Arildo Magno de Macedo

Protótipo de aplicação para análise de plágio entre múltiplos arquivos

Formiga - MG

2021

Arildo Magno de Macedo

Protótipo de aplicação para análise de plágio entre múltiplos arquivos

Monografia do pré-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

2021

Resumo

Com a inserção da tecnologia no âmbito acadêmico mostrou-se a possibilidade de realizar diversas atividades de maneira online e com isto alguns problemas se destacaram, dentre eles o plágio. Sendo assim, tornou-se de suma importância que este problema tenha um tratamento adequado. Examinando os arquivos e empregando técnicas de recuperação da informação, é analisada a probabilidade de plágio entre eles. Portanto é pretendido que na conclusão deste projeto tenha-se um sistema que retorne ao usuário a análise de seus arquivos com a probabilidade de plágio entre os mesmos.

Palavras-chave: Plágio, Projeto Conclusão Curso.

Abstract

With the insertion of technology in the academic environment, the possibility of carrying out various activities online was shown, and with this some problems stood out, including plagiarism. Therefore, it has become extremely important that this problem has an adequate treatment. By examining the files and employing information retrieval techniques, the likelihood of plagiarism among them is analyzed. Therefore, it is intended that at the conclusion of this project there is a system that returns to the user the analysis of their files with the probability of plagiarism among them.

Keywords: Plagiarism, Project Completion Course.

Sumário

1 Introdução

Conforme (MOOERS, 1951) Recuperação de informação é o nome dado ao processo ou método pelo qual um potencial usuário de informação é capaz de converter a sua necessidade de informação em uma lista real de citações a documentos em um acervo contendo informações úteis para ele. Para [Saracevic 1999], a Recuperação de Informação pode ser considerada a vertente tecnológica da Ciência da Informação e é resultado da relação desta com a Ciência da Computação. É notável o crescimento da tecnologia e sua inclusão em todo ambiente, na esfera acadêmica isto trouxe a possibilidade de realizar diversas atividades de maneira online, e com isso alguns problemas se destacaram, dentre eles o plágio.

Sendo assim, quando se trata da comparação entre dois ou mais arquivos, temos poucas alternativas no meio acadêmico para que os docentes possam utilizá-los para lhes auxiliar na correção de trabalhos. Neste contexto, o propósito deste projeto de um trabalho de conclusão de curso, é utilizar técnicas de recuperação de informação e analisar a similaridade entre dois ou mais arquivos. E visando a falta de alternativas para tal área, vamos desenvolver um produto com o foco mercadológico, também iremos desenvolver a plataforma para a web para que assim possamos atingir a maior parte de usuários possíveis e que seja de mais fácil acesso. Na plataforma web visamos fornecer ao usuário a possibilidade de enviar seus arquivos e que se obtenha a similaridade entre eles.

Na literatura podemos encontrar alguns trabalhos que abordam o assunto, como o de (H. et al., 2021) que utiliza da fórmula de TF-IDF aliada a inteligência artificial para realizar a detecção de plágio. E seu modo de analisar os dados via a fórmula de TF-IDF poderá ser de grande valia no presente trabalho. Para o desenvolvimento deste trabalho, será necessário conhecimento em tecnologias de programação web, como PHP/Laravel, JavaScript, HTML e CSS. Conhecimento de banco de dados e outras linguagens como Python.

Na seção 2 é apresentada a fundamentação teórica como as tecnologias utilizadas, na seção 3 é exibido como será realizado o desenvolvimento do trabalho e na seção 4 a conclusão que se espera que seja obtida do trabalho.

1.1 Justificativa

Como é dito em (P. et al., 2017) o plágio sempre foi um problema no meio acadêmico e parece estar aumentando. Sendo assim é excepcionalmente necessário que o assunto receba mais atenção e que seja devidamente tratado. Os docentes recebem diversos arquivos para serem avaliados, sendo assim é pertinente que se tenha alguma maneira de que se possa realizar uma análise de tais trabalhos buscando encontrar a probabilidade de plágio entre eles. Neste contexto, a justificativa deste trabalho é: gerar um sistema que auxilie um docente, a avaliar a probabilidade de haver plágio em seus arquivos.

O desenvolvimento deste trabalho no cunho pessoal irá possibilitar um avanço no conhecimento, tanto em programação quanto em técnicas de recuperação de informação. Além da possibilidade de lançar o sistema como um sistema comercial.

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver um sistema que realize a análise entre diversos arquivos e que retorne a probabilidade de haver algum plágio entre eles.

1.2.2 Objetivos Específicos

- Utilizar técnicas de recuperação de informação para analisar diversos arquivos.
- Criar um sistema web que receba diversos arquivos e que trate eles para que sejam analisados.

2 Fundamentação Teórica

A Recuperação de informação foi definida em (MOOERS, 1951) da seguinte forma: É o nome dado ao processo ou método pelo qual um potencial usuário de informação é capaz de converter a sua necessidade de informação em uma lista real de citações a documentos em um acervo contendo informações úteis para ele.

Em (COOPER, 1971) acabou por determinar o que seria o conceito de relevância, que é crucial na área de Recuperação de Informação, sendo muitas vezes utilizado na própria enunciação dos objetivos dessa área.O conceito de relevância foi passado a ser utilizado como uma medida de eficácia por (T., 1975).

No (M.K., 1991) foi estabelecido o termo documento que usamos na recuperação de informação como: um objeto que contém informação um documento.

Em (T., 1999) a Recuperação de Informação passou a ser considerada a vertente tecnológica da Ciência da Informação e é resultado da relação desta com a Ciência da Computação.

