UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Tácio Vinícius Bernardes Ribeiro

UM ESTUDO DE CASO NA AVALIAÇÃO AUTOMÁTICA DE QUESTÕES DISCURSIVAS COM ANALISE SEMÂNTICA LATENTE

Belém

2012

Tácio Vinícius Bernardes Ribeiro

UM ESTUDO DE CASO NA AVALIAÇÃO AUTOMÁTICA DE QUESTÕES DISCURSIVAS COM ANALISE SEMÂNTICA LATENTE

Dissertação de Mestrado apresentada para obtenção do grau de Mestre em Ciência da Computação. Programa de Pós-Graduação em Ciência da Computação, Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará.
Área de concentração: Inteligência Artificial
Orientador: Prof. Dr. Eloi Luiz Favero.
Co-orientador: João Carlos Alves dos Santos

Belém

2012

Ribeiro, Tácio Vinícius Bernardes

Um Estudo de Caso na Avaliação Automática de Questões Discursivas com Analise Semântica Latente / (Tácio Vinícius Bernardes Ribeiro); orientador, Eloi Luiz Favero. – 2012.

75 p. il. 28 cm

Dissertação (Mestrado) – Universidade Federal do Pará. Instituto de Ciências Exatas e Naturais. Programa de Pós-Graduação em Ciência da Computação. Belém, 2012.

1. Inteligência artificial. 2. Processamento de linguagem natural (Computação). 3. Avaliação educacional. I. Favero, Eloi Luiz, orient. II. Universidade Federal do Pará, Instituto de Ciências Exatas e Naturais, Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 22.ed. 006.3

Tácio Vinícius Bernardes Ribeiro

UM ESTUDO DE CASO NA AVALIAÇÃO AUTOMÁTICA DE QUESTÕES DISCURSIVAS COM ANALISE SEMÂNTICA LATENTE

Dissertação de Mestrado apresentada para obtenção do grau de Mestre em Ciência da Computação. Programa de Pós-Graduação em Ciência da Computação, Instituto de Ciências Exatas e Naturais. Universidade Federal do Pará.

Data da aprovação: Belém-Pa. 21 - 09 - 2012

Banca Examinadora

_____

Prof. Dr. Eloi Luiz Favero
(ORIENTADOR – UFPA)

_____

Prof. Dr. Bianchi Serique Meiguins
(MEMBRO-UFPA)

_____

Prof. Dr. Joaquim Carlos Barbosa Queiroz
(MEMBRO-UFPA)

VISTO:

_____

Prof. Dr. Sandro Ronaldo Bezerra Oliveira
(COORDENADOR DO PPGCC/ICEN/UFPA)

RESUMO

Este trabalho apresenta um estudo de caso de avaliação automática de questões discursivas (ensaios) baseada na técnica LSA (*Latent Semantic Analysis* ou Análise Semântica Latente), bem como uma ferramenta que apoia o ajuste automático de parâmetros para este domínio do problema. Nesta abordagem foram consideradas técnicas de pré-processamento (retirada de *stop words* e aplicação de *stemming*) em combinação com a técnica de n-gramas (unigramas e bigramas); função peso (ponderação); dimensão do espaço reduzido e medida de similaridade. Esse estudo de caso envolveu as provas das disciplinas de biologia e geografia do processo seletivo vestibular da Universidade Federal do Pará (UFPA) ocorrido em 2008, onde foi feita a comparação entre as notas calculadas automaticamente pela ferramenta e as notas atribuídas pelos avaliadores humanos durante o processo de correção do vestibular. Os melhores resultados alcançados nessa comparação foram os da disciplina de geografia com uma acurácia de 86,89% usando as seguintes técnicas: unigramas sem nenhuma técnica de processamento; redução ao espaço semântico de 3 dimensões; ponderação local binária; e a correlação de Pearson como similaridade. Destacam-se também os melhores resultados da disciplina biologia onde se alcançou uma acurácia 84,77% com a utilização das seguintes técnicas: bigramas com a técnica de remoção de *stop words*; redução do espaço semântico a dimensões 6; ponderação local da norma euclidiana dividida pela soma dos componentes; e o cosseno como medida de similaridade.

PALAVRAS-CHAVES: LSA, Avaliação Automática, Calibração de Parâmetros, Questões Discursivas.

ABSTRACT

This work presents a case study based in the LSA (Latent Semantic Analysis) for automatic evaluation in open ended questions (essays) and a tool that supports the automatic adjustment of parameters for each problem domain. This methodology uses some techniques like: preprocessing techniques (stop words removal, and stemming) together with n-grams techniques (unigrams and bigrams); weight function; reduced space dimension; and similarity measures. This case study was done involving admission tests of the biology and geography disciplines from Federal University of Pará (UFPA) occurred in 1998. In this case study it was realized a comparison between the grades given by the tool and the grade given by the human evaluators during the admission test assessment process. The best results reached in this comparison, are from the geography discipline with an accuracy of 86.89% using the follow techniques: unigrams without any preprocessing techniques; a reducing to the semantic space of 3 dimensions; local binary weighting; and the Person correlation similarity. Also noteworthy are the best results from biology discipline with accuracy of 84.77% using the follow techniques: bigrams with the stop words removing technique; a reducing to the semantic space of 6 dimensions; the Euclidian norm divided by the components sum as the local weighting; and the cosine as similarity measure.

KEYWORDS: LSA, Automatic Evaluation, Parameter calibration, Open Ended Questions.

LISTA DE SIGLAS

AVA Ambiente Virtual de Aprendizagem

LSA *Latent Semantic Analysis* (Análise Semântica Latente)

SVD *Singular Value Decomposition* (Decomposição em valores singulares)

UFPA Universidade Federal do Pará

IDF *Inverse document frequency* (Inverso da frequência do termo entre documentos)

SUMÁRIO

LISTA DE QUADROS

LISTA DE ILUSTRAÇÕES

1. INTRODUÇÃO

1.1 MOTIVAÇÃO

Durante a sua vida escolar, o aluno passa por um processo de avaliação de ensino aprendizagem contínuo, cumulativo e sistemático. Mesmo diante das concepções pedagógicas mais modernas, a aplicação de avaliações compostas por questões discursivas tem forte predominância, pois avaliam a capacidade de leitura, interpretação e construção do texto. No entanto, a tarefa de correção manual desse tipo avaliação, para um número considerável de alunos, demanda muito tempo do professor. Principalmente no contexto de um processo seletivo de grande alcance com avaliações compostas por questões discursivas, onde esta tarefa torna-se muito demorada e custosa. Nesta problemática, o desenvolvimento de ferramentas que automatizem a correção de respostas a questões discursivas torna-se relevante. Pesquisadores acreditam que os computadores podem ser utilizados para ajudar os professores na tarefa de avaliação(PÉREZ *et al.*, 2005), formando assim as bases do campo de pesquisa conhecido como "Avaliação assistida por computador" (*Computer-Assisted Assessment*). Esta área de pesquisa trata do desenvolvimento de ferramentas que automatizem a correção de respostas a questões discursivas, que é o tema abordado por este trabalho.

A partir dos anos 90 esse campo de pesquisa teve um avanço considerável devido à aplicação de técnicas de processamento de linguagem natural e de extração de informações (PÉREZ *et al.*, 2005), mas até hoje este problema ainda não foi totalmente resolvido (PÉREZ *et al.*, 2005). Com o crescimento da educação à distância, os ambientes virtuais de ensino como Moodle (MOODLE, 2012) e Blackboard (BLACKBOARD, 2012) vem sendo cada vez mais utilizados. Neste contexto cresce a importância do desenvolvimento de componentes de avaliação automática de questões discursivas, para que possam dar um *feedback* automático e imediato aos alunos a respeito do seu desempenho, também para questões discursivas.

