

Instituto Federal de Minas Gerais
Ciência da Computação



Projeto de graduação apresentado para a obtenção do
Diploma Nacional de Cientista da Computação

Arildo Magno de Macedo

Protótipo de aplicação para análise de plágio
entre múltiplos arquivos

Abstract

The present work is part of an undergraduate project, which aimed to develop a prototype to carry out the analysis of plagiarism between several files.

Resumo

O presente trabalho é parte de um projeto de graduação, que teve objetivo de desenvolver um protótipo para realizar a análise de plágio entre diversos arquivos.

Contents

1	Introdução	1
2	Justificativa	2
3	Objetivos	3
3.1	Objetivo Geral	3
3.2	Objetivo Especifico	3
3.2.1	Utilizar técnicas de recuperação de informação para analisar diversos arquivos	3
3.2.2	Criar um sistema web que receba diversos arquivos e que trate eles para que sejam analisados	3
4	Fundamentação Teórica	4
4.0.1	CSS	4
4.0.2	JavaScript	4
4.0.3	HTML	4
4.0.4	PHP	4
4.0.5	LARAVEL	4
4.0.6	PYTHON	5
4.0.7	MySQL	5
4.0.8	TF-IDF	5
5	Metodologia	6
5.0.1	Estudo das ferramentas:	6
5.0.2	Desenvolvimento do sistema:	6
5.0.3	Estudo das características qualitativas:	6
6	Cronograma	7
7	Resultados Esperados	8
8	Referências	9

Chapter 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 a evolução das tecnologias e sua integração em toda parte. A inserção da tecnologia no âmbito acadêmico trouxe a possibilidade de realizar diversas atividades de maneira online, e com isto alguns problemas se descatarem, dentre eles o plágio. Sendo assim tornou-se de suma importância que se tenha um sistema para que realize uma análise da similaridade entre arquivos.

Neste contexto, o propósito deste trabalho de conclusão de curso (TCC) é utilizar técnicas de recuperação de informação como TF-IDF e analisar a similaridade entre os arquivos. E então via uma plataforma desenvolvida para a web fornecer ao usuário a possibilidade de enviar seus arquivos e receber obtenha a similaridade entre eles.

O desenvolvimento de uma plataforma web facilitará a visualização das informações pelo usuário. O usuário do sistema terá uma visão mais clara e objetiva da relação entre os arquivos. 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.

Chapter 2

Justificativa

O plágio sempre esteve presente na sociedade, porém com o advento da tecnologia tal prática tornou-se mais simples e com isto, mais comum. 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 realizar 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.

Chapter 3

Objetivos

3.1 Objetivo Geral

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

3.2 Objetivo Especifico

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

Chapter 4

Fundamentação Teórica

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

4.0.1 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];

4.0.2 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;

4.0.3 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;

4.0.4 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];

4.0.5 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].

4.0.6 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

4.0.7 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 técnicas de recuperação de informação, tais como:

4.0.8 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]

Chapter 5

Metodologia

5.0.1 Estudo das ferramentas:

Para dar início no desenvolvimento do trabalho, será necessário realizar um estudo mais aprofundado sobre algumas técnicas de recuperação de informação, como tf-idf. As ferramentas de tf-idf já foram apresentadas na disciplina de Recuperação de Informação.

5.0.2 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.

5.0.3 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.

Chapter 6

Cronograma

Etapa	Descrição	MÊS							
		1	2	3	4	5	6	7	8
1	Elaboração da Proposta	x							
2	Pesquisa Bibliográfica sobre trabalhos relacionados		x						
3	Estudo de ferramentas na área			x					
4	Desenvolvimento dos algoritmos de recuperação de informação				x				
5	Desenvolvimento da aplicação web					x	x		
6	Execução de testes no sistema já implementado							x	
7	Redação da monografia							x	
8	Revisão e entrega oficial								x
9	Preparação para a defesa e apresentação pública do TCC								x

Chapter 7

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.

Chapter 8

Referências

[**FOLTÁINEK, MEUSCHKE, GIPP**] Academic Plagiarism Detection: A Systematic Literature Review, Department of Informatics, Mendel University in Brno, Czechia and University of Wuppertal Germany.

[**MALIK, BILAL, ILYAS, RAZZAQ, MAQBOOL, ABBAS**] Plagiarism Detection Using Natural Language Processing Techniques Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan.

[**CHAVAN, TAUFIK, KADAVE, CHANDRA**] Plagiarism Detector Using Machine Learning International Journal of Research in Engineering, Science and Management Volume 4, Issue 4, April 2021.

[**BUCKLAND**] Information as thing. Journal of the American Society of Information Science, v.42, n.5, 1991. p.351-360.

[**COOPER**] A Definition of Relevance for Information Retrieval. Information Storage and Retrieval, v.7, pp.19- 37, 1971.

[**MEADOW, C.T.; BOYCE, B.R.; KRAFT, D.H.; BARRY**] ext Information Retrieval System. 3rded. London UK: Elsevier, 2007.

[**MOOERS**] Zatocoding applied to mechanical organization of knowledge. American Documentation, v.2, n.1, 1951, p.20-32.

[**SARACEVIC**] T. Information Science. Journal of the American Society for Information Science, v.50, n.12, 1999.

[**VIEIRA**] La recuperación automática de información jurídica: metodología de análisis lógico-sintáctico para la lengua portuguesa.1994.