UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

Vinícius Leal Trindade

ANÁLISE E OTIMIZAÇÃO DE DESEMPENHO DA FERRAMENTA MISS MARPLE: DETECTOR DE PLÁGIO

Santa Maria, RS, Brasil

2015

Vinícius Leal Trindade

ANÁLISE E OTIMIZAÇÃO DE DESEMPENHO DA FERRAMENTA MISS MARPLE: DETECTOR DE PLÁGIO

Trabalho de Conclusão apresentado ao Curso de Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS) como requisito parcial para a obtenção do grau de **Bacharel em Ciência da Computação.**

Orientadora: Profa. Dra. Roseclea Duarte Medina

Trabalho de Graduação Nº 401
Santa Maria, RS, Brasil

2015

Vinícius Leal Trindade

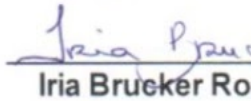
ANÁLISE E OTIMIZAÇÃO DE DESEMPENHO DA FERRAMENTA MISS MARPLE: DETECTOR DE PLÁGIO

Trabalho de Conclusão apresentado ao Curso de Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS) como requisito parcial para a obtenção do grau de **Bacharel em Ciência da Computação.**

Aprovado em 09 de dezembro de 2015:

Roseclea Duarte Medina, Dra. (UFSM)
(Presidente/Orientadora)

Giliane Bernardi, Dra. (UFSM)



Iria Brucker Roggia, Dra. (UFSM)

Santa Maria, RS
2015

DEDICATÓRIA

À Gabriela Gabbi

AGRADECIMENTOS

– A todos os meus familiares, sejam os que ainda se encontram presentes – em especial à minha mãe – sejam os que já se foram, por todo o apoio, material e emocional, sem o qual eu não teria chegado aonde cheguei.

– À minha orientadora Roseclea Duarte Medina e aos membros da banca, Profas. Giliane Bernardi e Iria Brucker Roggia, que me guiaram por todo esse tempo na construção deste trabalho e o tornaram possível.

– À Universidade Federal de Santa Maria por ter me acolhido por todos esses anos e ter me proporcionado a oportunidade que tive de crescer pessoal e profissionalmente.

– Aos colegas com os quais convivi, que se tornaram meus amigos e que me ajudaram direta ou indiretamente nesta jornada.

“O sucesso é simples. Primeiro, você decide com precisão o que quer; segundo, decide que pagará o preço para fazê-lo acontecer; então, paga esse preço.”

(Bunker Hunts)

RESUMO

ANÁLISE E OTIMIZAÇÃO DE DESEMPENHO DA FERRAMENTA MISS MARPLE: DETECTOR DE PLÁGIO

AUTOR: VINÍCIUS LEAL TRINDADE
ORIENTADORA: ROSECLEA DUARTE MEDINA

A detecção automática de plágio em trabalhos em formato digital como área de pesquisa encontra-se em desenvolvimento. O que se tem hoje em dia em termos de métodos e estudos que tratam desse assunto são recentes e em pequena quantidade em comparação com outras áreas da ciência da computação, ao passo que é cada vez maior a facilidade da prática do plágio em trabalhos acadêmicos devido à crescente quantidade de informação disponível na *internet* a cada ano. Dentro deste cenário, encontra-se o *Miss Marple* – Ferramenta de Detecção de Indícios de Plágio com Base no Método DIP – Detector de Indícios de Plágio. Tal ferramenta realiza detecção de indícios de plágio em textos em português através de buscas online de documentos suspeitos e posteriormente comparados em termos de similaridade de conteúdo. Porém, ao utilizá-la, a mesma apresentou dificuldades de desempenho durante sua execução, pois consumia tempos de processamento insatisfatórios comparados com outras ferramentas do gênero. Diante desta situação e após a realização de estudos bibliográficos na área de detecção de plágio, constatou-se que seria possível reestruturar seu código e otimizar seu desempenho, além de implementar a funcionalidade de detecção de plágio bilíngue, a qual a ferramenta ainda não dispunha. Após realizadas as devidas alterações, foram feitos testes de execução a partir de um conjunto de documentos selecionados a fim de validar a nova versão da ferramenta. Os resultados dos testes comprovaram que as alterações realizadas melhoraram o desempenho da ferramenta em relação à versão anterior, bem como a tornou mais funcional com o suporte à detecção de plágio bilíngue. O presente trabalho mostra como se deu o processo de desenvolvimento da nova versão da ferramenta desde a identificação dos problemas, o estudo bibliográfico, descrição das alterações realizadas, os testes e seus resultados.

Palavras-chave: Detecção de plágio. Análise de Documentos. Tradução de Máquina.

ABSTRACT

ANALYSIS AND OPTIMIZATION OF *MISS MARPLE* TOOL: PLAGIARISM DETECTOR

AUTHOR: VINÍCIUS LEAL TRINDADE
ADVISOR: ROSECLEA DUARTE MEDINA

The research area of automatic plagiarism detection of digital papers is still in development. The methods and studies about this subject these days are new and small amount compared to other areas of computer science, while the plagiarism practice in academic papers is more and more easy due the increasing amount of information available on the internet every year. Within this scenario is Miss Marple - Plagiarism Evidence Detection Tool Based on PED Method - Plagiarism Evidence Detector. Such tool performs detection of plagiarism evidence in portuguese texts through web searches for suspect documents and then compared in terms of content similarity. However, it showed poor performance during its execution due to a unsatisfactorily long processing time compared to other similar tools. Before this situation and after conducting bibliographical studies in plagiarism detection area, it was found that it would be possible to restructure its code and optimize its performance, as well as implement a bilingual plagiarism detection feature, which it did not have yet. After the necessary changes were, a set of runtime tests was performed with selected documents in order to validate the updated version of the tool. The test results proved that the changes that have been made improved the performance of the tool compared to its prior version as well as introduced the bilingual plagiarism detection support. This work shows how the development process of the new version was conducted, since the identification of the initial issues, the bibliographic study, description of the changes that were made, the tests and their results.

Keywords: Plagiarism Detection. Document Analysis. Machine Translation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	Application Program Interface
DIP	Detector de Indícios de Plágio
UFSM	Universidade Federal de Santa Maria
HTTP	Hypertext Transfer Protocol
PDF	Portable Document Format
HTML	Hypertext Markup language
RAM	Random Access Memory
TB	Terabyte
IDE	Integrated Development
MB	Megabyte
JDK	Java Development Kit

SUMÁRIO

INTRODUÇÃO

O uso da *internet* como fonte de pesquisa é algo que vem crescendo substancialmente ao longo do tempo. Informações que antigamente só eram acessíveis por meio de uma longa busca em livros e materiais impressos diversos e em lugares específicos, hoje em dia são acessadas em poucos segundos através de uma simples busca pela *internet*. Dentro da área acadêmica, isso gerou duas realidades distintas: De um lado, a produção de trabalhos acadêmicos se tornou muito mais prática e produtiva. De outro, uma crescente quantidade de trabalhos plagiados vem sendo constatada nos últimos anos e isso se torna um problema nas universidades, uma vez que o acesso irrestrito e anônimo à grande quantidade de materiais que a *internet* disponibiliza tornou a prática do plágio muito mais fácil e de difícil detecção.

Dada esta realidade, fica evidente a importância do desenvolvimento de ferramentas que forneçam soluções automatizadas de detecção de plágio, visto que o constante aumento do volume de informações de livre acesso na *internet* tende a estimular também o aumento da prática do plágio e, ao mesmo tempo, dificultar a sua detecção de forma manual, pois se torna inviável analisar manualmente uma quantidade tão grande de material como a encontrada na internet. Do ponto de vista do plágio bilíngue, a evolução e popularização de ferramentas online de tradução automática, como o *google tradutor,* também contribui fortemente com este cenário. Hoje em dia é possível traduzir textos com ótima qualidade de praticamente qualquer idioma do planeta em questão de segundos. Portanto, o desenvolvimento de ferramentas que sejam capazes de detectar essas práticas se tornou algo fundamental nos dias de hoje.

Dentro desse contexto, encontra-se a ferramenta *Miss Marple* – Ferramenta de Detecção de Indícios de Plágio com base no Método DIP – Detector de Indícios de Plágio (ARENHART, 2013) desenvolvida no projeto de mestrado do Programa de Pós-Graduação em Informática da Universidade Federal de Santa Maria. O software em questão realiza a análise e detecção de plágio, tanto do tipo direto quanto do tipo

plágio mosaico, em textos em português utilizando a internet como fonte de pesquisa. Durante sua utilização, constatou-se que o *Miss Marple* ocupa um tempo significativo de desempenho, fazendo com que a espera entre a submissão do arquivo a ser analisado e o resultado da análise seja muito elevado. O objetivo do presente trabalho é reduzir os problemas de desempenho encontrados através de uma reorganização de seu código fonte, baseado em técnicas e trabalhos correlatos desenvolvidos nesta área, bem como adicionar um módulo que realize busca de textos em inglês, funcionalidade que não se encontra no *Miss Marple* atualmente.

O trabalho se inicia, no primeiro capítulo, apresentando a fundamentação teórica sobre a qual o trabalho foi construído, feita por meio de um estudo bibliográfico de trabalhos que tratam da área de detecção de plágio em termos de conceituação, pesquisa e técnicas de detecção, bem como trabalhos correlatos ao *Miss Marple*, sob os quais este se baseou quando do seu desenvolvimento.

O segundo capítulo apresenta a metodologia usada no desenvolvimento deste trabalho. São apresentados os motivos que deram origem ao trabalho, as ferramentas utilizadas, o ambiente de desenvolvimento e como se deu o processo de testes e de validação da nova implementação.

O terceiro capítulo descreve como se deu o processo de implementação da nova versão da ferramenta. Tem-se primeiramente uma análise da arquitetura original do Miss Marple, identificando quais problemas ela apresenta, e logo após são descritas as medidas tomadas para corrigí-los e como se deu a implementação da detecção de plágio bilíngue.

O quarto capítulo apresenta a análise dos resultados obtidos a partir dos testes realizados, bem como a comparação de desempenho entre a versão original e a nova versão do programa, a fim de validar a proposta deste trabalho.

No quinto e último capítulo, tem-se as considerações finais sobre todo o processo de desenvolvimento do trabalho. São considerados as melhorias que a nova implementação trouxe diante dos objetivos traçados, as dificuldades encontradas ao longo do trabalho, sugestões para trabalhos futuros a partir de questões identificadas durante o desenvolvimento e quais foram as contribuições que este trabalho trouxe à área de detecção de plágio de uma forma geral.

