de processamento de dados, lógica fuzzy, bancos de dados e cálculos de similaridade é possível obter o grau de semelhança entre os documentos, detectando até plágios ofuscados. Na conclusão deste projeto é pretendido ter um protótipo de sistema que retorne ao usuário a análise de seus arquivos com a probabilidade de plágio entre os mesmos e as respectivas sentenças possivelmente plagiadas.

Palavras-chave: Plágio, Plágio Ofuscado, Lógica Fuzzy, Análise de Arquivos, Banco de dados Lexico, Similaridade.

Sumário

1 Introdução

O plágio no sentido de “roubo de propriedade intelectual” existe desde que os humanos produziram obras de arte e pesquisa. No entanto, o fácil acesso à Web, grandes bancos de dados e telecomunicações em geral, tornou o plágio um sério problema para editores, pesquisadores e instituições de ensino (MAURER; KAPPE; ZAKA, 2006).

A detecção de plágio é essencialmente uma tarefa dificil pois uma palavra pode ter vários signficados e sentidos possíveis. O plágio mais comum é o literal, que trata apenas de copiar a informação de determinado local e substituir em outro sem atribuir os devidos direitos autorais. Já o plágio ofuscado é mais complexo de ser detectado pois os textos plagiados são transformados em palavras e estruturas diferentes (ALZAHRANI; SALIM; PALADE, 2015).

O objetivo deste projeto é utilizar processamento de dados, lógica fuzzy, banco de dados léxico[1], e cálculos de similaridade para analisar a similaridade entre arquivos. Foram utilizadas metodologias já abordadas nos estudos de (WU; PALMER, 1994), (YERRA; NG, 2005) e (ALZAHRANI; SALIM; PALADE, 2015) para atender ao objetivo aqui fixado. Além de adaptação em métodos propostos nos trabalhos (ALZAHRANI; SALIM; PALADE, 2015) e (EZZIKOURI et al., 2018). O protótipo será desenvolvido para a plataforma “web” visando atingir a maior quantidade de usuários. Na plataforma “web” será fornecido ao usuário a possibilidade de enviar seus arquivos, e então verificar de forma visual a similaridade entre eles e suas respectivas sentenças.

Na literatura é possível encontrar alguns trabalhos que abordam o assunto, como o de (YERRA; NG, 2005) que utiliza de lógica fuzzy para detectar a similaridade entre páginas HTML, (EZZIKOURI et al., 2018) e sua abordagem ao utilizar lógica fuzzy aliada a banco de dados léxicos para detectar "cross-language-plagiarism"poderão ser úteis para o trabalho.

Tecnologias de programação web, como Django, Python, JavaScript, React, HTML e CSS. Conhecimento de banco de dados, pré-processamento textual e técnicas de análise de similaridade, foram definidos como base e fundamento tecnológico para o desenvolvimento deste trabalho.

Na próxima seção é apresentado os trabalhos relacionados, na seção 2 é exibida a fundamentação teórica, na seção 3 é exibido como será realizado o desenvolvimento do

[1] Bancos de dados lexicais contêm informações estruturadas sobre palavras de um idioma.

trabalho bem como as tecnologias pretendidas a serem utilizadas, na seção 4 o cronograma para o desenvolvimento do trabalho, e na seção 5 a conclusão do que se espera que seja obtido com o trabalho.

1.1 Justificativa

o plágio sempre foi um problema no meio acadêmico e parece estar aumentando, plágio é um problema não somente no meio acadêmico mas em qualquer área (MAURER; KAPPE; ZAKA, 2006). O que torna excepcionalmente necessário que o assunto receba mais atenção e que seja devidamente tratado. Neste contexto, a justificativa deste trabalho é: gerar um sistema que auxilie um usuário, a avaliar a probabilidade de haver plágio entre arquivos.

O desenvolvimento da presente proposta culminará na experiência prática com variadas tecnologias abordadas, porém não aprofundadas em cursos superiores de tecnologia da informação.

1.2 Objetivos

1.2.1 Objetivo Geral

Apresentar um sistema que realize a análise entre arquivos e retorne a similaridade entre eles.

1.2.2 Objetivos Específicos

- Utilizar técnicas de pré-processamento textual.
- Empregar lógica fuzzy na detecção de similaridade entre documentos.
- Empregar técnicas de análise de similaridade entre textos.
- Manipular dados em banco de dados léxicos.
- Criar um sistema web que receba arquivos e avalie a similaridade dos mesmos.