Ainda sobre essa área de pesquisa, (VALENTI, NERI e CUCCHIARELLI, 2003) afirma que existem várias abordagens para a solução do problema de avaliação automática para questões discursivas, entre essas abordagens pode-se destacar: a combinação de métodos baseados em palavras chaves com análise profunda de texto (BURSTEIN, LEACOCK e SWARTZ, 2000); técnicas de correspondência de padrões (MING, MIKHAILOV e LAY KUAN, 2000); a quebra de respostas em conceitos e as suas dependências semânticas (CALLEAR *et al.*, 2001); a combinação de classificadores Bayesianos com outras técnicas de aprendizado de máquina (ROSÉ *et al.*, 2003); pela redução da dimensão do espaço com o

LSA (Análise Latente Semântica); LSA incrementado com informações sintáticas e semânticas (LANDAUER, FOLTZ e LAHAM, 1998) (KANEJIYA, KUMAR e PRASAD, 2003).

Um dos métodos promissores para avaliação automática de questões discursivas é a LSA (*Latent Semantic Analysis*) (DEERWESTER *et al.*, 1990). Entretanto a eficácia dessa técnica depende fortemente do ajuste de seus parâmetros e do domínio de aplicação (LIFCHITZ, JHEAN-LAROSE e DENHIÈRE, 2009). Percebendo-se então a necessidade da ferramenta apoiar o ajuste de parâmetros para cada domínio de aplicação. Diante desses desafios, este trabalho visa contribuir para o meio acadêmico com um estudo de avaliação automática de questões discursivas, focando no ajuste dos parâmetros que influenciam na acurácia do método de forma automática, de modo que essa ferramenta necessite da menor quantidade de interação humana possível para a conclusão do seu processamento.

Para se trabalhar com ajuste de parâmetros para o método LSA precisa-se de uma base de respostas de questões discursivas reais. Felizmente na UFPA temos milhares de respostas de processos seletivos, onde as questões são de provas reais e já possuem avaliação de pelo menos dois avaliadores humanos. Neste trabalho exploraremos o uso destas questões.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Desenvolver um estudo de caso para avaliar a viabilidade de utilizar a técnica de analise semântica latente (LSA) para correção automática de questões com respostas discursivas (pequenos “ensaios”) do processo seletivo da UFPA. Construir uma ferramenta que permita experimentos diversos variando-se e/ou ajustando-se os parâmetros da técnica LSA buscando uma acurácia máxima.

1.2.2 Objetivos Específicos

O objetivo geral pode ser decomposto numa lista de objetivos específicos, a saber:

- Fazer um levantamento bibliográfico sobre avaliação automática, relatando sistemas e abordagens existentes bem como suas medidas de acurácia;
- Preparar uma revisão da literatura sobre o método LSA descrevendo os conceitos da abordagem, bem como as técnicas de ajustes de parâmetros,

criando um texto com os fundamentos teóricos para a equipe que trabalhará com LSA, no contexto do projeto onde este estudo esta sendo realizado;

- Compor uma base de questões de processos seletivos da UFPA para poder testar o método de LSA;
- Investigar e ajustar parâmetros para uso prático da técnica. A ferramenta desenvolvida deve apoiar o ajuste de parâmetros automático para o problema do estudo de caso;
- Realizar os experimentos com a técnica, investigando quais parâmetros trazem os melhores resultados.

Neste estudo utilizou-se uma questão conceitual da disciplina de biologia e outra questão analítico-discursiva da disciplina de geografia. Ambas as questões são oriundas do processo seletivo da Universidade Federal do Pará (UFPA).

Para validação dos resultados do estudo de caso, as notas dadas pela ferramenta foram comparadas com a média dada pelos avaliadores humanos durante a correção das provas de vestibular.

1.3 RELEVÂNCIA DO TRABALHO

Além das contribuições acadêmicas citadas acima, espera-se que o resultado desse trabalho possa vir a ser aplicado em processos seletivos, Ambientes Virtuais de Aprendizagens (AVAs) ou até mesmo no dia a dia do professor a fim de que professores e alunos pudessem usufruir dos seguintes benefícios:

- Em um AVA, os alunos poderiam ter um *feedback* imediato, para as avaliações e exercícios que envolvam questões discursivas;
- Em um processo seletivo (vestibular) poderia haver uma diminuição de custos financeiros e tempo para a realização de processos seletivos com avaliações discursivas. Considerando que a ferramenta poderia ser utilizada para fazer uma triagem inicial das provas que seriam corrigidas por um avaliador humano;
- No dia a dia do professor, o mesmo teria uma liberação da sua carga de trabalho, devido à diminuição do tempo gasto com a correção manual dessas respostas.

- Por outro lado, no projeto de pesquisa, contexto do trabalho, o estudo de caso permite um aprofundamento no uso da técnica LSA e nos seus fundamentos trazendo o conhecimento prático para toda a equipe;

Como principais resultados do estudo de caso obteve-se uma acurácia de 84,77% para a aplicação da ferramenta proposta nas questões da disciplina de Biologia da terceira fase processo seletivo seriado da Universidade Federal do Pará, enquanto que para as questões de Geografia do mesmo concurso, a maior acurácia atingida foi de 86,89%.

1.4 CONTEXTO DE REALIZAÇÃO DO TRABALHO

Este estudo está realizado dentro de um projeto de pesquisa maior que visa a "**Avaliação automática de questões não objetivas"** liderado pelo Prof. Eloi Favero onde também participam o Prof. João Carlos Alves dos Santos (estudante de doutorado) além de bolsistas de iniciação científica. O projeto já realizou pesquisas com avaliação de questões discursivas com bigramas(SANTOS *et al.*, 2007), Mapas Conceituais (CALDAS e FAVERO, 2009), avaliação de Programas (LINO, FAVERO e SILVA, 20007). Neste contexto o presente trabalho visa fazer um estudo de caso com a técnica de análise semântica latente.

1.5 ORGANIZAÇÃO DO TEXTO

Além deste Capítulo, este trabalho possui mais cinco seções, onde o Capítulo 2 apresenta o principal campo de pesquisa no qual este trabalho está envolvido, bem como os trabalhos relacionados à mesma; o Capítulo 3 apresenta a técnica LSA, bem como os conceitos associados à mesma; o Capítulo 4 apresenta a metodologia de uso do LSA utilizada neste trabalho; o Capítulo 5 apresenta as resultados obtidos nessa pesquisa; e o Capítulo 6, por fim, apresenta as conclusões.

2. AVALIAÇÃO ASSISTIDA POR COMPUTADOR PARA QUESTÕES DISCURSIVAS

2.1 INTRODUÇÃO

Avaliar as respostas discursivas de alunos é uma atividade que consome muito tempo do professor e faz com que eles tenham menos tempo para exercer as outras atividades que a profissão exige (PÉREZ *et al.*, 2005), chegando ao ponto de inclusive ser a atividade que consome mais o tempo do professor (MASON e GROVE-STEPHENSON, 2002). Diante dessa situação, muitos autores acreditam que a mesma pode ser resolvida através da utilização do computador como uma ferramenta de avaliação, mas sempre deixando claro que o objetivo não é substituir o professor com o computador, mas sim ajudar o professor com um programa de computador (MASON e GROVE-STEPHENSON, 2002). Essa tarefa é o principal objetivo do campo de pesquisa denominado "Avaliação Assistida por Computador" (*Computer Assisted Assessment*) (WHITTINGTON e HUNT, 1999). Este tema apresenta um problema que ainda não é considerado totalmente resolvido e também vem ganhando mais atenção graças ao sucesso dos AVA e os avanços das áreas de Extração de informação e Processamento de linguagem.