1 ESTUDO BIBLIOGRÁFICO

Este capítulo apresenta um estudo bibliográfico de trabalhos realizados nos últimos anos que tratam de conceituação, pesquisa e técnicas de detecção de plágio, bem como trabalhos correlatos ao *Miss Marple* que serviram como fundamentação teórica sobre a qual o *Miss Marple* foi desenvolvido e também serão a fundamentação do trabalho de implementação realizado pelo presente trabalho.

1.1 CONCEITUAÇÃO DE PLÁGIO

De uma forma geral, o plágio pode ser definido como sendo a divulgação de trabalhos ou de idéias de alguém como se fossem de sua própria autoria. A forma como isso é feito varia conforme o contexto e o tipo de publicação e utiliza diferentes estratégias. Segundo Kirkpatrick (2006), a prática do plágio pode assumir as seguintes classificações:

a) Plágio direto**:** Neste tipo de plágio, o conteúdo do texto é copiado palavra por palavra exatamente como foi originalmente escrito sem qualquer tipo de referência ao seu autor. Ao comparar o texto plagiado com o texto original, sua identificação é imediata, visto que se trata de uma simples cópia exata do texto original, sem nenhuma modificação;
b) Plágio mosaico: Ao contrário do plágio direto, o autor do plágio mosaico utiliza de artifícios como troca de palavras por sinônimos, reformulação de frases e parágrafos e mudança da estrutura geral do texto original para tentar mascarar a cópia do texto e fazê-lo se passar por um texto de sua autoria. É o tipo mais comum de plágio e o que requer técnicas mais sofisticadas de detecção em relação ao plágio direto, pois o sistema de detecção precisa identificar essas mudanças introduzidas no texto original para determinar a existência ou não do plágio;
c) Plágio bilíngue: Semelhante ao plágio direto, mas ao invés de se tratar de uma cópia exata, trata-se da tradução de um texto original escrito em outra

língua e apresentado como sendo de autoria do autor do plágio. Este método ainda pode ser combinado com o plágio mosaico, com o objetivo de dificultar a detecção do plágio. Um sistema detector de plágio bilíngue deve ser capaz de retraduzir o texto a um idioma comum (geralmente inglês) e compará-lo ao texto original da mesma forma que faz com o plágio direto, ou ainda aplicar técnicas de detecção de plágio mosaico;

1.2 METODOLOGIAS ANTIPLÁGIO

Diversas metodologias antiplágio foram sendo desenvolvidas e aprimoradas ao longo do tempo com base no tipo de plágio que deve ser analisado e os objetivos a que o sistema de detecção se propõe. No contexto do plágio em documentos de texto, existem dois pontos centrais que definem o tipo de análise de plágio que o sistema realiza: Um diz respeito ao critério adotado para determinar a existência ou não de indícios de plágio no documento a ser analisado. O outro diz respeito ao tipo de processamento ao qual o texto do documento suspeito será submetido para se chegar ao resultado desejado. Em relação ao critério de determinação de plágio, temos as seguintes categorias:

a) Extra corpus: Consiste em avaliar um documento suspeito por meio da comparação de seu conteúdo com o de um conjunto de documentos “externos”, considerados legítimos. Para cada documento do conjunto, a existência ou não de indícios de plágio no documento suspeito é determinada a partir do grau de similaridade entre ambos, calculado através de algum cálculo de similaridade. Esta estratégia é a mais comumente utilizada em sistemas de detecção de plágio hoje em dia (CARNAHAN *et al*, 2014; NASEEN; KURIAN, 2013).
b) Intra corpus: Na análise intra corpus, o texto suspeito é analisado em relação ao seu próprio conteúdo e seu grau de plágio é determinado em função de mudanças no “estilo de escrita” empregado pelo autor ao longo do texto, isto é, o sistema é capaz de identificar diversas características no modo de

escrever do autor do texto e sabe determinar quando esse estilo muda bruscamente a partir de algum trecho. Essas mudanças são quantificadas através de métodos estatísticos e usadas para determinar a probabilidade de indícios de plágio no texto analisado (CARNAHAN *et al*, 2014; POTTHAST *et al*, 2011).

No que diz respeito ao tratamento do conteúdo do texto, isto é, a forma com que o texto é processado pelo sistema para se determinar a existência de indícios de plágio, temos as seguintes metodologias:

a) Comparação de *Fingerprints*: São sentenças de texto convertidos em uma representação numérica gerada através de algoritmos de *hashing.* Assim*,* Cada sentença é representada por um *fingerprint* os termos que compõem a sentenças são chamados de *minutiae.* Durante o processo de comparação do texto suspeito com o conjunto de documentos externos em análises extra corpus, cada sentença do documento suspeito é convertido em *fingerprints* e seus *minutiae* são comparados com os dos documentos externos. O índice de similaridade entre ambos os documentos se dá em função da similaridade dos *minutiae* dos documentos (CARNAHAN *et al*, 2014; HOAD; ZOBEL, 2003);
b) Comparação de *Strings*: Este método segue os mesmos moldes da comparação de *fingerprints*, mas usa a representação textual pura para mensurar a similaridade dos documentos analisados. Para isso, elementos de cada palavra da sentença, como prefixos e sufixos devem ser processados e categorizados em estruturas de dados para serem posteriormente comparados. Devido a esse *overhead* de processamento prévio, é um método computacionalmente caro e pouco usado. (ZHAN *et al*, 2008);
c) Modelo Vetorial: No modelo vetorial, o sistema de detecção, analogamente ao modelo de *fingerprints* trata o texto, ou as sentenças do texto do documento suspeito como vetores de coeficientes em um espaço vetorial. Para tal, o sistema converte cada sentença em um vetor e compara através de um

cálculo de similaridade, (geralmente o cálculo cosseno) e gera um índice de similaridade que varia de 0% a 100%, ou 0 a 1. um índice de 0% indica textos completamente diferentes, enquanto 100% indica documentos idênticos (HOAD; ZOBEL, 2003; HUANG, 2008);

Essas técnicas descrevem como o sistema de detecção faz a análise central do texto do documento suspeito de plágio. De todas as implementações encontradas na literatura, as técnicas citadas são as mais utilizadas (AREFIN; MORIMOTO; SHARIF, 2013; KENT; SALIM, 2009; PEREIRA, 2010; PERTILE, 2011), sendo a metodologia extra corpus a mais popular. Dos métodos de análise de similaridade, a escolha varia conforme o contexto em que o sistema está inserido. O *Miss Marple* adota a abordagem *extra corpus* e a coleção de documentos externos encontra-se na *internet*, portanto o sistema usa a API *google search* para realizar as buscas de documentos similares. Já no processo de análise de similaridade, o texto é processado por um modelo vetorial e seu índice de similaridade é gerado pelo cálculo cosseno (ARENHARDT, 2013).

1.3 METODOLOGIAS ANTIPLÁGIO BILÍNGUE

Quando se fala em plágio bilíngue, fala-se necessariamente em técnicas de tradução de documentos. As metodologias de detecção de plágio bilíngue nada mudam no que diz respeito às técnicas de análise e processamento de texto do documento. O que as diferencia das metodologias tradicionais é a tradução prévia do texto a uma linguagem de referência antes de se buscar o plágio propriamente dito. Nesse sentido, o referencial teórico estudado aborda metodologias de tradução de textos eletrônicos que servem de fundamentação para a implementação do presente trabalho. Existem duas metodologias gerais adotadas em ferramentas de tradução de texto:

a) Tradução baseada em regras: A metodologia de tradução baseada em regras consiste em traduzir um texto de uma linguagem para outra através do mapeamento das regras gramaticais correspondentes entre essas

linguagens. Isso se dá por meio de uma base de dados que armazena a tradução direta das palavras de cada linguagem e suas propriedades morfológicas, sintáticas, semânticas (COSTA-JUSSÀ, 2012). A cada sentença do texto de origem, o sistema recupera as informações necessárias da base de dados e as mapeia para a linguagem de destino a partir de suas regras gramaticais. O problema dessa abordagem se dá pela quantidade de informações que precisam ser armazenadas para se implementar uma base de dados suficientemente completa entre dois idiomas para que a qualidade da tradução seja aceitável. Um sistema que se disponha a ter suporte a uma grande quantidade de idiomas pode se tornar inviável em função do custo de armazenamento que isso demandaria;

b) Tradução baseada em modelos estatísticos: Modelos estatísticos de tradução, ao invés de se basear em regras estáticas pré-definidas, realizam traduções dinamicamente por meio de máquinas de aprendizado. Dada uma sentença em uma linguagem de origem, o sistema de tradução determina qual a melhor tradução possível para a sentença através de uma análise estatística aplicada em um conjunto de possíveis traduções (LOPEZ, 2010). Essas traduções são inseridas no sistema manualmente e, a cada sentença traduzida, o sistema grava a opção escolhida e com isso “aprende” a traduzir outras sentenças que sejam semelhantes a esta. Isto faz com que o sistema se aperfeiçoe continuamente e faz melhor uso de sua base de informações. Este é o sistema utilizado atualmente por tradutores *web* como *google translate e bing translate* (KARAMI, 2010);

1.4 TRABALHOS CORRELATOS

Esta seção apresenta um resumo de três trabalhos correlatos que tratam da detecção de plágio e usam diferentes metodologias entre as citadas nos tópicos anteriores. Os dois primeiros são métodos de detecção e o último trata-se de um software de detecção online. Logo após, tem-se uma análise das vantagens e

desvantagens de cada um, como o presente trabalho e o próprio *Miss Marple* original se diferencia desses trabalhos.

1.4.1 Método DIP

O *Miss Marple* tem como base principal o método *DIP – Detector de Indícios de Plágio,* desenvolvido e publicado por (PERTILE, 2011) como dissertação de mestrado para o Programa de Pós-Graduação da Universidade Federal de Santa Maria. Ele é um método de detecção de plágio que adota a metodologia *extra corpus e* usa a *internet* como sendo a coleção de documentos de referência. O processo de buscas consiste em separar cada parágrafo do documento suspeito e pesquisá-los individualmente no repositório *web* do *google* através da *Google Web Search API*.

Ao final das pesquisas, os parágrafos são comparados com os *snippets* dos resultados encontrados, isto é, os pequenos trechos de texto retornados em cada resultado do *google*. O sistema então gera índices de similaridade para cada comparação, que são adicionados a um relatório juntamente com os endereços dos documentos correspondentes. Assim como o *Miss Marple*, não tem suporte a plágio bilíngue.