Mais recentemente (COADIC, 2004) definiu documento é o termo genérico que designa os objetos portadores de informação. Um documento é todo artefato que representa ou expressa um objeto, uma ideia ou uma informação por meio de signos gráficos e icônicos (palavras, imagens, diagramas, mapas, figuras, símbolos), sonoros e visuais (gravados em suporte de papel ou eletrônicos).

No (ARMSTRONG, 1991) o plágio foi descrito como um ato que abrange um espectro de ações onde o crédito original é desviado.

Sendo assim é possível que se utilize um documento na recuperação de informação para que represente um arquivo e seus dados, e através de sua representação seja possível analisar seu conteúdo com outros arquivos em busca de evidências de plágio.

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 algumas linguagens de programação como Php e em algumas técnicas de recuperação de informação, como tf-idf. As ferramentas de tf-idf já foram apresentadas na disciplina de 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 Php, como Next e React para o front-end e Laravel para o back-end, podendo sofrer alterações no decorrer do desenvolvimento. A princípio, foi definido tais frameworks e linguagens pela facilidade que elas apresentam e sua vasta documentação. Como apoio no desenvolvimento, será utilizado a linguagem de programação Python que possibilita o desenvolvimento de aplicações de forma rápida e segura, incentivando o uso das boas práticas de programação. Para a criação do painel de controle será feito o uso das ferramentas de desenvolvimento web, como JavaScript, HTML, CSS, React e Next. 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 do resultado de cada um.

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

É uma folha de estilo em cascata, que é utilizada para definir a aparência em páginas web, que utilizam HTML, XML e XHTML para o desenvolvimento [SILVA, 2008];

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. De acordo com [Dorado, 2005], JavaScript é implementado como parte do navegador permitindo melhorias nas interfaces do usuário e dar maior dinamismo nas páginas web;

3.2.3 HTML

É uma linguagem de marcação utilizada para o desenvolvimento de páginas web. Segundo [FLANAGAN; FERGUSON, 2002], HTML, CSS e JavaScript são os alicerces para a World Wide Web;

3.2.4 PHP

É uma linguagem de script feita para o desenvolvimento de páginas web, sendo executada do lado do servidor. Também é utilizada como linguagem de programação de propósito geral [GILMORE, 2011];

3.2.5 LARAVEL

É um framework escrito na linguagem PHP, que utiliza o padrão MVC e possui como principal característica o desenvolvimento de aplicações rápidas, performáticas e seguras[STAUFFER, 2016]; MySQL: É um sistema gerenciador de banco de dados relacional com código aberto, usado na maioria das aplicações gratuitas para gerir suas bases de dados[HEUSER, 2009].

3.2.6 PYTHON

Python é uma linguagem de programação de alto nível, interpretada de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. Foi lançada por Guido van Rossum em 1991

3.2.7 REACT

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

3.2.8 NEXT

Next.js é uma estrutura da web de desenvolvimento front-end React.

3.2.9 TF-IDF

Term frequency–inverse document frequency, que significa frequência do termo–inverso da frequência nos documentos, é uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos ou em um corpus linguístico. Ela é frequentemente utilizada como fator de ponderação na recuperação de informações e na mineração de dados. [RAJARAMANR, ULLMAN]

4 Cronograma

Etapa	**Descrição**	**MÊS**											
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	Elaboração da Proposta	x											
2	Pesquisa Bibliográfica	x											
3	Estudo de ferramentas na área		x										
4	Desenvolvimento dos algoritmos		x	x	x								
5	Desenvolvimento da aplicação web			x	x	x							
6	Execução de testes				x	x							
7	Redação da monografia					x	x	x	x	x	x	x	
8	Revisão e entrega oficial											x	
9	Preparação apresentação TCC												x

5 Resultados Esperados

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

Referências

ARMSTRONG, J. Plagiarism: what is it, whom does it offend, and how does one deal with it. *American Journal of Roentgenology, 1993;161: 479-484.*, 1991. Citado na página 7.

COADIC, Y.-F. L. *A Ciência da Informação. 2.ed. Brasília.* Dissertação (Mestrado) — Briquet de Lemos, 2004. Citado na página 7.

COOPER, W. A. *A Definition of Relevance for Information Retrieval.* Dissertação (Mestrado) — Information Storage and Retrieval, v.7, pp.19- 37, 1971. Citado na página 7.

H., C. et al. *Plagiarism Detector Using Machine Learning.* Dissertação (Mestrado) — International Journal of Research in Engineering, Science and Management, 2021. Citado na página 5.

M.K., B. *Information as thing.* Dissertação (Mestrado) — Journal of the American Society of Information Science, v.42, n.5, 1991. Citado na página 7.

MOOERS, C. *Zatocoding applied to mechanical organization of knowledge.* Dissertação (Mestrado) — American Documentation, v.2, n.1, p.20-32, 1951. Citado 2 vezes nas páginas 5 e 7.

P., F. et al. The ethical implications of plagiarism and ghostwriting in an open society. In: _____. [S.l.]: Walden University, LLC, Minneapolis, MN, 2017. cap. Volume9, p. 55–63. Citado na página 6.

T., S. *A review of and a framework for the thinking on the notion of information science.* Dissertação (Mestrado) — Journal of American Society for Information Science, v.26, n.6, p. 321-343, 1975. Citado na página 7.

T., S. *Plagiarism Detector Using Machine Learning Information Science.* Dissertação (Mestrado) — Journal of the American Society for Information Science, 1999. Citado na página 7.