Uma das dificuldades para avaliação de respostas a questões discursivas é a natureza subjetiva da atividade, que em vários casos faz com que uma mesma resposta receba notas com uma diferença grande, quando a avaliação é feita por mais de um avaliador. O que é percebido pelos alunos como uma grande fonte de injustiça (VALENTI, NERI e CUCCHIARELLI, 2003). Tal dificuldade reforça o fortalecimento desta área de pesquisa, já que um computador pode examinar e analisar respostas discursivas com muito mais detalhes que um ser humano e é totalmente livre de preconceito, mitos e vícios. Sendo assim irá sempre avaliar todas as resposta de uma mesma forma.

2.2 TRABALHOS CORRELATOS: **Avaliação Assistida por Computador**

2.2.1 Histórico

As pesquisas em "Avaliação Assistida por Computador" começaram em 1966 e teve como pioneiro o *Project Essay Grader* (PEG) (PAGE, 1966). Este se baseava na análise do estilo de escrita, pontuando assim o texto pela sua qualidade e não pelo seu conteúdo. Além disso, nessa primeira versão do projeto, os resultados tinham uma acurácia baixa e foi alvo de muitas críticas da comunidade acadêmica. Esses resultados fizeram com que a pesquisa nessa área ficasse quase estagnada até a década de 90.

Na década de 90, graças aos avanços de outras áreas de pesquisa, como por exemplo, o "Processamento de Linguagem Natural" e "Recuperação de Informação", as pesquisas na área de "Avaliação Assistida por Computador" voltaram a surgir. Em 1990 começou a ser desenvolvido o *Educational Testing Service* I (ETS) (WHITTINGTON e HUNT, 1999) que se baseava em medidas de qualidade de escrita mais diretas, em comparação ao PEG, no entanto ainda não considerava o conteúdo das respostas para atribuição das notas, envolvia trabalho manual e só funcionava em sentenças pequenas.

Em 1997 o *Project Essay Grader* evoluiu a ponto de mostrar resultados melhores e ficar comercialmente viável. Além disso, surgiram três novos projetos para contribuir com esta área: *Intelligent Essay Assessor (*IEA*)* (FOLTZ, LAHAM e LANDAUER, 1999); *E-Rater* (BURSTEIN *et al.*, 1998); *IntelliMetric* (VANTAGE LEARNING TECH, 2000). O IEA é um sistema baseado em LSA que foca-se no conteúdo do texto; O *E-Rater* é uma versão aprimorada do ETS que usa uma abordagem híbrida combinando Processamento de Linguagem Natural e Técnicas Estatísticas para análise do estilo do texto; e o *IntelliMetric* usa abordagens de Inteligência Artificial para ter acessos ao estilo e ao conteúdo.

Em 1998, Larkey (1998), fez o primeiro trabalho da área baseado em técnicas de categorização de textos, além de também usar técnicas de métodos de regressão linear e características de complexidade de textos. Também neste ano, os autores de *E-Rater* criaram o *C-Rater* (BURSTEIN, LEACOCK e SWARTZ, 2000), que ao contrário do primeiro focava na análise do conteúdo ao invés da análise do estilo. E em 2001, criaram o sistema *Criterion* (BURSTEINL, CHODOROW e LEACOCK, 2003), que era um sistema web que usava o *E-Rater* para avaliar as respostas que lhe eram submetidas.

Em 1999, CHRISTIE (1999) apresentou SEAR (*Schema Extract Analyze and Report*) que é baseado em técnicas de reconhecimentos de padrões para avaliação de forma e conteúdo.

A partir do ano 2000, surgiram vários sistemas como o IEMS (MING, MIKHAILOV e LAY KUAN, 2000) que se baseava na técnica Indextron (MIKHAILOV, 1998) e o Apex Assessor (DESSUS, LEMAIRE e VERNIER, 2000) que era baseado em LSA. CALLEAR *et al.* (2001) desenvolveu o *Automated Text Marker* (ATM), o qual seguiria uma linha de pesquisa até então não explorada, na qual o sistema procura por conceitos e suas dependências

para atribuir duas notas independentes. MASON (2002) apresentou o sistema chamado de *Paperless school free text marking engine* (PS-ME) que era baseado em técnicas de linguagem natural e MITCHELL (2002) desenvolveu Automark, que é um sistema que emprega técnicas de processamento de linguagem natural para fazer uma busca textual inteligente de acordo com um esquema de marcação de respostas pré-definido.

Em 2003, dois novos sistemas foram apresentados à comunidade acadêmica: *Auto-marking*, baseado em técnicas de processamento de linguagens natural e reconhecimento de padrões, e o CarmelTC (ROSÉ *et al.*, 2003), que avalia a escrita dos alunos usando métodos de classificação como Naive Bayes e métodos de aprendizado de máquina.

2.3 ABORDAGENS PARA AVALIAÇÃO ASSISTIDA POR COMPUTADOR

Na literatura, várias técnicas têm sido aplicadas para a resolução desse problema, com resultados cada vez melhores. Estas podem ser agrupadas em seis abordagens segundo KARANIKOLAS (2010):

- Estruturas superficiais
- Reconhecimento de padrões
- Categorização de texto
- Classificação Bayesiana
- LSA
- Abordagens Combinadas

Cada abordagem será melhor detalhada a seguir.

A abordagem baseada em estruturas superficiais não considera o conteúdo dos textos a serem avaliados. O sistema mais representativo para essa abordagem foi o *Project Essay Grader* (PEG) com medidas de variáveis intrínsecas: fluência, gramática, pontuação e outros.

A abordagem de reconhecimento de padrões para a avaliação automática de questões discursivas é baseada no reconhecimento de palavras no documento. O sistema mais representativo para essa abordagem é o *Indextron*.

A abordagem de categorização de texto utiliza algum tipo de classificador binário para diferenciar a resposta boa da resposta ruim. Essa resposta do classificador é usada para classificar o ensaio. Estes classificadores se guiam pela análise das ocorrências de certas

palavras nos documentos. Desta categoria, como principal representante têm-se o TCT (LARKEY, 1998) que considera conteúdo e estilo.

A abordagem de classificação Bayesiana possui como o sistema mais representativo dessa categoria o BETSY(RUDNER e LIANG, 2002), que usa características relacionadas ao conteúdo, estilo e outras. No entanto, requere uma grande quantidade de número de ensaios de treinamento.

Além disso, existem abordagens mais avançadas que combinam inteligência artificial e processamento de linguagem natural. O sistema mais representativo é o *E-rater* que considera organização, estrutura do texto e conteúdo para avaliar os ensaios.

Finalmente tem-se a abordagem baseada em LSA que é uma análise baseada somente no conteúdo do ensaio. É uma técnica baseada em matrizes que descobre relacionamentos ocultos (latentes) entre palavras, palavras e parágrafos e entre parágrafos.

2.4 PRINCIPAIS SISTEMAS

Agora serão apresentados os sistemas mais representativos da área de Avaliação Assistida por Computador, sendo que existem sistemas que consideram apenas o estilo para avaliar o ensaio, enquanto outras consideram apenas o conteúdo, existindo também abordagens que consideram ambas as características.