1.4.2 Cross-Language Plagiarism Detection

Desenvolvido por (PEREIRA, 2010) como dissertação de mestrado para o programa de Pós-Graduação em Computação da Universidade Federal do Rio Grande do Sul, este trabalho apresenta um sistema capaz de detectar plágio em documentos escritos em qualquer idioma. Para isso, o sistema conta com um método de detecção de idiomas integrado, a fim de identificar o idioma do texto submetido de forma automática e transparente, e um tradutor automático, responsável pela tradução do texto antes de realizar a análise de plágio.

Assim como o método DIP, adota a metodologia *extra corpus*, mas ao invés de buscar os documentos externos na *internet*, usa uma coleção de documentos de

referência local. Além disso, a seleção dos trechos do texto suspeito que serão enviados para a análise se dá por meio de um modelo de classificação capaz de diferenciar uma sentença plagiada de uma sentença original. Esta é uma abordagem tipicamente utilizada em análises intra corpus, mas adaptada a um modelo extra corpus. Assim como o *Miss Marple*, também realiza *stemming* e remoção de *stopwords* do texto antes de enviá-lo para a análise.

Após a extração dos trechos selecionados, o sistema realiza as pesquisas na base local e determina a similaridade entre os documentos através do cálculo cosseno. Ao final do processamento, os resultados são reunidos e exibidos ao usuário como uma única passagem de texto.

1.4.3 Turnitin

Turnitin.com é um sistema online e pago de detecção de plágio com foco em trabalhos escolares. O sistema recebe um documento do usuário e usa o método de *fingerprints* para realizar buscas tanto na *internet* quanto em repositórios locais. Os repositórios locais são compostos pelos próprios documentos enviados pelos usuários, o que aumenta a precisão dos resultados ao longo do tempo. Ao final do processamento o sistema emite um relatório informando o nível de originalidade do documento submetido. Entre os sistemas pagos, é considerado o que possui a maior taxa de detecção de plágio. (KENT; SALIM, 2009; PERTILE, 2011).

Dados estes trabalhos, percebe-se algumas características que os diferem entre si. O método DIP se destaca por utilizar a *internet* como *corpus*, mas não possui repositório local e o tratamento do texto não é tão eficiente quanto o do *Miss Marple*, além de não realizar busca bilíngue. Por outro lado, o método descrito no trabalho Cross-Language Plagiarism Detection tem essa funcionalidade como característica principal, tratando inclusive o plágio em múltiplos idiomas. Em compensação, o método utiliza apenas uma base local de pesquisa, sendo esta menos abrangente do que a *internet*. Já o *turnitin*, dispõe tanto da *internet* quanto de

bases locais para pesquisa, mas é uma sistema pago, o que restringe bastante seu uso.

O *Miss Marple* reúne características dos três sistemas, pois faz uso tanto da *internet* quanto de base local para pesquisa, além de realizar *stemming* e remoção de *stopwords*, características que também são encontras no trabalho Cross-Language Plagiarism. Mas ao contrário deste, não detecta plágio bilíngue. O presente trabalho, por sua vez, busca preencher mais esta lacuna ao incluir o plágio bilíngue entre suas funcionalidades, bem como tornar o sistema como um todo mais eficiente, visto os problemas identificados inicialmente.

2 METODOLOGIA

Este capítulo apresenta a metodologia usada neste trabalho para alcançar os objetivos desejados com base no estudo bibliográfico exposto até aqui e na estrutura sobre a qual o Miss Marple foi construído.

2.1 DEFINIÇÃO DO TRABALHO

O ponto de partida para a realização deste trabalho se deu a partir de testes de execução do Miss Marple realizados em um ambiente computacional. A partir desses testes, constatou-se que o tempo de processamento que o sistema consome na execução da análise de um único arquivo fica muito acima do que seria aceitável. Enquanto nas ferramentas estudadas, como a implementação do método DIP e o *turnitin*, o tempo de análise de um documento dificilmente passa dos 20 minutos, O Miss Marple facilmente consome o dobro desse tempo ou mesmo chega a ultrapassar mais de 1 hora para a mesma análise até que o todos os arquivos sejam baixados e analisados. Esta e uma diferença considerável de performance, ainda mais se tivermos vários documentos suspeitos a serem testados em uma sessão de análise.

Com o objetivo de identificar quais eram as causas desse problema e de que forma isso poderia ser corrigido, foi feito um estudo bibliográfico que analisou tanto a fundamentação teórica do Miss Marple e do método DIP, no qual ele se baseia, quanto os métodos e trabalhos correlatos da área de detecção de plágio, descritos no capítulo 1. Desse estudo, também surgiu a idéia de adicionar ao *Miss Marple* a funcionalidade de detecção de plágio bilíngue.

Após o levantamento desses parâmetros, definiu-se um projeto de aperfeiçoamento do Miss Marple através de uma reimplementação de algumas estruturas do Miss Marple responsáveis pelos problemas de tempo de execução que foram observados e na adição da funcionalidade de detecção de plágio bilíngue. Tal

implementação reúne pontos positivos tanto dele quanto do método DIP a fim de contornar os problemas identificados inicialmente.

2.2 MODELAGEM DO PROGRAMA

A fim de mostrar as diferenças entre as duas versões do programa e ilustrar como se dá o funcionamento da nova implementação, foi construído um modelo gráfico de implementação do trabalho que mostra seu fluxo de execução desde a submissão do documento suspeito por parte do usuário até o relatório final com resultados da análise de plágio. Tais modelos são mostrados no capítulo 3 deste trabalho.

2.3 AMBIENTE DE DESENVOLVIMENTO E FERRAMENTAS UTILIZADAS

Como se trata de uma contribuição sobre um projeto já estabelecido, tanto a linguagem de programação quanto as bibliotecas usadas na implementação original foram mantidas, sendo acrescentado apenas o que foi usado na implementação da detecção bilíngue. O Miss Marple foi programado na linguagem JAVA e utiliza as seguintes bibliotecas:

a) *Apache Lucene*: API de busca e indexação de documentos, realiza no Miss Marple a tarefa o *stemming* do texto submetido pelo usuário;
b) *Apache HTTP Core*: Trabalha na conversão dos resultados retornados pelo *Google Search* em páginas HTML;
c) *Apache PDF Box*: Biblioteca usada para a criação e tratamento de documentos em PDF;
d) *DOCX4J*: Biblioteca usada na manipulação de textos com extensão docx;
e) *POI*: Biblioteca de manipulação de arquivos do MS Word;
f) *Google Web Search*: API do *google* usado no processo de buscas de documentos na *Web*;
g) *Microsoft Translator*: API usada na tradução de textos;

O ambiente de desenvolvimento em termos de hardware e software é constituído da seguinte forma:

a) Computador Desktop com processador Intel Core i5-2500K de 3.3GHz;
b) 8GB de memória RAM e 1TB de disco rígido;
c) Sistema Operacional Windows 7 64 bits;
d) IDE NetBeans com Java 8 JDK 1.8;
e) *Internet* banda larga de 15MB de velocidade;

2.4 METODOLOGIA DE TESTES E VALIDAÇÃO

Ao final da implementação, foram realizados testes a fim de validar a solução proposta e sua integração no software *Miss Marple*. Os testes consistiram de execuções da versão antiga e da versão otimizada a partir de um conjunto de documentos a fim de comparar o desempenho de ambas as versões e analisar a precisão da resposta do programa em relação aos seguintes parâmetros:

a) Tempo de execução do programa: Verificar se a nova versão efetivamente tem melhor desempenho que a versão antiga;
b) Quantidade de arquivos baixados: Verificar a diferença na quantidade de arquivos baixados entre as duas versões;

Após a conclusão dos testes, o programa foi avaliado em relação aos objetivos do trabalho e às contribuições trazidas por esta implementação ao software *Miss Marple*.

3 IMPLEMENTAÇÃO

Este capítulo descreve como a implementação do trabalho foi feita a fim de se atingir os objetivos traçados no capítulo anterior. Para isso, foi feita uma análise da arquitetura do Miss Marple, mostrando as fases de processamento pelas quais o documento suspeito de plágio passou desde sua submissão até o resultado da análise de plágio. Logo após são descritos quais são os fatores que causam os problemas apresentados na sua execução, o que foi feito para corrigí-los e como a detecção de plágio bilíngue foi implementada.

3.1 ARQUITETURA *MISS MARPLE*

A execução do Miss Marple desde a submissão do documento suspeito até o resultado da análise está dividida em 4 passos, mostradas nos tópicos seguintes.

3.1.1 Pré-processamento

A fase de pré-processamento é a primeira fase pela qual o documento submetido passa e realiza uma série de transformações no texto com o objetivo de remover palavras desnecessárias, reduzir o texto a um formato que seja mais adequado ao mecanismo de buscas na web e assim aumentar a quantidade de documentos candidatos retornados. Tais transformações são:

a) Remoção de *stopwords: Stopwords* são palavras que não possuem relevância à estrutura de em um texto, pois não carregam informações que definem o seu conteúdo. Podemos citar como sendo *stopwords*: advérbios, artigos, pronomes, conjunções e preposições. Ao removê-las, teremos um texto mais conciso e buscas mais precisas. Nesta fase também são eliminados caracteres especiais;

b) *Stemming: Stemming* é um processo de redução das palavras ao seu radical, facilitando o processo de busca. Isso aumenta a relevância dos resultados encontrados pois palavras que podem ter sido alteradas pelo autor do plágio por outras que possuem o mesmo radical não seriam encontradas sem esse processo;
c) *Tokenização:* Após passar pelos processos de remoção de *stopwords* e *stemming*, o processo de *tokenização* trata de dividir o texto em diversos trechos, cada um com um número de termos, ou *tokens* (palavras), a fim de enviar ao motor de buscas. Esse processo tem como objetivo otimizar o número de resultados relevantes encontrados em função da quantidade de termos que compõem o trecho. Segundo (PERTILE, 2011), um trecho ótimo contém entre 20 e 50 termos;

3.1.2 Recuperação dos documentos candidatos

Após a fase de pré-processamento, cada trecho extraído do texto é pesquisado na internet a fim de encontrar documentos candidatos a serem fonte de plágio. O Miss Marple executa as buscas através da API *google search e* faz o download de cada arquivo encontrado, formando um repositório local.

3.1.3 Análise de Plágio

Após as buscas, cada arquivo baixado é comparado com o arquivo submetido pelo usuário e então recebe um índice de similaridade que indica a probabilidade de ter sido usado como fonte de plágio no trabalho analisado. Esse índice é gerado através do cálculo cosseno.