1.3 Trabalhos Relacionados

Zadeh (1965) definiu que um conjunto fuzzy é uma classe de objetos que contém graus de pertinencia. Onde o conjunto é caracterizado por uma função de pertinência que atribui a cada objeto dele um grau que varia de 0 a 1.

Kraft e Buell (1983) fizeram um trabalho substancial utilizando os subconjuntos fuzzy para recuperação de informação, realizando consultas booleanas em documentos.

Miyamoto e Nakayama (1986) proposuseram a recuperação de informação bibliográfica fuzzy baseada em thesaurus fuzzy e em pseudothesaurus, em seu trabalho as relações fuzzy são geradas a partir de um modelo de conjunto fuzzy que descreve a associação de uma palavra-chave aos seus conceitos.

Miyamoto (1990) propõe um modelo de conjunto fuzzy para recuperação de informação desenvolvendo métodos e algoritmos para recuperação de informação baseados no modelo de conjunto fuzzy.

Ogawa, Morita e Kobayashi (1991) propuseram um sistema fuzzy de recuperação de documentos utilizando uma matriz de conexão de palavras-chave para representar semelhanças entre palavras-chave.

Armstrong (1993) realizaram um trabalho no estudo do que é plágio, concluíram que o plágio engloba um espectro de ações em que o crédito é desviado. Pode incluir levantamento literal direto de passagens sem atribuição; reformulação de ideias do original no próprio estilo do suposto autor; paráfrase não creditada do trabalho de outra pessoa.

Brin, Davis e García-Molina (1995a) desenvolveram uma nova abordagem utilizando um sistema de biblioteca digital. Propuseram um sistema de registro de documentos e detecção de cópias, sejam cópias completas ou cópias parciais, descreveram algoritmos para tal detecção e métricas necessárias para avaliar os mecanismos de detecção (abrangendo precisão, eficiência e segurança), em seu trabalho descrevem um protótipo funcional chamado COPS.

Shivakumar e Garcia-Molina (1995) apresentaram um novo esquema para detecção de cópias baseado na comparação das ocorrências de frequência de palavras do novo documento com as de documentos registrados, chamado de SCAM. Também foi feita uma comparação experimental entre o novo esquema de detecção proposto e o COPS (BRIN; DAVIS; GARCÍA-MOLINA, 1995b).

Campbell, Chen e Smith (2000) Apresentaram um sistema de detecção de cópia para automatizar a detecção de duplicação em documentos digitais baseado em sentenças.

Yerra e Ng (2005) propuseram uma nova abordagem para detectar documentos Web semelhantes, especialmente documentos HTML. A abordagem determina a razão de chances de dois documentos quaisquer fazendo uso dos graus de semelhança dos documentos

e exibe as localizações de sentenças semelhantes detectadas nos documentos.

Zhang et al. (2010) apresentaram um novo algoritmo para realizar a tarefa de detecção de duplicatas parciais. Além das semelhanças entre documentos, o algoritmo pode localizar simultaneamente as partes duplicadas. A idéia principal é dividir a tarefa de detecção de duplicatas parciais em duas subtarefas: detecção de quase duplicatas em nível de sentença e correspondência de sequência.

Alzahrani, Salim e Palade (2015) elaboraram um detector de plágio para a lingua inglesa que trata os casos de plágio altamente ofuscados. Um modelo de similaridade baseado em semântica difusa, e banco de dados léxico é apresentado.

Ezzikouri et al. (2018) propuseram detector de similaridade semântica difusa para CLPD(cross-language-plagiarism-detection) usando a taxonomia WordNet e três abordagens semânticas Wu e Palmer, Lin e Leacock-Chodorow para documentos árabes.

Além disto, atualmente soluções semelhantes ao projeto proposto podem ser encontrados na web, porém, nenhuma delas realizam análise em multiplos arquivos simultaneamente. Tais como:

- **CheckPlagiarism**: Verifica a semelhança de texto entre duas urls ou em dois arquivos e mostra o conteúdo correspondente (CHECKPLAGIARISM, s.d).
- **Prepostseo**: Compara dois documentos ou páginas da web lado a lado para descobrir o conteúdo plagiado (PREPOSTSEO, s.d).
- **CopyLeaks**: Realiza a comparação de dois documentos de texto que podem estar em formatos diferentes (COPYLEAKS, s.d).