2.4.1 Project Essay Grader

O Project Essay Grader baseia-se na análise do estilo de características linguísticas superficiais em um bloco de texto. Essa análise consiste em buscar várias características numéricas no texto (chamadas de "*proxes*"), as quais representam características intrínsecas mais abstratas ("*trins*") que seriam analisadas por um avaliador humano e podem ser diretamente medidas pelo computador. Quadro 2-1 apresenta alguns exemplos de *proxes* com seus respectivos *trins*

Quadro 2-1 Exemplos de proxes e trins usadas pelo Project Essay Grader

Proxes	**Trins**
Tamanho do texto.	Representa a fluência o aluno.
Número de preposições, pronomes relativos e outras partes da	Indicadores da complexidade da estrutura dos períodos.

linguagem.	
Variação no tamanho das palavras.	Indicam a riqueza de vocabulário (já que palavras maiores geralmente são menos incomuns).

Para avaliar as respostas dos alunos, o *Project Essay Grader* utiliza as notas de respostas já previamente avaliadas, juntamente com seus respectivos *proxes* para calcular os coeficientes para a equação de regressão que será usada para calcular a nota final do aluno. Esse projeto atingiu como seu melhor resultado, uma correlação de 87% para um corpus de mais de 500 ensaios.

2.4.2 *Educational Testing Service I* (ETS I)

O ETS I (WHITTINGTON e HUNT, 1999) utiliza técnicas léxico-semânticas para construir um sistema de avaliação baseado em pequenos conjuntos de dados. Este sistema guarda as informações léxicas e semânticas dos dados de treinamento, que servem de base para a avaliação dos textos submetidos a esse método. Esses dados de treinamento são processados pelo software *Microsoft Natural Language Processing* (MsNLP). Após esse processamento os sufixos são removidos manualmente, assim como várias *stop words* pré-listadas. Após esse processamento, as regras gramaticais também eram construídas manualmente para cada categoria de resposta.

Com a utilização dessa metodologia o sistema supracitado conseguiu atingir uma acurácia de 80% durante a avaliação do corpus de teste e chegou a 90% quando o mesmo avalia juntamente os corpora de teste e treinamento.

O sistema tem grandes limitações no que tange o número de palavras suportado (trabalha apenas com textos contendo 15 a 20 palavras) e a quantidade de trabalho manual envolvido. No entanto, a respeito da grande quantidade de trabalho manual envolvido, os autores argumentam que apesar do custo, o tempo gasto compensa.

2.4.3 *Intelligent Essay Assessor* (IEA)

O *Intelligent Essay Assessor* (IEA) (HEARST, 2000) é baseado em LSA e seu principal objetivo é avaliar o conhecimento contido no ensaio. Dentre suas vantagens, observa-se a independência de linguagem utilizada no sistema, com a restrição de que não pode processar estruturas morfológicas de linguagem muito complexas; o sistema em questão requer um treinamento inicial, mas esse não necessita ser supervisionado; a única entrada

necessária é o conjunto de textos de referência sobre o tópico a ser avaliado; outra vantagem desse sistema é verificação de ensaios anômalos, o qual tem a função de alertar o professor quando um ensaio é muito diferente dos demais, não havendo assim a possibilidade do sistema avaliá-lo de maneira confiável. Neste caso, o professor deve avaliar o referido ensaio, pois o aluno pode estar tendo dificuldades ou então pode estar tentando fazer algum tipo de "trapaça".

O IEA utiliza a ponderação do inverso da frequência do termo entre documentos; o cosseno como medida de similaridade e nenhuma técnica de processamento de linguagem natural como a remoção de *stop words*. Ressalte-se que o referido sistema não leva em consideração a ordem das palavras, sendo assim não consegue interpretar significados em que a ordem das palavras é um fator discriminante.

Os módulos que formam o sistema são:

i) Módulo de conteúdo: É módulo mais importante, utiliza LSA para extrair a pontuação do ensaio;

ii) Módulo de Mecânica: pontuação e ortografia são analisadas para pontuar a mecânica do ensaio.

iii) Módulo de Estilo: leva em consideração a coerência do ensaio, que é medida também utilizando LSA; é avaliada também a gramática do ensaio (é medida através semelhança das estruturas gramaticais entre o modelo e a sentença a ser avaliada).

O sistema IEA possui resultado bom nos assuntos de ciências, estudos sociais, medicina e negócios. Dentre os resultados publicados, destacam-se os resultados da aplicação desse sistema em textos das disciplinas de psicologia, medicina e história, alcançando uma acurácia de 80% a 90%, em comparação às notas do professor. Além disso, outros testes conduzidos com ensaios do teste norte americano GMAT (*Graduate Management Admission Test*), alcançando 85% a 91% de acurácia em comparação aos avaliadores humanos.

2.4.4 E-Rater *(Electronic Essay Rater)*

O *E-rater* (BURSTEIN *et al.*, 1998) foi uma evolução do sistema ETS I, e em 1999 tornou-se o segundo sistema a ser usado como avaliador no exame norte americano GMAT (*Graduate Management Admission Test*) e tem como objetivo gerar uma pontuação baseada

na organização, estrutura sentencial e conteúdo do ensaio. É um sistema relativamente complexo e requer mais treinamento em comparação a muitos outros sistemas disponíveis (pelo menos 200 textos sobre o assunto da questão) e também pode ser usado para avaliar a escrita de falantes não nativos da língua inglesa. (BURSTEIN e CHODOROW, 1999)

O referido sistema baseia-se na combinação de técnicas estatísticas e de processamento de linguagem natural para extrair características linguísticas (uso de vocabulário, variedade sintática, entre outras) dos ensaios a serem avaliados. A avaliação destes é feita baseada em um banco de dados contendo um conjunto de ensaios avaliados por especialistas humanos. Nesta avaliação, para um ensaio receber uma nota maior este precisará se manter dentro do assunto da questão; possuir uma estrutura de argumentação bem organizada e coerente; e mostrar uma variedade de palavras.

Dentre as limitações desse sistema, tem-se: a impossibilidade de avaliar textos com um número muito pequeno de palavras (BURSTEIN, LEACOCK e SWARTZ, 2000), a possibilidade de o sistema ser enganado por um texto que contenha uma escrita gramaticalmente correta, mas que tenha um conteúdo sem significado. (VALENTI, NERI e CUCCHIARELLI, 2003).

O *E-rater*, é formado por cinco módulos. Os três primeiros módulos identificam as características que podem ser usadas como critério guia para variedade sintática, organização de ideias e uso de vocabulário de um ensaio. Um quarto módulo independente é utilizado para selecionar e pesar variáveis preditivas para a avaliação do ensaio. Finalmente o último módulo é utilizado para calcular a nota final.

Como melhor resultado, o *E-rater* foi treinado com 270 respostas que já foram manualmente avaliadas por avaliadores humanos treinados. Mais de 750000 ensaios GMAT (*Graduate Management Admission Test*) foram avaliados, comparando as notas dadas pelos humanos e as notas dadas pelo *E-rater* dentre 15 questões, os resultados empíricos ficaram entre 87% a 94%.

2.4.5 IntelliMetric

O IntelliMetric (VANTAGE LEARNING TECH, 2000) é um sistema comercial criado entre 1988 e 1997, com o foco na tentativa de simular um avaliador humano através da avaliação do conteúdo, estilo e da organização e convenções de cada resposta.

Este sistema necessita de uma fase inicial de treinamento com respostas já previamente avaliadas por avaliadores humanos.

O IntelliMetric é capaz de considerar centenas características do texto para avaliar um texto, no entanto ele seleciona o mais apropriado para o assunto em estudo. Dentre essas características destaca-se o foco e a unidade do texto (que indicam o propósito e a ideia principal do texto); a organização e a estrutura (que indicam a lógica do discurso) ou as convenções (que indicam a conformidade com as regras da linguagem inglesa).