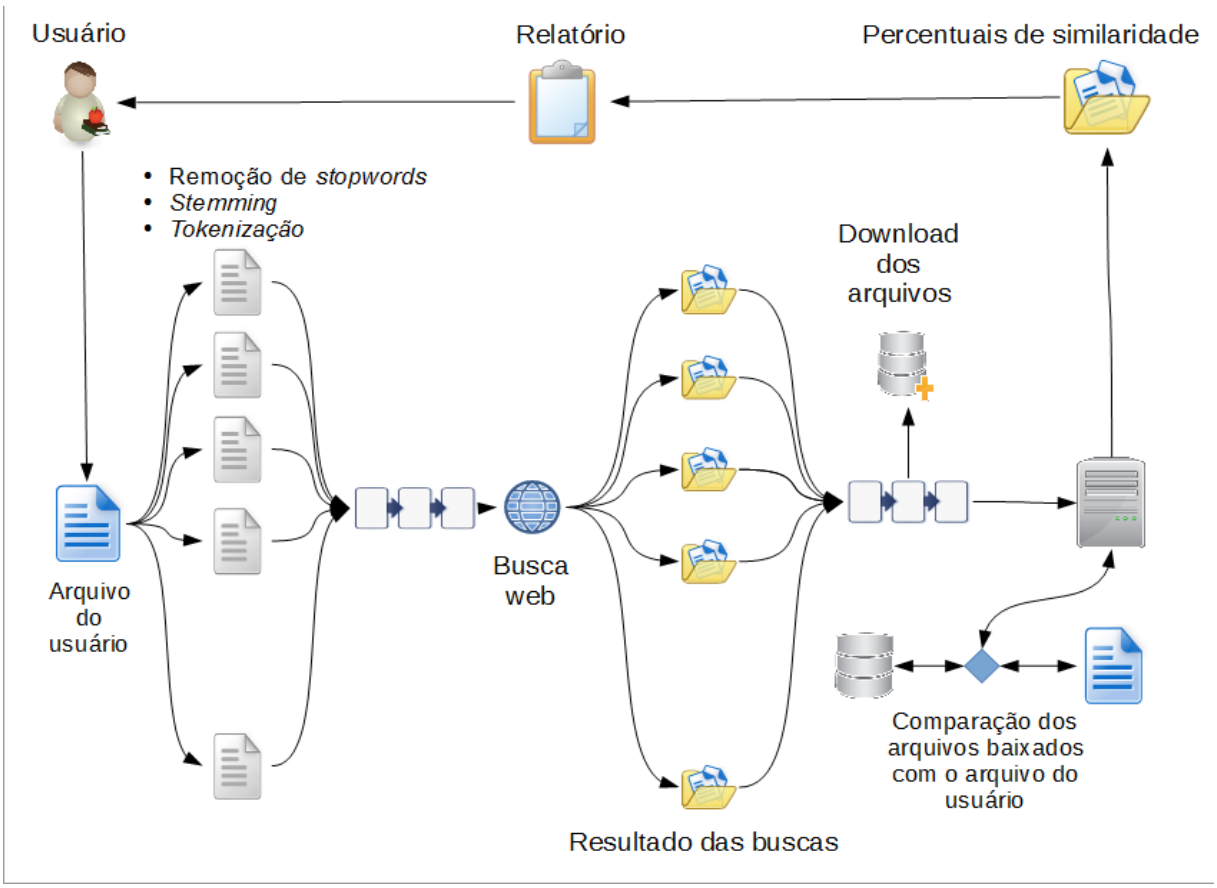
3.1.4 Pós-Processamento

Após o processo de análise e cálculo de similaridade, o pós-processamento trata da formatação e exibição dos resultados da análise ao usuário. O Miss Marple,

realiza isso na forma de um relatório que lista todos os arquivos baixados que obtiveram um índice de similaridade maior ou igual a 60% e, portanto possuem grandes chances de ter seu conteúdo plagiado de alguma forma. A partir daí, cabe ao usuário analisar os arquivos manualmente e determinar se o conteúdo semelhante realmente se trata de plágio.

Na figura 1 tem-se um diagrama com a modelagem atual do Miss Marple, que exemplifica os passos de execução do sistema desde a submissão do documento do usuário até a geração do relatório final contendo o resultado da análise.

Figura 1 – Diagrama de execução do Miss Marple



Fonte: Próprio autor

3.2 DESEMPENHO E NOVA IMPLEMENTAÇÃO

Ao analisar a estrutura de implementação do Miss Marple e sua fundamentação teórica, constatou-se que a causa dos problemas de desempenho apresentados se deve a como o processo de downloads funciona, pois a forma como ele está implementado permite que uma quantidade extremamente alta de arquivos precisem ser baixados. Como pode-se ver no diagrama da figura 1, todos os arquivos encontrados na busca são baixados, sem exceção. Essa carga excessiva de downloads faz com que o tempo de execução fique sobrecarregado. Além disso, muitas vezes ocorre de um mesmo documento se encontrar em mais de um resultado diferente e o sistema acaba fazendo downloads duplicados em função disso, diminuindo ainda mais a eficiência do tempo de execução.

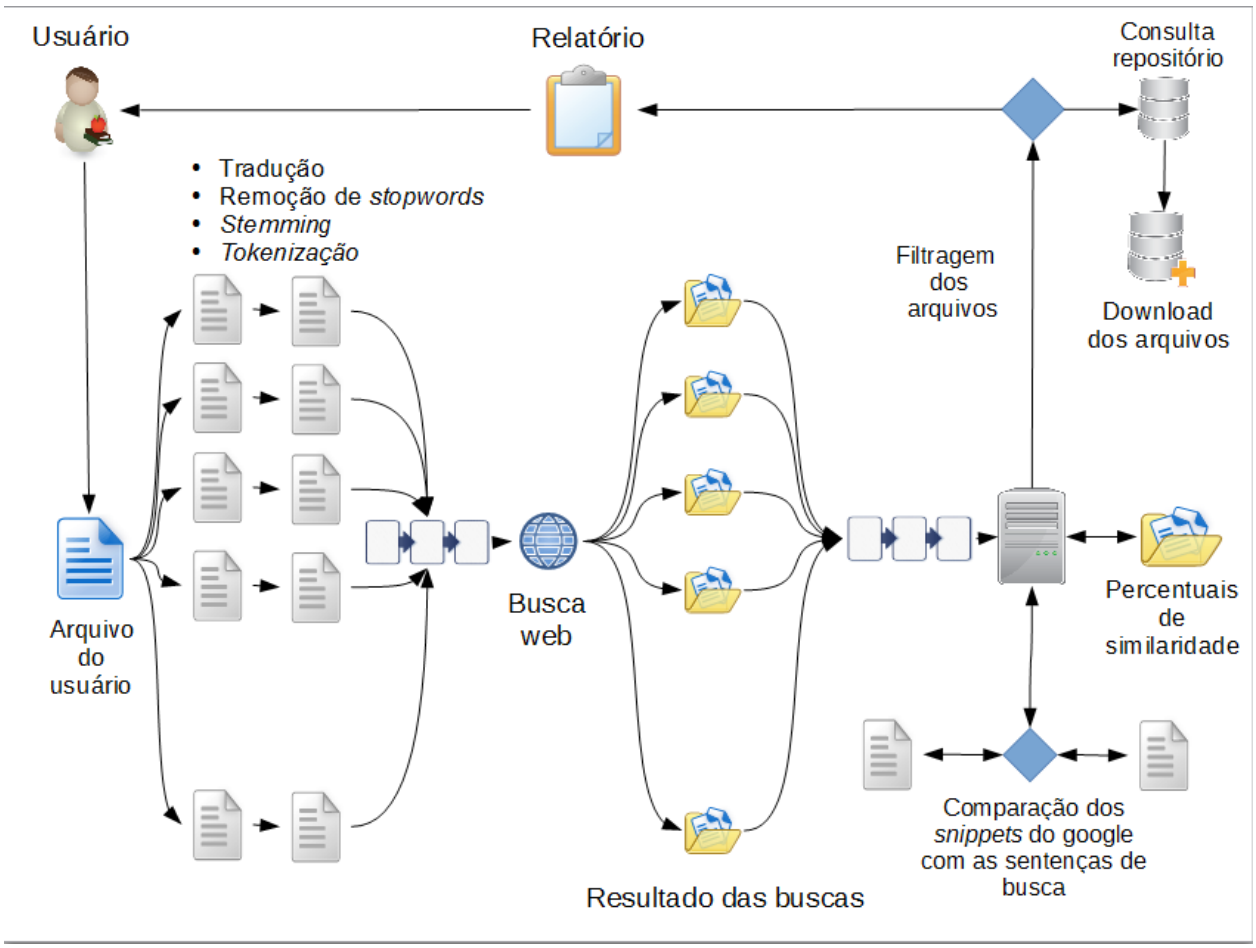
Para resolver esses problemas, foi necessário reimplementar o processo de downloads, de forma que ele seja mais eficiente. Abaixo estão listadas as principais mudança feitas na arquitetura do *Miss Marple* e na figura 2 tem-se o diagrama de execução após as mudanças:

a) Lentidão no processo de downloads: Trocou-se o sistema de download atual pelo mesmo sistema de *snippets* encontrado no método DIP. Dessa forma, são baixados apenas aqueles documentos cuja similaridade entre o texto pesquisado e o *snippet* retornado for maior que 60%, índice esse considerado pelo Miss Marple como de alta probabilidade de plágio. Isso faz com que a quantidade de arquivos baixados seja muito menor que a da implementação original;
b) Download de documentos duplicados: Para eliminar o download desnecessário de documentos duplicados, o repositório formado pelos arquivos baixados em análises anteriores passou a ser usado como uma base secundária de pesquisa. Assim, a cada documento encontrado na *internet*, sua existência é consultada no repositório. Em caso positivo, ele é reaproveitado. Essa consulta é feita através de um índice, criado através da biblioteca *Lucene*, de todas as urls baixadas e seus arquivos

correspondentes. Se a url do arquivo a ser baixado estiver indexado, significa que seu arquivo correspondente também se encontra no repositório;

c) Detecção bilíngue: Também foi adicionada a funcionalidade de detecção de plágio bilíngue. Para isso, criou-se uma fase adicional que traduz o documento submetido pelo usuário antes de passar para a fase de pré-processamento atual. Isso implicou também em adaptar os processos de remoção de *stopwords* e de *stemming* à língua inglesa, visto que a implementação original processa apenas textos em português;