2 Fundamentação Teórica

Bancos de dados ou bases de dados são conjuntos de arquivos relacionados entre si com registros sobre pessoas, lugares ou coisas. São coleções organizadas de dados que se relacionam de forma a criar algum sentido informação e dar mais eficiência durante uma pesquisa ou estudo científico (HEUSER, 2009).

Um lexicon é o vocabulário de uma língua ou ramo do conhecimento. Em linguística, um lexicon é o inventário de lexemas de uma língua (FELDMAN; SANDRA; TAFT, 1999).

Um lexema é uma unidade abstrata básica de significado, uma unidade de análise morfológica em linguística que corresponde aproximadamente a um conjunto de formas assumidas por uma única palavra raiz. Por exemplo, em inglês, run, runs, run e running são formas do mesmo lexema, que pode ser representado como run (CRYSTAL, 1995).

A lógica difusa ou lógica fuzzy é a forma de lógica multivalorada, na qual os valores verdade das variáveis podem ser qualquer número real entre 0 correspondente ao valor falso e 1 correspondente ao valor verdadeiro (ARTERO, 2009).

Em lógica, a lógica multivalorada ou lógica plurivalente é um cálculo proposicional em que há mais de dois valores verdade. Tradicionalmente, na Lógica aristotélica, existem apenas dois possíveis valores isto é, "verdadeiro"e "falso"para cada proposição (HURLEY, 2006).

Lógica é o estudo do raciocínio correto ou bons argumentos. Muitas vezes é definido em um sentido mais estrito como a ciência de inferências dedutivamente válidas ou de verdades lógicas (HURLEY, 2006).

Desenvolvimento web é o termo utilizado para descrever o desenvolvimento de sites, na Internet ou numa intranet (CAMPBELL, 2017).

Processamento textual refere-se à teoria e prática de automatizar a criação ou manipulação de texto eletrônico. O termo processamento refere-se ao processamento automatizado ou mecanizado. Usando processamento de linguagem natural NLP e aprendizado de máquina, subcampos da inteligência artificial, as ferramentas de processamento de texto são capazes de entender automaticamente a linguagem humana e extrair valor dos dados de texto (JAMES, 1994).

O processamento de linguagem natural NPL refere-se ao ramo da ciência da computação preocupado em dar aos computadores a capacidade de entender texto e palavras.

Taxonomia é a disciplina que define os grupos de organismos com base em características comuns e dá nomes a esses grupos (TAXONOMY, s.d).

Hiperônimo é uma palavra que pertence ao mesmo campo semântico de outra mas com o sentido mais abrangente, podendo ter várias possibilidades para um único hipônimo. Por exemplo, a palavra flor está associada a todos os tipos de flores: rosa, dália, violeta, etc (HIPERONIMOS, s.d).

Hipônimo têm sentido mais restrito que os hiperônimos, ou seja, hipônimo é um vocábulo mais específico. Por exemplo: Observar, examinar, olhar, enxergar são hipônimos de ver (HIPONIMOS, s.d).

Um Synset é um grupo de elementos de dados que são considerados semanticamente equivalentes para fins de recuperação de informações. Um synset ou conjunto de sinônimos é definido como um conjunto de um ou mais sinônimos que são intercambiáveis em algum contexto sem alterar o valor de verdade da proposição na qual estão inseridos.

Path Similarity é o cálculo de semelhança baseada em caminho, é uma medida de similaridade que utiliza as informações do caminho mais curto entre dois synsets em uma WordNet.(SIMILARITYPATH, s.d)

Wu Palmer Similarity, é um algoritmo de Path Similarity que calcula o parentesco de dois synsets nas taxonomias WordNet. A pontuação de parentesco pode ser entre 0 e 1. Seu cálculo de semelhança é feito com base em quão semelhantes são os sentidos das palavras e onde os Synsets ocorrem em relação uns aos outros na árvore de hiperônimos. (SIMILARITYPATH, s.d)

WordNet é um banco de dados lexical de relações semânticas entre palavras. (WORDNETGLOBAL, s.d). WordNet vincula palavras em relações semânticas, incluindo sinônimos, hipônimos e merônimos. Os sinônimos são agrupados em synsets com definições curtas e exemplos de uso. WordNet pode assim ser visto como uma combinação e extensão de um dicionário e tesauro.(WORDNET, s.d).