Segundo VANTAGE LEARNING TECH (2000), o sistema IntelliMetric utiliza outros sistemas proprietários para executar a avaliação automática como o *CogniSearch* e o *Quantum Reasoning Technologies*.

Esse sistema foi extensivamente usado em escolas, universidades e empresas. Dentre seus resultados, tem-se a avaliação de 594 textos escritos por estudantes de 11 anos, usando 100 respostas para treinamento, conseguindo uma acurácia de 98%. Também foi utilizado para avaliar textos em Hebreu, atingindo uma correlação de 84%.

2.4.6 Sistema de Larkey (1998)

Esse sistema é baseado em técnicas de categorização para avaliar as questões discursivas como “bons” ou “ruins” considerando conteúdo e estilo (LARKEY, 1998). Nesse sistema, a avaliação é feita por três módulos, onde a nota final é dada por um desses ou pela combinação de dois ou mais. Os módulos são:

i) **Classificadores Bayesianos:** para cada documento é atribuída uma probabilidade de pertencer a uma categoria de documento previamente especificada. Para isso, dois passos são executados: o primeiro é a seleção de características que remove as *stop words*, faz o *stemming* (remoção de afixos das palavras) e procura pelas características mais representativas usando redes Bayesianas. Por fim, faz o treinamento usando o modelo binário onde 0 significa que a característica não está no texto e 1 o oposto.

ii) **Procura pelos ensaios de referência:** O sistema procura os sistemas de referência mais similares ao ensaio que está sendo avaliado.

iii) **Cálculo de características de complexidades:** O sistema calcula automaticamente características do texto em avaliação. Dentre essas características, destaca-se o número de

caracteres no documento, número de palavras diferentes no documento, tamanho médio de cada oração e número de palavras com mais de sete caracteres.

A avaliação final do ensaio é oriunda da regressão linear feita com os valores resultantes dos módulos apresentados acima.

Esse sistema foi aplicado em questões de estudos sociais, física e direito. Nessas questões o sistema alcançou uma acurácias de 60% e 100% (considerando uma tolerância de erro de 1 ponto), quando todos os critérios de avaliação foram usados. Para avaliação de questões de opiniões gerais o resultado foi uma acurácia de 55% e 97% (considerando uma tolerância de erro de 1 ponto). A correlação obtida foi de 80% e no caso das questões de opiniões gerais foi de 88%.

2.4.7 *C-Rater* (*Conceptual Rater*)

C-Rater é um protótipo baseado em processamento de linguagem natural que tem o objetivo de avaliar pequenas respostas associadas a questões de conteúdo, como aquelas que aparecem na seção de revisão dos livros didáticos. *C-rater* adota várias técnicas de processamento de linguagem natural, e é muito similar ao *E-rater*, sendo que a principal diferença entre ambos é que o *E-rater* foca no estilo e o *C-rater* foca no conteúdo, ou seja, o *C-rater* identifica se a resposta contém a informação específica necessária para estar correta. Além disso, o *C-rater* necessita de um conjunto de dados de treinamento menor em comparação ao *E-rater* e é capaz de tolerar palavras sinônimas, erros de ortografia e variações sintáticas e flexionais.

O principal objetivo do *C-Rater* é distinguir se o a resposta do estudante está certa ou errada baseada no conteúdo da mesma. Trata-se de um sistema de avaliação totalmente automática, excetuando apenas o a construção da modelo de referência, o qual necessita de intervenção humana.

C-rater foi testado em pequenas tarefas formativas como por exemplo, pequenas perguntas localizadas no fim de cada capítulo de livros escolares. Quando usado em estudos de menor escala com um ambiente virtual de aprendizado, este alcançou mais de 80% de acurácia (LEACOCK, 2004). Quando usado em avaliações de grande escala, como a avaliação de 170 mil pequenas respostas a 19 questões de leitura e compreensão de texto e 5 questões de álgebra, o resultado foi 85% de acurácia.

Segundo LEACOCK (2004), os maiores problemas do *C-rater* são: impossibilidade de avaliar respostas cujo significado depende da flexão do verbo, devido à utilização do *stemming* por este; impossibilidade de lidar com respostas que contenham uma citação; alguns erros de ortografia não são corretamente corrigidos e o sistema não consegue lidar com expressões idiomáticas.

2.4.8 SEAR (*Schema Extract Analyze and Report*)

O SEAR foi desenvolvido por CHRISTIE (1999) e tem por objetivo avaliar o conteúdo e o estilo do texto.

No caso da avaliação do estilo são necessários quatro passos: i) pré-determinar as métricas candidatas; ii) utilizar textos já avaliados por avaliadores humanos como modelo; iii) calibrar o sistema a fim de encontrar uma acurácia aceitável em relação às avaliações dos avaliadores humanos; iv) Processar a resposta do aluno, procurando pelas métricas definidas e aplicando uma ponderação para cada métrica para computar a nota final do aluno.

Para o conteúdo, não são necessários treinamento e calibração, mas o professor precisa criar alguns esquemas de referência. Ressalte-se ainda que o sistema usa técnicas de extração de informação para preencher os esquemas dos estudantes com os dados dos mesmos para compará-los com as referencias.

2.4.9 IEMS (*Intelligent Essay Marking System*)

O IEMS (MING, MIKHAILOV e LAY KUAN, 2000) é baseado na rede neural indexadora de padrões Indextron (MIKHAILOV, 1998), que faz o reconhecimento de padrões e nesses casos os padrões são as palavras dos textos. Pode ser utilizado como ferramenta de avaliação; como ferramenta de diagnóstico e para propósitos de tutoria em muitos problemas de conteúdo. Além disso, necessita de um treinamento demorado e não incremental. Esse sistema foi aplicado em disciplinas como biologia, psicologia, história ou anatomia. Dentre os resultados desse sistema, foi dado um texto base de 800 palavras e em seguida foram avaliados 85 resumos de estudantes universitários com no máximo 180 palavras sobre o texto base, obtendo-se uma acurácia de 80%.

2.4.10 Apex Assessor

Esse sistema é integrado dentro do Ambiente Virtual de Aprendizagem Apex (DESSUS, LEMAIRE e VERNIER, 2000) e é usado para fazer a avaliação de aprendizagem no fim de cada capítulo de um livro didático. Por ser um sistema baseado em LSA, este necessita apenas textos de treinamento que não precisam ter sido avaliados previamente.

O Apex Acessor é dividido em três módulos principais:

1) Módulo de avaliação baseado em conteúdo: compara a representação LSA da resposta do aluno com o modelo LSA;

2) Módulo de avaliação detalhado: para cada parágrafo no texto do aluno, é mostrado qual o conceito mais similar;

3) Módulo de coerência: mede a distância semântica entre os períodos como o LSA, sendo que se a distância entre duas sentenças consecutivas estiver abaixo de um limite pré-estabelecido, o sistema considera essa situação como quebra de coerência. Nesses casos, o aluno é advertido.

Ressalte-se que foram incluídos conteúdos textuais não técnicos nesse sistema, para que o mesmo pudesse também lidar com termos de fora do domínio do problema que viessem a aparecer nas respostas dos alunos.

Para análise do desempenho do sistema, utilizou-se 31 ensaios de um curso de pós-graduação em sociologia da educação que foram avaliados por professores e compararam as notas deste com a nota dada pelo sistema. Como resultado, obteve-se uma correlação de 59% com $p<0,001$.