Figura 2 – Diagrama de execução após a reimplementação



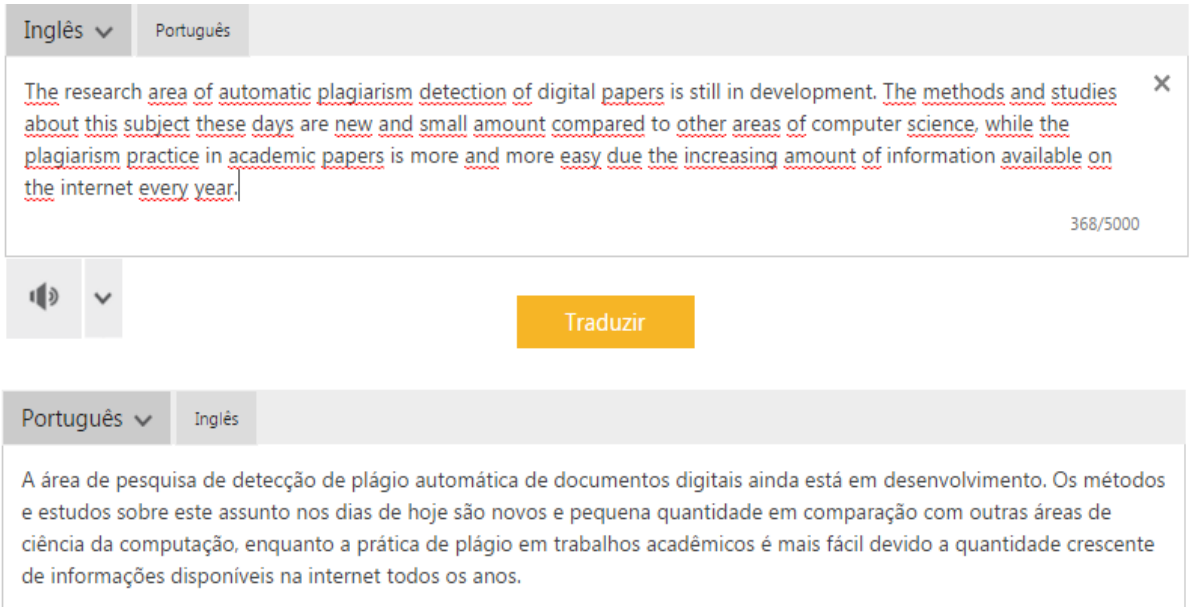
Fonte: Próprio autor

3.3 DIFICULDADES ENCONTRADAS

Das funcionalidades pretendidas neste trabalhos, todas puderam ser implementadas e foram postas em funcionamento com êxito. Porém, houve algumas dificuldades que se apresentaram ao longo do processo de desenvolvimento que trouxeram algumas limitações ao sistema. Uma delas, que já existia, é o fato de que a API de buscas utilizada, *Google Web Search API* é uma API gratuita que foi descontinuada por parte do *google* em detrimento de uma nova API paga. Ela continua funcionando normalmente, mas possui um limite de 1000 buscas diárias. Como o ambito deste trabalho é acadêmico e não comercial, optou-se por trabalhar nesta mesma API, ainda que com essas limitações. Pode-se pensar em migrar para a nova versão, caso haja esse interesse em trabalhos futuros.

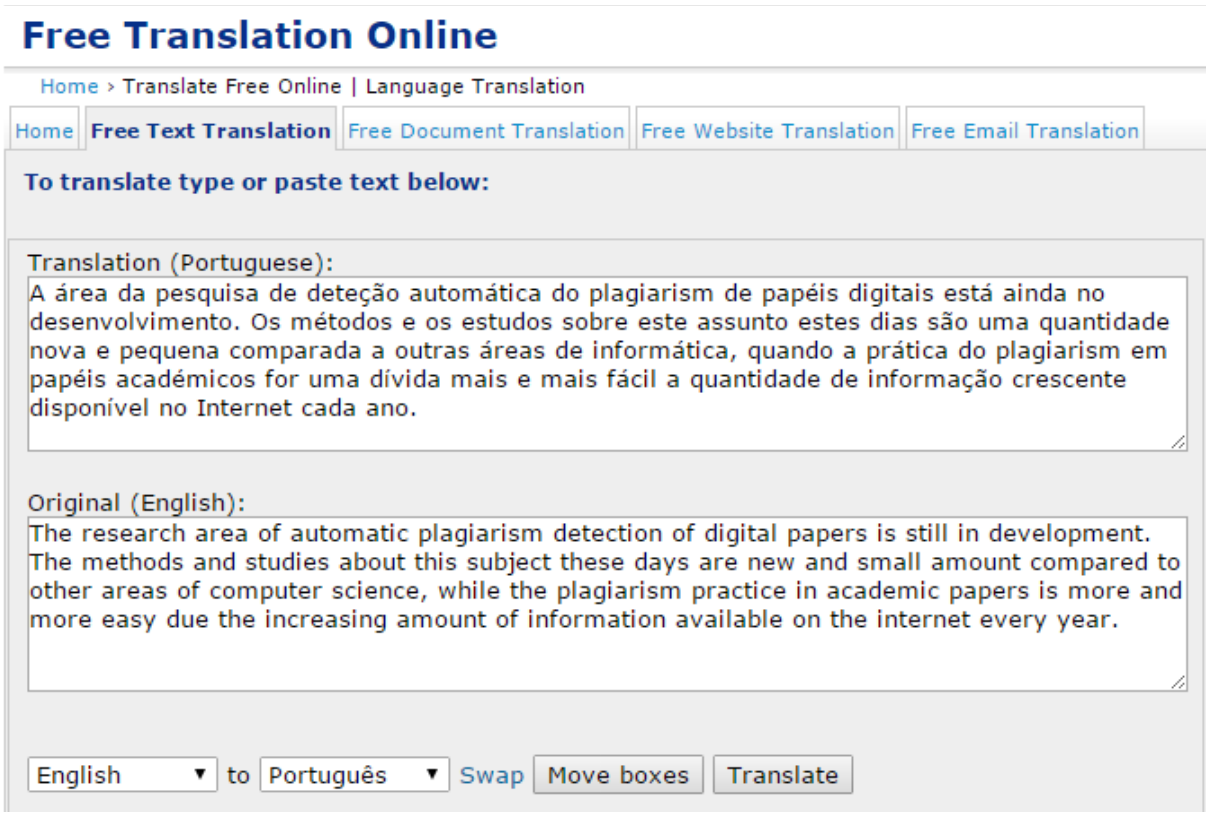
O mesmo tipo de problema ocorreu com a *Google Translate*, API de tradução do *Google* que pretendia-se usar como ferramenta de tradução para a detecção bilíngue. Já houve versões gratuitas desta API, mas hoje em dia existem apenas versões pagas, o que inviabilizou seu uso neste trabalho. Para contornar este problema, optou-se por utilizar a *Microsoft Translate*, API da Microsoft que realiza a mesma tarefa que a *Google Translate*. Ainda assim, sua versão gratuita limita a tradução a 2 milhões de caracteres por mês e exige autenticação. Existem outras ferramentas de tradução gratuitas e ilimitadas disponíveis, mas, após testes, constatou-se que a precisão da tradução deixou muito a desejar em comparação com a *Microsoft Translate*. Para exemplificar esta diferença de precisão, tem-se na figuras 3 um trecho do abstract deste trabalho traduzido do inglês para o português pelo Microsoft translator e nas figuras 4 e 5 o mesmo trecho traduzido nas ferramentas gratuitas WorldLingo e Translator2.

Figura 3 – Tradução realizada pelo Microsoft Translator



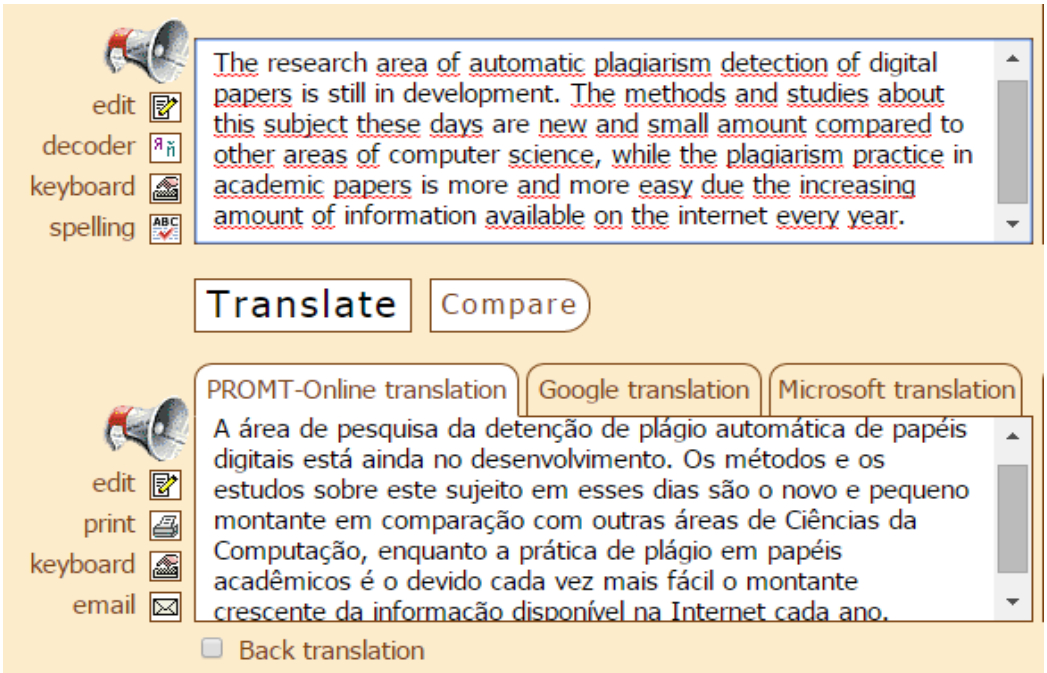
Fonte: https://www.bing.com/translator

Figura 4 – Tradução realizada pela ferramenta WorldLingo



Fonte: http://www.worldlingo.com/pt/products_services/worldlingo_translator.html

Figura 5 – Tradução realizada pela ferramenta Translator2



Fonte: http://translation2.paralink.com/

4 RESULTADOS

Neste capítulo são apresentados os resultados obtidos a partir dos testes realizados ao final da implementação do trabalho, bem como a comparação de desempenho entre a versão original e a nova versão do programa, a fim de validar a proposta deste trabalho.