OpenWordnet-PT é um wordnet de acesso aberto para português, originalmente desenvolvido como uma projeção sintática do Universal WordNet (UWN) de Melo e Weikum. (PAIVA; RADEMAKER; MELO, 2012)

3 Metodologia

3.0.1 Estudo das ferramentas

Para dar início no desenvolvimento do trabalho, será necessário realizar um estudo mais aprofundado em manipulação de dados.

3.0.2 Desenvolvimento do sistema

Inicialmente, o algoritmo proposto será desenvolvido para a plataforma web usando frameworks e bibliotecas de Javascript e Python como React e Django, podendo sofrer alterações no decorrer do desenvolvimento. A princípio, foi definido tais frameworks e linguagens pela sua curva de aprendizado, sua vasta documentação e integração com bibliotecas de manipulação de dados. Deste modo, visando fornecer ao usuário uma interface para enviar seus arquivos, onde eles passarão pelos processos de processamento de dados, cálculo de similaridade das sentenças, cálculo da similaridade entre os documentos, e geração do log final com a porcentagem de probabilidade de plágio dos arquivos e as sentenças possivelmente plagiadas como é demonstrado na figura 1.

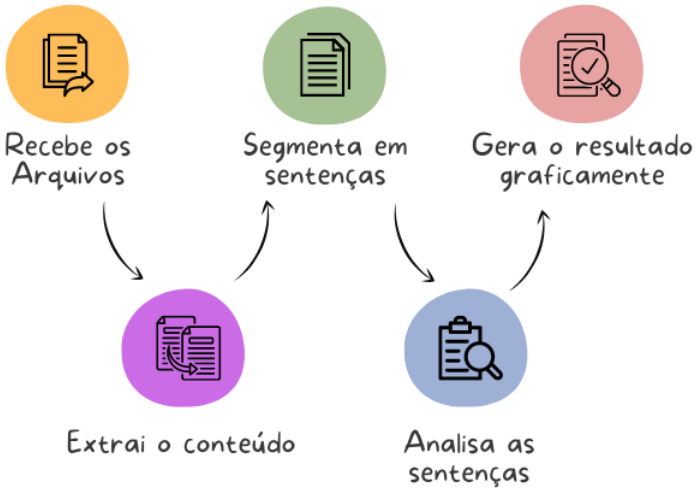

Figura 1 – Estrutura geral do modelo para detecção de plágio

3.0.3 Estudo das características quantitativas

Nessa etapa, o autor realizará comparações entre diversos arquivos e analisará a corretude dos resultados.

3.1 Materiais

Para o desenvolvimento do módulo web, será necessário a utilização de algumas tecnologias de programação, desenvolvimento de interfaces, bancos de dados, e bibliotecas, tais como:

- **Cascading Style Sheets (CSS)**: mecanismo para adicionar estilos a um documento web.
- **JavaScript**: linguagem de programação interpretada, capaz de executar scripts do lado do cliente, sem a necessidade do script ser executado pelo servidor.
- **HyperText Markup Language (HTML)**: linguagem de marcação utilizada na construção de páginas na Web. Documentos HTML podem ser interpretados por navegadores. A tecnologia é fruto da junção entre os padrões HyTime e SGML.
- **React**: biblioteca JavaScript de código aberto com foco em criar interfaces de usuário em páginas web.
- **Recharts**: biblioteca de gráficos combináveis construída em componentes React.
- **Echarts**: biblioteca de gráficos e visualização para navegador.
- **Material-UI**: oferece um conjunto abrangente de ferramentas de interface do usuário para ajudá-lo a enviar novos recursos mais rapidamente.
- **Python**: linguagem de programação de alto nível, interpretada de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte.
- **Django**: framework para desenvolvimento rápido para web, escrito em Python, que utiliza o padrão model-template-view.
- **Banco de dados**: são conjuntos de arquivos relacionados entre si com registros sobre pessoas, lugares ou coisas.
- **SQLite**: biblioteca em linguagem C que implementa um banco de dados SQL embutido.
- **OpenWordnet-PT**: Wordnet de Acesso Aberto para Português.
- **Wn Library**: uma biblioteca Python para explorar informações em Wordnets.