Vale à pena mencionar que um dos problemas encontrados no sistema, é que algumas respostas muito curtas conseguiam alcançar notas altas. Para resolver isso, os autores do sistema modificaram o mesmo para ser mais rígido com texto que tivessem menos de 300 palavras.

2.4.11 PS-ME (*Paperless school free text marking engine*)

O PS-ME (MASON e GROVE-STEPHENSON, 2002) foi um sistema desenvolvido para ser integrado a um ambiente virtual de aprendizagem. Essa ferramenta introduziu uma novidade dentro do campo de pesquisa, que foi a existência de um texto base de referência

negativa contendo um conjunto de afirmações falsas geralmente compostas pelos erros mais comuns cometidos pelos alunos.

Esse sistema aplica técnicas de processamento de linguagem natural para avaliar as respostas dos alunos da seguinte forma: o aluno submete o ensaio para o servidor juntamente com informações sobre a tarefa que lhe foi pedida, a fim de identificar corretamente o texto que servirá como base para a avaliação. Devido ao seu grande consumo de processamento, o PS-ME não calcula as notas dos alunos em tempo real. A nota final do aluno é resultante da comparação da resposta deste com todas as respostas base relevantes ao assunto em questão, sendo que cada comparação ganha um peso específico na nota do aluno, no caso do texto base negativo, esse peso é negativo.

O processo de funcionamento do PS-ME pode ser resumido como: primeiramente seleciona-se um texto base de fontes como livros, enciclopédias ou sites relevantes. Em seguida, obtém-se um exemplo de uma resposta já avaliada por um especialista. Depois, utiliza-se esse exemplo para ser avaliado automaticamente, e em seguida tenta-se combinar os parâmetros até que se encontre a melhor configuração para que a nota automática se aproxime o máximo possível da nota do especialista. Após isso, o sistema executa a avaliação automática no texto do aluno e envia os resultados ao servidor.

É importante destacar que O PS-ME não retorna somente as notas para os alunos, mas retorna também um *feedback* formativo, sobre as diferentes áreas do assunto em questão.

Dentre as principais dificuldades desse sistema, tem-se a dificuldade pra conseguir textos base; e também erros de gramática e ortografia podem afetar bastante os resultados.

Esse sistema foi aplicado no exame GCSE (*General Certificate of Secondary Education*) e na área comercial tem sido usado por publicadoras, porém até o momento ainda não foram encontrados publicações do desempenho desse sistema.

2.4.12 Automark

O Automark (MITCHELL *et al.*, 2002) foi concebido primeiramente como projeto acadêmico, mas em 2002 começou a ser usado comercialmente dentro do ambiente virtual de aprendizagem *ExamOnline*. O objetivo desse sistema é avaliar o estilo e o conteúdo da resposta do aluno para dizer para dizer se o texto é aceitável ou não de acordo com o critério especificado pelo professor ao sistema. A ferramenta utiliza técnicas de extração da

informação e de processamento de linguagem natural para ignorar alguns erros de ortografia, sintática e semântica que não devem ser levados em consideração.

O Automark procura por conteúdos específicos dentro das respostas, os quais são especificados na forma de vários modelos esquemas de marcação. Cada modelo representa uma forma válida ou então uma resposta inválida.

O processo de avaliação no Automark ocorre da seguinte forma: primeiramente a resposta do aluno é pré-processada para padronizá-la em termos de pontuação e ortografia. Então um analisador de orações identifica os constituintes sintáticos principais do texto juntamente com o relacionamento entre eles. Em seguida, o módulo identificador de padrões procura por estruturas equivalentes nos modelos de esquemas de marcação e nos constituintes sintáticos do texto do aluno e, finalmente, o modulo de *feedback* processa o resultado do módulo anterior.

O Automark foi testado com respostas de alunos de 11 anos e a correlação alcançada variou entre 93% a 96%.

2.4.13 Auto-marking

O Auto-marking (SUKKARIEH, PULMAN e RAIKES, 2003) é uma ferramenta desenvolvida com o objetivo de avaliar exames de pequeno porte, onde cada exercício recebia uma nota de 0 até 2. Onde 0 significava “incorreto”, 1 significava “parcialmente correto” e 2 significava “correto e completo”. Esse sistema é baseado em técnicas de processamento linguagem natural e de reconhecimento de padrões.

O Auto-marking é formado pelos seguintes módulos:

i) Customização e módulo de processamento superficial: primeiramente o sistema usa um modelo oculto de Markov para etiquetação da língua falada e uma maquina de estados para remover os substantivos e verbos das frases. Em alguns casos, é necessária intervenção manual.

ii) Módulo reconhecedor de padrões: é muito similar ao módulo usado no Automark (MITCHELL *et al.*, 2002), ou seja o especialista humano tem que projetar os padrões de extração de informação que servirão de base para a avaliação das respostas dos alunos.

iii) Módulo de avaliação: usando regras definidas pelo especialista no módulo anterior, o sistema atribui a cada resposta a sua respectiva classe.

Esse sistema foi aplicado em respostas do exame GCSE (*General Certificate of Secondary Education*) como uma acurácia de 88% em comparação as respostas do professor.

Dentre os problemas do sistema, destaca-se o fato do sistema ser aconselhado pelos autores do mesmo a não ser usado em situações que envolvam opiniões subjetivas gerais. Outro problema é que o sistema não consegue lidar com textos que contenham inferências com informação contraditória ou inexistente.

2.4.14 CarmelTC

A ferramenta CarmelTC (ROSÉ *et al.*, 2003) é parte integrante do ambiente virtual de aprendizado chamado Carmel. O funcionamento deste sistema é uma combinação de métodos de classificação baseada em aprendizado de máquina e a classificação baseada em Rainbow Naive Bayes (MCCALLUM e NIGAM, 1998).

O CarmelTC, além de dar uma nota ao estudante, pode ser usado para encontrar quais conceitos são usados corretamente no texto do aluno.

O processo de avaliação dos alunos ocorre da seguinte forma: primeiramente o sistema quebra o texto em orações. Após isso, usa-se a rede bayesiana para buscar a provável característica correta que representa cada oração, a fim de gerar um vetor indicando a presença ou ausência de cada característica correta. Finalmente, o terceiro passo induz as regras para identificar classes de orações baseadas nesses vetores como o algoritmo de aprendizado em árvore ID3 (QUINLAN, 1993).

O algoritmo foi testado com 126 ensaios de física, e os resultados foram de 90% de acurácia.

2.4.15 BETSY (*Bayesian Essay Test Scoring sYstem*)

O sistema BETSY (RUDNER e LIANG, 2002) tem como objetivo determinar para uma determinada resposta discursiva em qual classificação esta se enquadra dentro de uma escala nominal composta por quatro itens (extensiva, essencial, parcial e insatisfatória) considerando estilo e conteúdo.

O sistema BETSY é baseado em redes bayesianas (Naive Bayes). Sendo que nesse sistema há a possibilidade de se escolher qual modelo estatístico será usado no processo de avaliação das respostas. Os modelos disponíveis para escolha são: O modelo de Bernoulli e o modelo multivariado de Bernouilli. Sendo que também possui a possibilidade de usar *stemming* e remoção de *stop words* para melhorar os resultados do método.

O sistema foi usado para avaliar testes de biologia no colégio Maryland, onde este foi calibrado com 462 ensaios com duas categorias. Depois de feita a calibração, o sistema foi aplicado em 80 ensaios já previamente avaliados por especialistas, onde cada categoria deveria ter 40 ensaios avaliados. Dentro deste cenários a ferramenta obteve uma acurácia de 80%, usando o modelo de Bernoulli, e uma acurácia de 74% usando o modelo multivariado de Bernoulli.