4.1 TESTES

Conforme descrito no capítulo 2, os testes consistiram de um conjunto de execuções do programa em ambas as versões a partir de um conjunto de documentos selecionados cujos resultados foram analisados em relação a dois parâmetros: tempo de execução e número de arquivos baixados, sendo que quanto menor for o tempo de processamento e a quantidade de arquivos baixados, e quanto maior for a diferença em relação à versão antiga, melhor será a performance da nova versão. Como a versão antiga não realiza detecção bilíngue, para validar essa nova funcionalidade, foram selecionados artigos acadêmicos em inglês que fizeram parte do estudo bibliográfico deste trabalho. Como tais artigos encontram-se na *internet*, estes artigos foram propositalmente traduzidos do inglês para o português por ferramentas de tradução *online*, fora da ferramenta, e então enviados para a análise. O Miss Marple, por sua vez, fez a tradução inversa, do português para o inglês, e em seguida fez as buscas do texto traduzido. Ao final da análise, verificou-se se os artigos originais em inglês foram encontrados pelas buscas e baixados no repositório dos arquivos candidatos. Se encontrados, significa que a detecção bilíngue funciona como o esperado. Além dessa verificação, também foram avaliados os parâmetros considerados nos outros testes. Sendo assim, foram realizados três conjuntos de testes: dois para avaliar a detecção de plágio padrão, um para cada versão da ferramenta, e outro para avaliar a detecção de plágio bilíngue.

Em cada conjunto de testes, foram realizadas 5 análises consecutivas em cada um dos documentos selecionados e em seguida usou-se a média dos valores

destas 5 análises para se chegar a um resultado definitivo. Esta metodologia se deu em função de algumas oscilações encontradas nos resultados das análises causadas por algumas características da API de buscas *Google Web Search*, bem como fatores externos, como velocidade de conexão com a *internet*. No que diz respeito à API de buscas, observa-se uma variação na quantidade de resultados encontrados (e consequentemente baixados) a cada análise realizada. Ou seja, se em uma primeira análise uma determinada quantidade de arquivos é retornada, em análises posteriores esse número pode variar para mais ou para menos, geralmente com uma diferença de 1 a 3 arquivos, salvo raras exceções. Esta é uma característica intrínseca à API de buscas, a qual não se tem controle dentro da ferramenta e que ocorre desde a antiga versão do *Miss Marple*, que já foi validada anteriormente. Por isso, como uma forma de se obter resultados mais precisos, optou-se por realizar repetidas análises e adotar a média dos valores destas como resultado definitivo. Essas oscilações também ocorrem em função da velocidade de conexão com a *internet* no momento da análise*,* o que influencia diretamente no tempo de análise do documento, pois o tempo para se baixar os arquivos encontrados depende dela.

Os documentos selecionados para fazerem parte do conjunto de testes da detecção padrão foram extraídos da Biblioteca Digital de Trabalhos de Graduação, repositório online mantido pela UFSM que contém os trabalhos de graduação do curso de Ciência da Computação desta universidade. São dez trabalhos em arquivos pdf com uma média de quarenta e cinco páginas cada. Para os testes da detecção de plágio foram escolhidos dez arquivos em inglês utilizados no estudo bibliográfico deste trabalho.

4.2 RESULTADOS DOS TESTES

Abaixo estão relacionados os resultados de cada um dos testes realizados com os arquivos selecionados. Tem-se primeiramente os dados tabelados de forma numérica e logo após tem-se gráficos comparativos construídos a partir dos

resultados das tabelas a fim de comparar o desempenho de ambas as versões da ferramentas em função dos parâmetros analisados.

4.2.1 Tabelas de dados

Nas quatro primeiras tabelas tem-se os dados referentes à execução do Miss Marple em sua implementação original. As três primeiras apresentam os dados contabilizados de cada uma das cinco análises realizadas e na quarta tem-se as médias destes valores. Os dados contabilizados são: o tempo de execução do programa, o número de arquivos baixados durante o processo de análise e, destes, quantos deles possuem *score* de similaridade maior ou igual a 60%. Da quinta à oitava, tem-se as análises dos mesmos documentos na nova versão da ferramenta, com os mesmos dados contabilizados, com a adição de um parâmetro extra na última coluna da sétima tabela que verifica se aqueles arquivos com *score* de similaridade acima de 60% também foram encontrados na versão original do *Miss Marple*, a fim de confirmar se a precisão de análise encontrada na versão original se mantém na nova versão. Nas quatro últimas, os dados são da detecção de plágio bilíngue, sendo que na última tabela verifica-se se o arquivo original em inglês foi encontrado pelas buscas.

Tabela 1 – Tempo de processamento da versão antiga

Documento	Análise 1	Análise 2	Análise 3	Análise 4	Análise 5
DocumentoTeste1.pdf	55 min	57 min	56 min	55 min	56 min
DocumentoTeste2.pdf	53 min	55 min	56 min	54 min	54 min
DocumentoTeste3.pdf	1h 03 min	1h 02 min	1h 06 min	1h 05 min	1h 05 min
DocumentoTeste4.pdf	49 min	47 min	48 min	46 min	46 min
DocumentoTeste5.pdf	55 min	55 min	56 min	54 min	54 min
DocumentoTeste6.pdf	49 min	48 min	48 min	47 min	47 min
DocumentoTeste7.pdf	45 min	45 min	43 min	44 min	44 min
DocumentoTeste8.pdf	35 min	36 min	35 min	35 min	34 min

DocumentoTeste9.pdf	52 min	51 min	50 min	51 min	53 min
DocumentoTeste10.pdf	36 min	36 min	35 min	35 min	34 min

Tabela 2 – Número de arquivos baixados pela versão antiga nas 5 análises realizadas

Documento	Análise 1	Análise 2	Análise 3	Análise 4	Análise 5
DocumentoTeste1.pdf	111	114	112	112	113
DocumentoTeste2.pdf	81	82	80	82	80
DocumentoTeste3.pdf	109	111	112	110	110
DocumentoTeste4.pdf	96	96	98	97	97
DocumentoTeste5.pdf	128	130	129	130	128
DocumentoTeste6.pdf	84	86	85	84	84
DocumentoTeste7.pdf	87	85	84	86	84
DocumentoTeste8.pdf	70	70	72	73	72
DocumentoTeste9.pdf	93	91	92	92	90
DocumentoTeste10.pdf	70	69	68	70	69

Fonte: Próprio autor

Tabela 3 – Número de arquivos baixados pela versão antiga com score acima de 60%

Documento	Análise 1	Análise 2	Análise 3	Análise 4	Análise 5
DocumentoTeste1.pdf	7	11	9	8	8
DocumentoTeste2.pdf	8	7	8	11	9
DocumentoTeste3.pdf	7	10	12	9	9
DocumentoTeste4.pdf	2	2	4	3	3
DocumentoTeste5.pdf	6	8	7	9	7
DocumentoTeste6.pdf	7	10	8	8	9
DocumentoTeste7.pdf	5	8	7	6	6

DocumentoTeste8.pdf	4	5	5	6	6
DocumentoTeste9.pdf	10	7	8	9	7
DocumentoTeste10.pdf	6	4	4	6	5

Fonte: Próprio autor

Tabela 4 – Médias dos testes de detecção de plágio versão antiga

Documento	Tamanho do arquivo	Tempo médio de análise	Média de arquivos baixados	Média de score > 60%
DocumentoTeste1.pdf	4,233 KB	56 min	112	9
DocumentoTeste2.pdf	3,394 KB	54 min	81	9
DocumentoTeste3.pdf	2,132 KB	1h 04min	111	9
DocumentoTeste4.pdf	3,143 KB	47 min	97	3
DocumentoTeste5.pdf	1,255 KB	55 min	129	7
DocumentoTeste6.pdf	1,336 KB	48 min	85	8
DocumentoTeste7.pdf	1,042 KB	44 min	85	6
DocumentoTeste8.pdf	887 KB	35 min	71	5
DocumentoTeste9.pdf	4,657 KB	50 min	92	10
DocumentoTeste10.pdf	412 KB	51 min	70	5
Média	2,249 KB	50 min	93	7

Fonte: Próprio autor

Tabela 5 – Tempo de processamento da nova versão

Documento	Análise 1	Análise 2	Análise 3	Análise 4	Análise 5
DocumentoTeste1.pdf	10 min	12 min	12 min	11 min	11 min
DocumentoTeste2.pdf	6 min	7 min	9 min	9 min	8 min
DocumentoTeste3.pdf	8 min	11 min	10 min	8 min	8 min
DocumentoTeste4.pdf	3 min	3 min	4 min	5 min	5 min
DocumentoTeste5.pdf	11 min	10 min	12 min	10 min	9 min

DocumentoTeste6.pdf	10 min	9 min	7 min	10 min	11 min
DocumentoTeste7.pdf	8 min	8 min	9 min	8 min	7 min
DocumentoTeste8.pdf	7 min	4 min	5 min	6 min	6 min
DocumentoTeste9.pdf	7 min	10 min	9 min	10 min	10 min
DocumentoTeste10.pdf	6 min	5 min	5 min	7 min	6 min

Fonte: Próprio autor

Tabela 6 – Número de arquivos baixados pela nova versão

Documento	Análise 1	Análise 2	Análise 3	Análise 4	Análise 5
DocumentoTeste1.pdf	9	12	11	10	10
DocumentoTeste2.pdf	8	9	12	12	10
DocumentoTeste3.pdf	7	10	9	8	8
DocumentoTeste4.pdf	3	2	3	5	4
DocumentoTeste5.pdf	10	9	10	9	8
DocumentoTeste6.pdf	10	8	7	9	10
DocumentoTeste7.pdf	9	7	8	7	6
DocumentoTeste8.pdf	6	4	4	5	5
DocumentoTeste9.pdf	7	9	8	9	10
DocumentoTeste10.pdf	5	4	5	6	6

Fonte: próprio autor

Tabela 7 – Número de arquivos baixados pela nova versão com score acima de 60%

Documento	Análise 1	Análise 2	Análise 3	Análise 4	Análise 5	Constam na 1ª versão
DocumentoTeste1.pdf	9	12	11	10	10	sim
DocumentoTeste2.pdf	8	9	12	12	10	sim
DocumentoTeste3.pdf	9	10	12	9	8	sim
DocumentoTeste4.pdf	3	2	3	5	4	sim

DocumentoTeste5.pdf	10	9	10	9	8	sim
DocumentoTeste6.pdf	10	8	7	9	10	sim
DocumentoTeste7.pdf	9	7	8	7	6	sim
DocumentoTeste8.pdf	6	4	4	5	5	sim
DocumentoTeste9.pdf	7	9	8	9	10	sim
DocumentoTeste10.pdf	5	4	5	6	6	sim

