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 veio a possibilidade de realizar diversas atividades de maneira e com isto alguns problemas se destacaram, dentre eles o plágio. Tendo isto em vista, tornou-se de suma importância com que tenha um tratamento adequado a tal ato. Com análises técnicas sobre dois ou mais arquivos, utilizando técnicas da recuperação da informação é análisada a probabilidade plágio entre eles. É pretendido que se tenha na conclusão deste projeto um sistema que retorne ao usuário a análise dentre seus arquivos.

Palavras-chave: Plágio, Pré-Projeto.

Abstract

With the insertion of technology in the academic sphere came the possibility of carrying out various activities in a manner and with this some problems stood out, among them plagiarism. In view of this, it has become of paramount importance that you have an adequate treatment for such an act. With technical analysis on two or more files, using information retrieval techniques, the likelihood of plagiarism between them is analyzed. It is intended to have at the conclusion of this project a system that returns to the user the analysis of its files.

Keywords: Plagiarism, Pre-Project.

Sumário

1 INTRODUÇÃO 5
1.1 Justificativa 5
1.2 Objetivos 6
1.2.1 Objetivo Geral 6
1.2.2 Objetivos Específicos 6

2 FUNDAMENTAÇÃO TEÓRICA 7

3 METODOLOGIA 9
3.1 Desenvolvimento 9
3.1.1 Estudo das ferramentas 9
3.1.2 Desenvolvimento do sistema 9
3.1.3 Estudo das características qualitativas 9
3.2 Materiais 9
3.2.1 CSS 9
3.2.2 JavaScript 10
3.2.3 HTML 10
3.2.4 PHP 10
3.2.5 LARAVEL 10
3.2.6 PYTHON 10
3.2.7 MySQL 10
3.2.8 TF-IDF 11

4 CRONOGRAMA 12

5 RESULTADOS ESPERADOS 13

REFERÊNCIAS 14

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. A inserção da tecnologia no âmbito acadêmico trouxe a possibilidade de realizar diversas atividades de maneira online, e com isto alguns problemas se descataram, dentre eles o plágio.

Sendo assim, quando se trata da comparação entre dois ou mais arquivos, temos muitas poucas alternativas no meio acadêmico para que os docentes possam utilizar 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 como TF-IDF e análisar 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 desenvoler 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 análisar 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 técnologias 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. O termo home office se tornou cada vez mais recorrente nos dias atuais, e com ele a entrega de atividades e trabalhos online. Os docentes recebem diversos arquivos para serem avaliados, posto isto é pertinente que se tenha alguma maneira de que se possa realziar um análise previa de tais trabalhos buscando encontrar a probabilidade de plágio entre eles. Neste contexto a justificativa deste trabalho é: gerar um sistema que auxilie qualquer usuário como um docente por exemplo, a avaliar a probabilidade de haver plágio em 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 comerciável

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver um sistema que realiza 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 análisar diversos arquivos
- Criar um sistema web que receba diversos arquivos e que trate eles para que sejam análisados

2 Fundamentação Teórica

(MOOERS, 1951) definiu a Recuperação de informação 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.

(COOPER, 1971) determinou o conceito de relevância é crucial na área de Recuperação de Informação, sendo muitas vezes utilizado na própria enunciação dos objetivos dessa área.

Em (T., 1975) foi abordado o conceito de relevância como uma medida da eficácia de um contato entre uma fonte e um destino em um processo de comunicação. E, uma vez que uma medida é uma relação, a relevância é também uma relação;

(M.K., 1991) definiu o termo documento que usamos na recuperação de informação como: um objeto que contém informação um documento. Assim, o termo informação poderia também designar "algo atribuído a um objeto, tal como dado e documento que se referem à informação, porque deles se espera que sejam informativos".

(VIEIRA, 1994) definiu Recuperação de informação como um processo de comunicação em que o emissor e o receptor se relacionam para cobrir uma necessidade de informação. Ao fazer uma pergunta ao sistema o homem funciona como emissor e o computador como receptor. Em contrapartida, o computador, ao apresentar a sua resposta passa a ser o emissor e o homem o receptor. Esta interação só é possível através do uso da linguagem.

Para (T., 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.

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).

Em (MEADOW et al., 2007) disse que a recuperação é um processo de comunicação, pelo qual autores e criadores de registros se comunicam com os leitores, mas indiretamente e possivelmente com um longo intervalo de tempo entre a criação de uma mensagem ou texto e a sua entrega para o usuário de um sistema de recuperação de informação. [...] As linguagens e os canais de tal sistema de comunicação são bastante diferentes de outros modelos bem conhecidos, tais como a radiodifusão ou a comunicação ponto-a-ponto.

Sendo assim Um sistema de recuperação de informação é um ambiente linguístico mediador da comunicação entre um estoque de informação e os seus requisitantes. Da qual esta comunicação pode ser a busca por uma informaçaõ.

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 sobre algumas ténicas 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 na linguagem de programação PHP, podendo sofrer alteração no decorrer no desenvolvimento. A princípio, foi definido essa linguagem por causa da facilidade que a linguagem apresenta e da vasta documentação presente no site da linguagem. Como apoio no desenvolvimento, será utilizado o framework Laravel 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 PHP, JavaScript, HTML, CSS e Laravel. 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 os arquivos e com o resultado de cada irá comparar com os demais e irá exibir o resultado ao usuário

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 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].

Para o desenvolvimento do algoritmo será feito o estudo e análise de algumas tecnicas de recuperação de informação, tais como:

3.2.8 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 realiza uma análise sobre a probabilidade de haver plágio entre diversos arquivos. E que com que o autor obtenha conhecimento mais aprofundado sobre tais temas e contextos adequados para aplicar cada um dos algoritmos.

Referências

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.

MEADOW, C. et al. *Text Information Retrieval System. 3rded.* Dissertação (Mestrado) — London UK:Elsevier, 2007. Citado na página 7.

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 5.

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.

VIEIRA, S. *La recuperación automática de información jurídica:metodologia de análises lógico-sintáctico para la lengua portuguesa*. Dissertação (Mestrado) — Universidad Complutense de Madrid, 1994. Citado na página 7.