4 Cronograma

Etapa	Descrição	Mês								
		3	4	5	6	7	8	9	10	11
1	Estudo Geral	X								
3	Implementação dos Algoritmos		X	X						
4	Implementação Sistema Web				X	X	X			
5	Testes no Sistema Web						X			
7	Escrita da Monografia						X	X	X	
8	Preparar Apresentação									X

5 Resultados Esperados

Espera-se, que ao final deste projeto, tenha-se um sistema que realize uma análise da probabilidade de haver plágio entre arquivos. E que o autor obtenha conhecimento mais aprofundado sobre tais temas.

Referências

ALZAHRANI, S. M.; SALIM, N.; PALADE, V. Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. *Journal of King Saud University - Computer and Information Sciences*, v. 27, n. 3, p. 248–268, 2015. ISSN 1319-1578. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1319157815000361>. Citado 2 vezes nas páginas 4 e 7.

ANDERSON, M. S.; STENECK, N. H. The problem of plagiarism. *Urologic Oncology: Seminars and Original Investigations*, v. 29, n. 1, p. 90–94, 2011. ISSN 1078-1439. Plagiarism. Disponível em: <https://www.sciencedirect.com/science/article/pii/S107814391000270X>. Citado na página 2.

ARMSTRONG, J. D. Plagiarism: what is it, whom does it offend, and how does one deal with it? *AJR. American journal of roentgenology*, v. 161 3, p. 479–84, 1993. Citado na página 6.

ARTERO, A. O. *Inteligência Artificial: Teórica e Prática*. São Paulo: Livraria da Física, 2009. Citado na página 8.

BRIN, S.; DAVIS, J.; GARCÍA-MOLINA, H. Copy detection mechanisms for digital documents. In: *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: Association for Computing Machinery, 1995. (SIGMOD '95), p. 398–409. ISBN 0897917316. Disponível em: <https://doi.org/10.1145/223784.223855>. Citado na página 6.

BRIN, S.; DAVIS, J.; GARCÍA-MOLINA, H. Copy detection mechanisms for digital documents. Association for Computing Machinery, New York, NY, USA, v. 24, n. 2, p. 398–409, may 1995. ISSN 0163-5808. Disponível em: <https://doi.org/10.1145/568271.223855>. Citado na página 6.

CAMPBELL, D. M.; CHEN, W. R.; SMITH, R. D. Copy detection systems for digital documents. In: *Proceedings of the IEEE Advances in Digital Libraries 2000*. USA: IEEE Computer Society, 2000. (ADL '00), p. 78. ISBN 0769506593. Citado na página 6.

CAMPBELL, J. *Web Design: Introductory*. 20 Channel Center Street Boston, MA 02110 USA: Cengage Learning, 2017. Citado na página 8.

CHECKPLAGIARISM. s.d. <https://www.check-plagiarism.com/plagiarism-comparison-search>. Accessed: 2022-06-23. Citado na página 7.

COPYLEAKS. s.d. <https://app.copyleaks.com/text-compare>. Accessed: 2022-06-23. Citado na página 7.

CRYSTAL, D. *The Cambridge Encyclopedia of the English Language*. Cambridge University Press Assessment Shaftesbury Road Cambridge CB2 8EA: Cambridge University Press, 1995. 118 p. ISBN 0521401798. Citado na página 8.

EZZIKOURI, H. et al. Fuzzy cross language plagiarism detection (arabic-english) using wordnet in a big data environment. In: *Proceedings of the 2018 2nd International*

Conference on Cloud and Big Data Computing. New York, NY, USA: Association for Computing Machinery, 2018. (ICCBDC'18), p. 22–27. ISBN 9781450364744. Disponível em: <https://doi.org/10.1145/3264560.3264562>. Citado 2 vezes nas páginas 4 e 7.

FELDMAN, L.; SANDRA, D.; TAFT, M. Morphological structure, lexical representation and lexical access. *The American Journal of Psychology*, v. 111, p. 445, 11 1999. Citado na página 8.

HEUSER, C. A. *Projeto de Banco de Dados*. Porto Alegre: Bookman, 2009. Citado na página 8.

HIPERONIMOS. s.d. <https://mundoeducacao.uol.com.br/gramatica/hiponimos-hiperonimos-holonimos-meronimos.htm>. Accessed: 2022-06-22. Citado na página 9.