2.5 RESUMO COMPARATIVO

O Quadro 2-2 abaixo apresenta um resumo comparativo de todos os sistemas abordados anteriormente, onde estão destacados, para cada sistema, a principal técnica utilizada por este; os melhores desempenhos alcançados; e a descrição do corpus utilizados para alcançar os melhores resultados.

Quadro 2-2 Exemplos de proxes e trins usadas pelo Project Essay Grader

Sistema	**Técnica**	**Desempenho**	**Corpus**
Project Essay Grader	Análise de características linguísticas superficiais	87% (correlação)	Corpus de 495 e 599 ensaios
ETS I	Técnicas léxico-semânticas	90% de acurácia.	Não especificado, mais é o mesmo corpora de teste e treinamento.
Intelligent Essay Assessor	LSA	Acurácia de 80% a 90%	Textos das disciplinas de psicologia, medicina e história, GMAT.
E-Rater	Técnicas estatísticas e Processamento de Linguagem Natural.	Acurácia de aproximadamente 87% a 94%.	Mais de 750000 ensaios GMAT
IntelliMetric	-	Acurácia de 98% e correlação de 84%.	594 textos de estudos sociais, física e direito escritos por estudantes de 11 anos.
Larkey (1998)	Técnicas de categorização	Correlação de 80%.	Questões de estudos sociais, física, direito e opiniões gerais
C-Rater	Técnicas de processamento de linguagem natural	85% de acurácia.	170 mil pequenas respostas localizadas no fim de cada capítulo de livros escolares

SEAR	Técnicas de extração de informação	-	-
IEMS	Rede neural indexadora de padrões Indextron	Acurácia de 80%.	85 resumos de estudantes universitários sobre o texto base com no máximo 180 palavras.
Apex Assessor	LSA	Correlação de 59%	31 ensaios de um curso de pós-graduação em sociologia da educação.
PS-ME	Técnicas de processamento de linguagem natural	-	GCSE (*General Certificate of Secondary Education*)
Automark	Técnicas de extração da informação e de processamento de linguagem natural	Correlação de 93% a 96%.	Respostas de alunos de 11 anos
Auto-marking	Técnicas de processamento linguagem natural e de reconhecimento de padrões	Acurácia de 88%.	GCSE (General Certificate of Secondary Education)
CarmelTC	Métodos de classificação baseada em aprendizado de máquina e Rainbow Naive Bayes.	Acurácia de 90%.	126 ensaios de física.
BETSY	Naive Bayes	Acurácia de 80%	80 testes de biologia no colégio Maryland.

Por fim, analisando o Quadro 2-2 destaca-se que dentre os 15 sistemas analisados, apenas 2 são baseados na técnica LSA e estes alcançaram como resultado uma acurácia de 80% a 90% e uma correlação de 59%.

3. FUNDAMENTAÇÃO TEÓRICA

3.1 LSA (*Latent Semantic Analysis*)

Análise Semântica Latente, LSA (DEERWESTER *et al.*, 1990) é uma técnica estatístico matemática para extração e inferência de relações de contexto em passagens do discurso (LANDAUER, FOLTZ e LAHAM, 1998), que possibilita a comparação de dois trechos de texto usando uma medida de similaridade (KANEJIYA, KUMAR e PRASAD, 2003). Em outras palavras, essa técnica objetiva simular a experiência que as pessoas têm ao analisar um texto em sua língua.

A representação de significado de palavras e sentenças que pode ser obtido pelo LSA têm sido capaz de simular uma variedade de habilidades que as pessoas possuem como, por exemplo, reconhecimento de vocabulário, categorização de palavras, compreensão de discurso, etc. (LANDAUER, FOLTZ e LAHAM, 1998).

Para se usar o LSA, o primeiro passo é representar a coleção de textos a serem analisados em uma matriz termo-documento. Esta por sua vez é uma matriz que contém como elementos as frequências do número de vezes que cada palavra aparece em um determinado trecho de texto (que pode ser uma oração, um parágrafo ou até mesmo um documento inteiro), sendo que as linhas representam as palavras e colunas representam os documentos (ou trechos de texto), conforme pode se observar no exemplo apresentado nos Quadros 3-1 e 3-2 onde um trecho de texto (colunas da tabela) é considerado como um documento. Entretanto, percebe-se que neste exemplo algumas palavras foram omitidas na transição para a matriz termo documento. Esta omissão é opcional e as palavras que foram consideradas para a construção desta tabela-termo documento estão destacadas em itálico.

Quadro 3-1 Exemplos de dados textuais: títulos de memorandos técnicos. Adaptado de LANDAUER, FOLTZ e LAHAM (1998).

C1: *Human* machine *interface* for ABC *computer* applications

C2: A *survey* of *user* opinion of *computer system response time*

C3: The *EPS user interface* management *system*

C4: *System* and *human system* engineering testing of *EPS*

C5: Relation of *user* perceived *computer* response *time* to error measurement

M1: The generation of random, binary, ordered *trees*

M2: The intersection *graph* of paths in *trees*

M3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering

Quadro 3-2 Tabela termo-documento formada pelos dados textuais (Matriz {X})

	C1	C2	C3	C4	C5	M1	M2	M3	M4
human	1	0	1	0	0	0	0	0	0
interface	1	1	0	0	0	0	0	0	0
computer	1	1	1	0	1	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

O próximo passo é a aplicação da técnica de decomposição de valores singulares, do inglês *Singular Value Decomposition* (SVD). Esta técnica consiste em decompor uma matriz singular no produto de outras três. Sendo que uma dessas matrizes descreve as entidades das linhas da matriz original como vetores de valores de fatores ortogonais derivados (denominada de matriz $\{T_0\}$); outra matriz descreve as entidades das colunas da matriz original como vetores de fatores ortogonais derivados da mesma maneira (denominada de matriz $\{D_0\}$); e a terceira é uma matriz diagonal contendo os valores de escala (denominada

de matriz {S_0}) de tal forma que quando as três matrizes são multiplicadas, a matriz original é reconstruída (LANDAUER, FOLTZ e LAHAM, 1998), ou seja, considerando a matriz original como {X}, esta pode ser decomposta da seguinte forma: {X} = {T_0} * {S_0} * {D_0}'. Uma ilustração dessa decomposição pode ser vista logo abaixo na Figura 3-1.

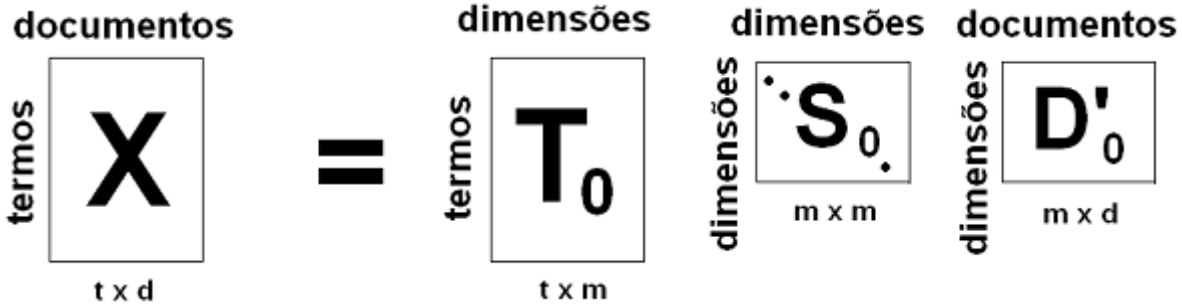


Figura 3-1Ilustração de funcionamento da SVD(MORGADO, 2010)

Para ilustrar melhor o processo de SVD, serão mostradas logo a seguir as matrizes {T_0}, {S_0}, {D_0}' resultantes da decomposição da matriz {X} nos Quadros 3-3, 3-4 e 3-5, respectivamente.