Fonte: próprio autor

Tabela 8 – Médias dos testes de detecção de plágio da nova versão

Documento	Tamanho do arquivo	Tempo de análise	Arquivos baixados	Score > 60%
DocumentoTeste1.pdf	4,233 KB	11 min	10	10
DocumentoTeste2.pdf	3,394 KB	8 min	10	10
DocumentoTeste3.pdf	2,132 KB	9 min	8	8
DocumentoTeste4.pdf	3,143 KB	4 min	3	3
DocumentoTeste5.pdf	1,255 KB	10 min	9	9
DocumentoTeste6.pdf	1,336 KB	9 min	9	9
DocumentoTeste7.pdf	1,042 KB	8 min	7	7
DocumentoTeste8.pdf	887 KB	6 min	5	5
DocumentoTeste9.pdf	4,657 KB	9 min	9	9
DocumentoTeste10.pdf	412 KB	6 min	5	5
Média	2,249 KB	8 min	8	8

Fonte: Próprio autor

Tabela 9 – Tempo de processamento da detecção de plágio bilíngue

Documento	Análise 1	Análise 2	Análise 3	Análise 4	Análise 5
DocumentoTraduzido1.pdf	14 min	13 min	15 min	17 min	15 min
DocumentoTraduzido2.pdf	10 min	8 min	8 min	7 min	9 min

DocumentoTraduzido3.pdf	13 min	14 min	13 min	12 min	13 min
DocumentoTraduzido4.pdf	9 min	10 min	10 min	8 min	8 min
DocumentoTraduzido5.pdf	12 min	10 min	10 min	11 min	11 min
DocumentoTraduzido6.pdf	10 min	8 min	9 min	8 min	8 min
DocumentoTraduzido7.pdf	6 min	6 min	4 min	5 min	5 min
DocumentoTraduzido8.pdf	7 min	5 min	6 min	6 min	7 min
DocumentoTraduzido9.pdf	7 min	7 min	9 min	8 min	9 min
DocumentoTraduzido10.pdf	10 min	12 min	11 min	11 min	11 min

Fonte: Próprio autor

Tabela 10 – Quantidade de arquivos baixados pela detecção de plágio bilíngue

Documento	Análise 1	Análise 2	Análise 3	Análise 4	Análise 5
DocumentoTraduzido1.pdf	8	8	10	12	11
DocumentoTraduzido2.pdf	8	7	8	7	7
DocumentoTraduzido3.pdf	10	11	10	10	11
DocumentoTraduzido4.pdf	6	7	7	6	7
DocumentoTraduzido5.pdf	7	5	5	5	6
DocumentoTraduzido6.pdf	7	7	8	7	7
DocumentoTraduzido7.pdf	7	6	4	4	5
DocumentoTraduzido8.pdf	9	7	7	8	6
DocumentoTraduzido9.pdf	7	8	8	6	7
DocumentoTraduzido10.pdf	9	11	10	10	11

Fonte: Próprio autor

Tabela 11 – Número de arquivos baixados na detecção de plágio bilíngue com score acima de 60%

Documento	Análise 1	Análise 2	Análise 3	Análise 4	Análise 5
DocumentoTraduzido1.pdf	8	8	10	12	11
DocumentoTraduzido2.pdf	8	7	8	7	7
DocumentoTraduzido3.pdf	10	11	10	10	11
DocumentoTraduzido4.pdf	6	7	7	6	7
DocumentoTraduzido5.pdf	7	5	5	5	6
DocumentoTraduzido6.pdf	7	7	8	7	7
DocumentoTraduzido7.pdf	7	6	4	4	5
DocumentoTraduzido8.pdf	9	7	7	8	6
DocumentoTraduzido9.pdf	7	8	8	6	7
DocumentoTraduzido10.pdf	9	11	10	10	11

Fonte: Próprio autor

Tabela 12 – Médias dos testes de detecção de plágio bilíngue

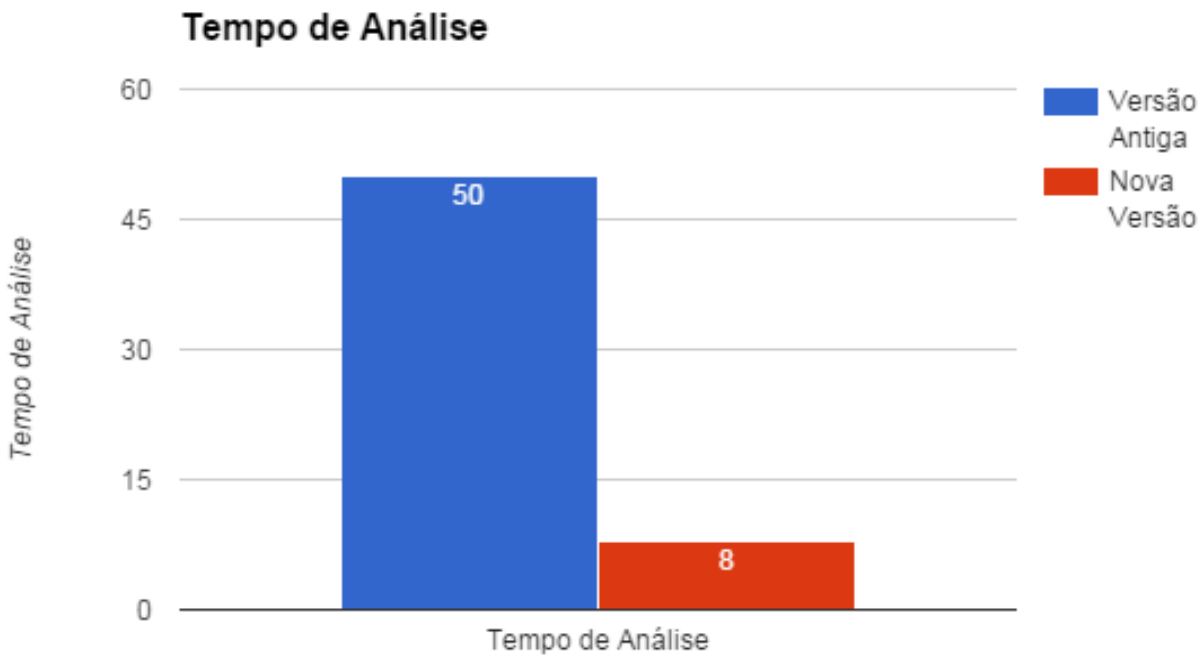
Documento	Tamanho do arquivo	Tempo de análise	Arquivos baixados	Score > 60%	Achou documento
DocumentoTraduzido1.pdf	614 KB	15 min	10	10	sim
DocumentoTraduzido2.pdf	422 KB	7 min	7	7	sim
DocumentoTraduzido3.pdf	400 KB	10 min	10	10	sim
DocumentoTraduzido4.pdf	578 KB	9 min	7	7	sim
DocumentoTraduzido5.pdf	177 KB	10 min	6	6	sim
DocumentoTraduzido6.pdf	92 KB	9 min	7	7	não
DocumentoTraduzido7.pdf	245 KB	5 min	5	5	sim
DocumentoTraduzido8.pdf	887 KB	7 min	10	10	sim
DocumentoTraduzido9.pdf	66 KB	8 min	7	7	não
DocumentoTraduzido10.pdf	258 KB	11 min	9	9	sim
Média	374 KB	9 min	8	8	

Fonte: Próprio autor

4.2.2 Gráficos comparativos

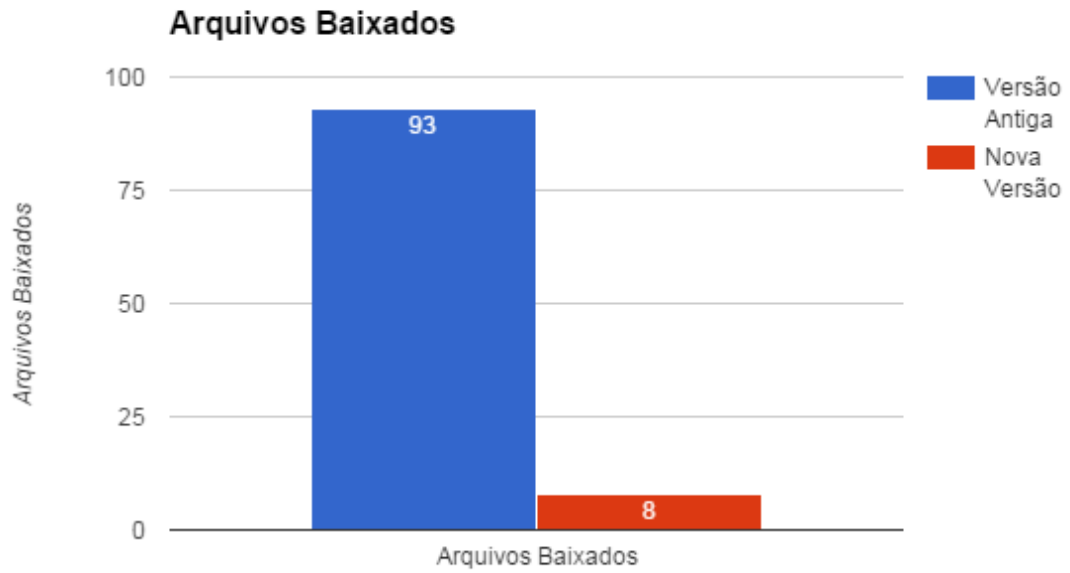
Os gráficos comparativos apresentam a diferença de desempenho encontrada entre as duas versões da ferramenta após a execução dos 3 tipos de teste e foram construídos a partir da média dos tempos de desempenho e da quantidade de arquivos baixados, obtidos em cada versão da ferramenta e na execução da detecção de plágio bilíngue. Estas médias foram extraídas da última linha das tabelas de resultados finais de cada versão. Os dois primeiros gráficos comparam os tempos de análise e o número de arquivos baixados na detecção padrão das duas versões da ferramenta. Nos dois últimos a comparação é entre a detecção padrão da nova versão da ferramenta e a detecção bilíngue.

