Arildo Magno de Macedo

Protótipo de Aplicação para Detecção de Plágio Ofuscado

Formiga - MG

2022

Arildo Magno de Macedo

Protótipo de Aplicação para Detecção de Plágio Ofuscado

Monografia do 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

Com o crescimento da Internet, um problema já existente se agravou, cópias de documentos tornam-se cada vez mais comuns. Neste artigo, propomos um algoritmo para detectar a probabilidade de plágio entre documentos. Composto de pré-processamento dados, conjuntos nebulosos, bancos de dados léxicos 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 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, Conjuntos Nebulosos, Análise de Arquivos, Banco de dados Lexico, Similaridade.

Abstract

With the growth of the Internet, an existing problem has worsened, copies of documents become more and more common. In this article, we propose an algorithm to detect the probability of plagiarism between documents. Composed of pre-processing data, fuzzy sets, lexical databases and similarity calculations, it is possible to obtain the degree of similarity between documents, even detecting obfuscated plagiarism. At the conclusion of this project, it is intended to have a system that returns to the user the analysis of their files with the probability of plagiarism between them and the respective sentences possibly plagiarized.

Keywords: Plagiarism, Obfuscated Plagiarism, Fuzzy Sets, Archive Analysis, Lexical Database, Similarity.

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 propósito deste projeto é utilizar processamento de dados, conjuntos nebulosos, banco de dados léxico, e cálculos de similaridade para analisar a similaridade entre arquivos. A abordagem se baseia na semelhança entre as palavras e utilizando conjuntos nebulosos e um banco dados léxico da lingua portuguesa pode ser obtido a similaridade das mesmas. O sistema 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 obter a porcentagem de 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 conjuntos nebulosos para detectar similaridade entre páginas HTML, (EZZIKOURI et al., 2018) e sua abordagem ao utilizar conjuntos nebulosos e banco de dados léxicos para detectar "cross-language-plagiarism"poderão ser de grande vália para o trabalho.

Para o desenvolvimento deste trabalho, será necessário conhecimento em tecnologias de programação web, como Django, Python, JavaScript, HTML e CSS, conhecimento de banco de dados, recuperação da informação, e noção matemática.

Na próxima seção é apresentada a fundamentação teórica, na seção 3 é exibido como será realizado o desenvolvimento do 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 qo 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, e com o fácil acesso à Web, grandes bancos de dados e telecomunicações em geral, tornou o plágio um sério problema (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 deste trabalho no cunho pessoal irá possibilitar um avanço no conhecimento, tanto em programação quanto em manipulação de dados e conjuntos nebulosos.

1.2 Objetivos

1.2.1 Objetivo Geral

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

1.2.2 Objetivos Específicos

- Utilizar técnicas de analise de dados em arquivos.
- Empregar conjuntos nebulosos na detecção de similaridade entre documentos.
- Manipular dados em banco de dados léxicos.
- Criar um sistema web que receba arquivos e julgue os mesmos.

2 Fundamentação Teórica

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.

3 Metodologia

3.1 Desenvolvimento

3.1.1 Estudo das ferramentas

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

3.1.2 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 facilidade que elas apresentam, sua vasta documentação e integração com bibliotecas de manipulação de dados. Para a criação do menu de interação será feito o uso das ferramentas de desenvolvimento web, como JavaScript, HTML, CSS, React e Django. O objetivo é construir um painel amigável e de fácil utilização.

3.1.3 Estudo das características qualitativas

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

3.2 Materiais

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

3.2.1 CSS

Cascading Style Sheets (CSS) é um mecanismo para adicionar estilos a um documento web;

3.2.2 JavaScript

JavaScript: É uma linguagem de programação interpretada, capaz de executar scripts do lado do cliente, sem a necessidade do script ser executado pelo servidor.

3.2.3 HTML

HTML é uma 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.

3.2.4 Django

Django é um framework para desenvolvimento rápido para web, escrito em Python, que utiliza o padrão model-template-view.

3.2.5 React

O React é uma biblioteca JavaScript de código aberto com foco em criar interfaces de usuário em páginas web.

3.2.6 Banco de dados

Bancos de dados ou bases de dados são conjuntos de arquivos relacionados entre si com registros sobre pessoas, lugares ou coisas.

4 Cronograma

Etapa	Descrição	Mês									
		3	4	5	6	7	8	9	10	11	12
1	Estudo Geral	X									
2	Implementação Algoritimos		X								
3	Testes Algoritmos		X								
4	Implementação FrontEnd			X							
5	Implementação Backend			X							
6	Testes Sistema Web				X	X					
7	Escrita Monográfica				X	X	X	X	X	X	
8	Testes Stress Web						X	X			
9	Preparar Apresentação									X	X

5 Resultados Esperados

Espera-se, ao final deste trabalho de conclusão de curso, 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 5 e 8.

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 7.

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 7.

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 7.

CAMPBELL, D.; CHEN, W.; SMITH, R. Copy detection systems for digital documents. In: . [S.l.: s.n.], 2000. p. 78–88. ISBN 0-7695-0659-3. Citado na página 7.

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 5 e 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 7.

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 5 e 6.

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 7.

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 7.

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 7.

SHIVAKUMAR, N.; GARCIA-MOLINA, H. Scam: A copy detection mechanism for digital documents. In: *DL*. [S.l.: s.n.], 1995. Citado na página 7.

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 5 e 8.

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 7.

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 8.