HIPONIMOS. s.d. <https://brasilescola.uol.com.br/gramatica/hiponimos-hiperonimos.htm>. Citado na página 9.

HURLEY, P. *A Concise Introduction to Logic*. Werner Siemens: Cengage Learning, 2006. Citado na página 8.

JAMES, A. *Natural Language Understanding*. 330 Hudson in New York City, New York USA: Pearson, 1994. Citado na página 8.

KRAFT, D. H.; BUELL, D. A. Fuzzy sets and generalized boolean retrieval systems. *Int. J. Man Mach. Stud.*, v. 19, p. 45–56, 1983. Citado na página 6.

MAURER, H.; KAPPE, F.; ZAKA, B. Plagiarism – a survey. *Journal of Universal Computer Science*, Verlag der Technischen Universität Graz, v. 12, n. 8, p. 1050–1084, 2006. ISSN 0948-695X. Citado 2 vezes nas páginas 4 e 5.

MIYAMOTO, S. Information retrieval based on fuzzy associations. *Fuzzy Sets and Systems*, v. 38, n. 2, p. 191–205, 1990. ISSN 0165-0114. Fuzzy Information and Database Systems. Disponível em: <https://www.sciencedirect.com/science/article/pii/016501149090149Z>. Citado na página 6.

MIYAMOTO, S.; NAKAYAMA, K. Fuzzy information retrieval based on a fuzzy pseudothesaurus. *IEEE Transactions on Systems, Man, and Cybernetics*, v. 16, n. 2, p. 278–282, 1986. Citado na página 6.

OGAWA, Y.; MORITA, T.; KOBAYASHI, K. A fuzzy document retrieval system using the keyword connection matrix and a learning method. *Fuzzy Sets and Systems*, v. 39, n. 2, p. 163–179, 1991. ISSN 0165-0114. Applications of fuzzy systems theory. Disponível em: <https://www.sciencedirect.com/science/article/pii/016501149190210H>. Citado na página 6.

PAIVA, V. de; RADEMAKER, A.; MELO, G. de. Openwordnet-pt: An open Brazilian Wordnet for reasoning. In: *Proceedings of COLING 2012: Demonstration Papers*. Mumbai, India: The COLING 2012 Organizing Committee, 2012. p. 353–360. Published also as Techreport http://hdl.handle.net/10438/10274. Disponível em: <http://www.aclweb.org/anthology/C12-3044>. Citado na página 9.

PREPOSTSEO. s.d. <https://www.prepostseo.com/plagiarism-comparison-search>. Accessed: 2022-06-23. Citado na página 7.

SHIVAKUMAR, N.; GARCIA-MOLINA, H. Scam: A copy detection mechanism for digital documents. *proc DL*, 1995. Citado na página 6.

SIMILARITYPATH. s.d. <https://wn.readthedocs.io/en/latest/api/wn.similarity.html>. Accessed: 2022-06-22. Citado na página 9.

TAXONOMY. s.d. <https://www.etymonline.com/word/taxonomy>. Accessed: 2022-06-22. Citado na página 9.

WORDNET. s.d. <https://wordnet.princeton.edu/news-0>. Accessed: 2022-06-22. Citado na página 9.

WORDNETGLOBAL. s.d. <http://globalwordnet.org/resources/wordnets-in-the-world>. Accessed: 2022-06-22. Citado na página 9.

WU, Z.; PALMER, M. Verb semantics and lexical selection. *ArXiv*, abs/cmp-lg/9406033, 1994. Citado na página 4.

YERRA, R.; NG, Y.-K. Detecting similar html documents using a fuzzy set information retrieval approach. In: *2005 IEEE International Conference on Granular Computing*. [S.l.: s.n.], 2005. v. 2, p. 693–699 Vol. 2. Citado 2 vezes nas páginas 4 e 6.

ZADEH, L. Fuzzy sets. *Information and Control*, v. 8, n. 3, p. 338–353, 1965. ISSN 0019-9958. Disponível em: <https://www.sciencedirect.com/science/article/pii/S001999586590241X>. Citado na página 6.

ZHANG, Q. et al. Efficient partial-duplicate detection based on sequence matching. In: . New York, NY, USA: Association for Computing Machinery, 2010. (SIGIR '10), p. 675–682. ISBN 9781450301534. Disponível em: <https://doi.org/10.1145/1835449.1835562>. Citado na página 7.