Figura 6 – Gráfico comparativo das médias dos tempos de análise em minutos



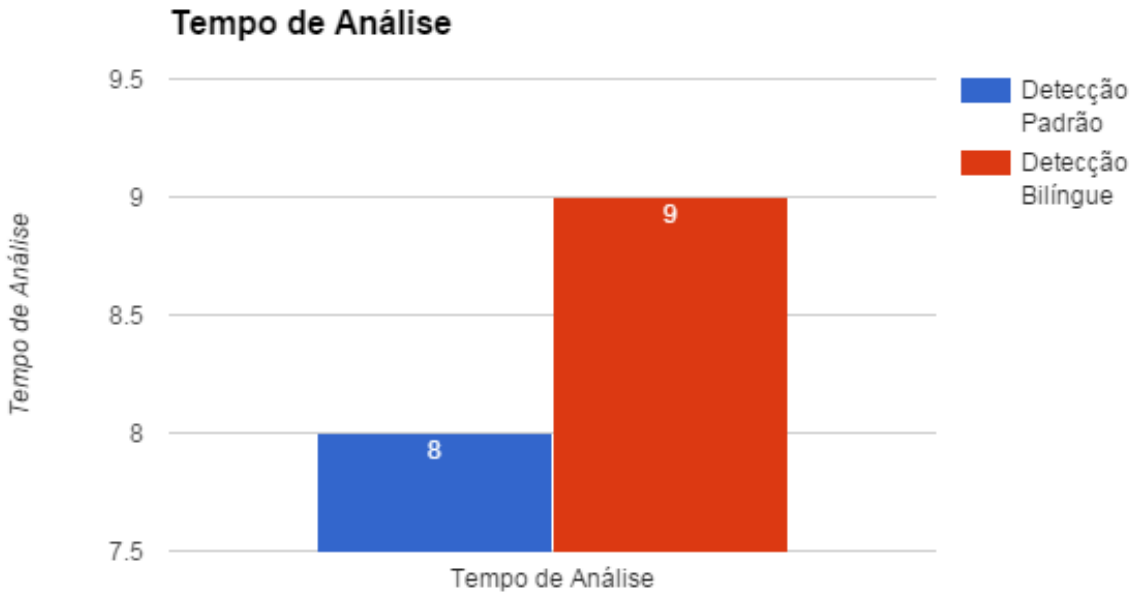
Fonte: Próprio autor

Figura 7 – Gráfico comparativo das médias do número de arquivos baixados



Fonte: Próprio autor

Figura 8 – Gráfico comparativo das médias do tempo de análise na detecção padrão e na detecção bilíngue



Fonte: Próprio autor

Figura 9 – Gráfico comparativo das médias do número de arquivos baixados pela detecção padrão e pela detecção bilíngue

Fonte: Próprio autor

4.3 INTERPRETAÇÃO DOS RESULTADOS

Diante dos resultados obtidos, nas tabelas de dados e nos gráficos comparativos, percebemos uma drástica queda no tempo de análise e na quantidade de arquivos baixados na nova versão da ferramenta em comparação com sua versão anterior. Em relação ao tempo de análise, essa queda foi de cerca de 84%, passando de uma média de 50 minutos para 8 minutos de análise. Em se tratando da quantidade de arquivos baixados, a média passou de 93 para 8 arquivos, uma queda de mais de 90%. Esta diferença de performance pode ser explicada principalmente pela nova política de download de arquivos adotada na nova implementação, em que apenas arquivos com *score* de similaridade maiores ou iguais a 60% são baixados para o repositório. Isto implica em uma carga de downloads muito menor, como pode ser visto na quantidade de arquivos baixados na nova versão. Ao analisarmos a última coluna das tabelas de dados, podemos verificar a quantidade de arquivos irrelevantes baixados na versão anterior. De uma

média de 93 arquivos, apenas 7 possuem 60% de similaridade ou mais, isto é, têm grandes chances de serem fontes de plágio e são relevantes para a análise. Os demais arquivos não chegam a atingir este score e, portanto, não há a necessidade de serem baixados. Na nova versão da ferramenta, esta diferença desaparece, pois todos os arquivos baixados necessariamente possuem score maior ou igual a 60%. Outro fator que contribui para uma menor quantidade de downloads é a introdução da verificação prévia da existência do arquivo a ser baixado no repositório, fazendo com que não haja o download de documentos duplicados. Também constata-se que a precisão de análise encontrada na versão original da ferramenta mantém-se a mesma, pois os mesmos arquivos relevantes encontrados por esta versão também são encontrados pela versão atualizada.

Em relação à detecção de plágio bilíngue, observa-se também uma melhor performance, afinal, trata-se da nova versão da ferramenta. Porém, ao compararmos com a detecção de plágio comum, de textos em português, observa-se um ligeiro aumento no tempo de análise enquanto o número de arquivos baixados atinge a mesma média, mesmo com arquivos consideravelmente menores como pode ser visto na segunda coluna da tabela. Isto pode ser explicado ao se levar em conta que a quantidade de material de lingua inglesa encontrada na *internet* é superior em relação à quantidade de material de língua portuguesa. Isso tende a fazer com que mais resultados sejam encontrados pelo motor de busca e, consequentemente, mais arquivos sejam baixados, elevando o tempo de análise. Em relação à precisão da análise bilíngue, dos 10 arquivos, 8 foram encontrados durante a busca, o que corresponde a 80% do total. Ainda assim, os resultados obtidos foram bastante satisfatórios.

Em ambos os casos, pode-se constatar que o processo de download é o principal responsável pelo tempo de análise de um documento e pela performance da ferramenta como um todo. Isto se mostrou bem claro nos dados obtidos nos testes realizados e no ganho de performance obtido após a reestruturação deste processo.

5 CONCLUSÃO

Ao final deste trabalho, ao se analisar os resultados obtidos frente às propostas apresentadas, pode-se concluir que, de uma forma geral, os objetivos almejados foram alcançados de forma bastante satisfatória, seja em relação à melhoria de desempenho do programa, objetivo este que motivou o início deste trabalho, assim como o objetivo de se implementar a detecção de plágio bilíngue, tendo-se assim uma ferramenta mais eficiente tanto em termos de desempenho quanto de funcionalidades.

Das dificuldades encontradas ao longo do processo de desenvolvimento, surgiram questões não abordadas neste trabalho que podem vir a ser estudadas em trabalhos futuros. Uma delas é a limitação da quantidade de buscas diárias imposta pela versão gratuita da API de buscas *Google Search API*, utilizada neste trabalho pelo custo benefício que proporciona e pelo âmbito acadêmico do trabalho. O mesmo ocorre com a API de tradução *Microsoft Translator*, que limita a tradução a 2 milhões de caracteres por mês. Outra questão levantada é a possibilidade de extender a busca *web* às imagens encontradas junto ao texto do documento que está sendo analisado. Afinal, assim como o plágio do próprio texto, há a possibilidade de também haver imagens plagiadas no documento**.**

Por fim, do ponto de vista de pesquisa, este trabalho foi uma contribuição a uma área ainda pouco explorada e que tem uma crescente demanda no que diz respeito ao desenvolvimento de técnicas e de ferramentas, tendo em vista que a detecção automática de plágio em trabalhos acadêmicos e científicos de uma forma geral se faz cada vez mais necessária em função da grande expansão e popularização que a internet vem sofrendo ao longo dos últimos anos.

REFERÊNCIAS

AREFIN, MOHAMMAD SHAMSUL., MORIMOTO, YASUHIKO., SHARIF, MOHAMMAD AMIR. BAENPD: **A Bilingual Plagiarism Detector**. Disponível em: <http://ojs.academypublisher.com/index.php/jcp/article/view/jcp080511451156/6788> Acesso em: ago. 2015.

ARENHARDT, CATIANE PRISCILA BARBOSA. ***MISS MARPLE* – Desenvolvimento de Ferramenta Para Auxiliar Na Verificação De Indícios De Plágio Com Base No Método DIP – Detector De Indícios de Plágio**. Dissertação de Mestrado – Universidade Federal de Santa Maria, Rio Grande Do Sul, 2012.

CARNAHAN, NOAH. *et al*. **Plagiarism Detection**. 2014. Disponível em: <http://www.cs.carleton.edu/cs_comps/1314/dlibenno/final-results/plagcomps.pdf> Acesso em: out. 2015.

COSTA-JUSSÀ, MARTA R. **Study and Comparison of Rule-Based and Statistical Catalan-Spanish Machine Translation Systems**. Disponível em: <http://cai.type.sk/content/2012/2/study-and-comparison-of-rule-based-and-statistical-catalan-spanish-machine-translation-systems/> Acesso em: out. 2015.

DANILOVA, VERA. **Cross-Language Plagiarism Detection Methods**. Disponível em: <http://www.aclweb.org/anthology/R13-2008> Acesso em: set. 2015.

PERTILE, SOLANGE DE LURDES. **Desenvolvimento e Aplicação de um Método Para Detecção de Indícios de Plágio**. Dissertação de Mestrado – Universidade Federal de Santa Maria, Rio Grande do Sul, 2011.

HOAD, TIMOTHY C., ZOBEL, JUSTIN. **Methods For Identifying Versioned and Plagiarised Documents**. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.442.6430&rep=rep1&type=pdf> Acesso em: set. 2015.

HUANG, ANNA. **Similarity Measures For Text Document Clustering**. Disponível em: <http://nzcsrsc08.canterbury.ac.nz/site/proceedings/Individual_Papers/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf> Acesso em: set. 2015.

KARAMI, OMID. **The Brief View on Google Translate Machine**. 2010. Disponível em: <http://logic.at/lvas/185054/GoogleTranstaeMachineBriefView_OmidKarami.pdf> Acesso em: out. 2015.

KENT, KOK CHOW., SALIM, NAOMIE. **Web Based Cross Language Plagiarism Detection**. Disponível em: <http://arxiv.org/ftp/arxiv/papers/0912/0912.3959.pdf> Acesso em: ago. 2015.

KIRKPATRICK, J. **Teaching acknowledgement practice using the internet-based plagiarism detection service. Marketing Education Review**. 2006.

LOPEZ, ADAM. **Statistical Machine Translation**. 2010. Disponível em: <http://www.mtmarathon2010.info/web/Program_files/survey.pdf > Acesso em: out. 2015.

NASEEM RASIA., KURIAN, SHEENA. **Extrinsic Plagiarism Detection Detection in Text Combining Vector Space Model and Fuzzy Semantic Similarity Scheme**. 2013. Disponível em: <http://www.iracst.org/ijacea/papers/vol2no62013/1vol2no6.pdf> Acesso em: out. 2015.

PEREIRA, RAFAEL COREZOLA. **Cross-Language Plagiarism Detection**. Dissertação de Mestrado – Universidade Federal do Rio Grande do Sul, 2010.

POTTHAST *et al*. (2011). **Cross-Language Plagiarism Detection. Language Resources and Evaluation**. 45(1):45-62. doi:10.1007/s10579-009-9114-z.

STEIN, *et al*. **Cross-Language Plagiarism Detection**. Disponível em: <https://riunet.upv.es/bitstream/handle/10251/37479/Cross-Language%20Plagiarism%20Detection.pdf?sequence=2> Acesso em: set. 2015.

ZHAN *et al*. **Plagiarism detection using the Levenshtein distance and Smithwaterman algorithm**. Disponível em: <http://dx.doi.org/10.1109/ICICIC.2008.422> Acesso em: out. 2015.