Quadro 3-3 Matriz {T_0}

0,22	-0,11	0,29	-0,41	-0,11	-0,34	0,52	-0,06	-0,41
0,2	-0,07	0,14	-0,55	0,28	0,5	-0,07	-0,01	-0,11
0,24	0,04	-0,16	-0,59	-0,11	-0,25	-0,3	0,06	0,49
0,4	0,06	-0,34	0,1	0,33	0,38	0	0	0,01
0,64	-0,17	0,36	0,33	-0,16	-0,21	-0,17	0,03	0,27
0,27	0,11	-0,43	0,07	0,08	-0,17	0,28	-0,02	-0,05
0,27	0,11	-0,43	0,07	0,08	-0,17	0,28	-0,02	-0,05
0,3	-0,14	0,33	0,19	0,11	0,27	0,03	-0,02	-0,17
0,21	0,27	-0,18	-0,03	-0,54	0,08	-0,47	-0,04	-0,58
0,01	0,49	0,23	0,03	0,59	-0,39	-0,29	-0,25	-0,23
0,04	0,62	0,22	0,00	-0,07	0,11	0,16	-0,68	0,23
0,03	0,45	0,14	-0,01	-0,30	0,28	0,34	0,68	0,18

Quadro 3-4 Matriz {S_0}

3,34	0	0	0	0	0	0	0	0
0	2,54	0	0	0	0	0	0	0
0	0	2,35	0	0	0	0	0	0
0	0	0	1,64	0	0	0	0	0
0	0	0	0	1,5	0	0	0	0

0	0	0	0	0	1,31	0	0	0
0	0	0	0	0	0	1,31	0	0
0	0	0	0	0	0	0	0,56	0
0	0	0	0	0	0	0	0	0,36

Quadro 3-5 Matriz {D_0}'

0,2	0,61	0,46	0,54	0,28	0	0,01	0,02	0,08
-0,06	0,17	-0,13	-0,23	0,11	0,19	0,44	0,62	0,53
0,11	-0,5	0,21	0,57	-0,51	0,1	0,19	0,25	0,08
-0,95	-0,03	0,04	0,27	0,15	0,02	0,02	0,01	-0,03
0,05	-0,21	0,38	-0,21	0,33	0,39	0,35	0,15	-0,6
-0,08	-0,26	0,72	-0,37	0,03	-0,3	-0,21	0	0,36
0,18	-0,43	-0,24	0,26	0,67	-0,34	-0,34	0,25	0,04
-0,01	0,05	0,01	-0,02	-0,06	0,45	-0,76	0,45	-0,07
-0,06	0,24	0,02	-0,08	-0,26	-0,62	0,02	0,52	-0,45

Após a decomposição da matriz {X}, é possível reconstruí-la usando um número menor de valores de escala da matriz {S_0}. Esse número de valores da matriz {S_0} que são usados para reconstruir a matriz {X} é denominado de "dimensão". Sendo que os valores da matriz diagonal {S_0} estão ordenados de cima para baixo, de modo que para usar uma dimensão 2, por exemplo, deve-se manter os dois primeiros elementos da diagonal da matriz {S_0} e substituir os demais elementos da diagonal por 0. Essa nova dimensão é usualmente representada pela letra *K* e a redução da dimensão da matriz {S_0} é geralmente citada na literatura como redução ao espaço semântico.

Para melhor exemplificar será mostrado a seguir um exemplo número para *k*=2 usando as matrizes {X}, {T_0}, {S_0} e {D'$_0$}.

Para a redução do espaço semântico a duas dimensões devem ser consideradas apenas os dois primeiros elementos da diagonal da matriz {S_0}, conforme pode ser visto no Quadro 3-6. Após isso a matriz {X} deve ser reconstruída através das multiplicações das matrizes {T_0}, {S_0} e {D'$_0$}, o que resultará numa nova matriz {X}, a qual pode ser vista no Quadro 3-7.

Quadro 3-6 Matriz {S_0} para k=2.

3,34	0	0	0	0	0	0	0	0
0	2,54	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Quadro 3-7 Nova matriz {X} para k=2.

0,19	0,47	0,38	0,26	0,29	-0,03	-0,06	-0,08	-0,03
0,23	0,59	0,44	0,28	0,36	0,01	0,04	0,05	0,09
0,46	1,17	0,89	0,59	0,72	0	0	0	0,10
0,39	1,00	0,76	0,5	0,62	0	0,01	0,01	0,09
0,46	1,14	0,92	0,63	0,71	-0,06	-0,14	-0,19	-0,06
0,26	0,69	0,5	0,32	0,42	0,02	0,06	0,08	0,13
0,26	0,69	0,5	0,32	0,42	0,02	0,06	0,08	0,13
0,21	0,51	0,43	0,30	0,32	-0,05	-0,12	-0,16	-0,09
0,17	0,52	0,28	0,14	0,30	0,13	0,29	0,41	0,38
-0,02	0,12	-0,15	-0,18	0,04	0,27	0,60	0,84	0,70
0	0,19	-0,16	-0,21	0,08	0,34	0,76	1,06	0,89
0	0,15	-0,11	-0,15	0,06	0,24	0,54	0,76	0,64

Vale a pena salientar que o número de dimensões usadas no LSA para a redução ao espaço semântico é geralmente selecionado de maneira empírica (LANDAUER, FOLTZ e LAHAM, 1998), o que pode ser problema em potencial, considerando que a seleção de uma dimensionalidade ótima implicaria indução correta dos relacionamentos latentes com consequente influência no resultado final.

A nova matriz é reconstruída em um espaço semântico k-dimensional denominada de {X}’ possui valores diferentes da matriz {X} original. Essa mudança de valores caracteriza o mecanismo no qual o LSA usa para fazer inferência ou indução.

Segundo BERRY *et al.* (1995) esta redução da dimensão, trazem as seguintes vantagens:

- Facilidade no tratamento computacional dos dados.
- Melhoria nas relações entre os textos através da identificação de estruturas semânticas ocultas nas relações entre palavras e textos;
- Redução do ruído;

- Variabilidade na aplicação das palavras.
- No espaço semântico ocorre uma melhor similaridade entre dois textos.

Após a redução da dimensionalidade, a matriz é refeita e a partir desta reconstrução é medida a similaridade entre os documentos, onde a similaridade igual a um indica alta similaridade e uma similaridade igual à zero indica baixa. Nesses casos os dois textos não estão relacionados.

Para medição dessa similaridade existem algumas medidas disponíveis na literatura, dentre essas se destacam cosseno, correlação de Pearson e Spearman, as quais são apresentadas no capítulo 3.5.

Em resumo, o método LSA é apenas um formalismo matemático que para ser usado na prática precisa ser alimentado com a matriz de entrada e com um conjunto de mecanismos auxiliares. Para a matriz de entrada os textos podem ser processados de diferentes formas melhorando o desempenho do método. Neste trabalho explorou-se: a remoção de stop words, e *stemming* em unigramas e bigramas. Por outro lado, como mecanismos auxiliares temos a escolha do número de dimensões; as técnicas de ponderação, que visam alterar os dados numéricos das matrizes de tal forma que as relações entre as palavras fiquem mais evidentes; e as métricas de similaridade que tem a finalidade de medir a semelhança entre os textos, onde cada similaridade verifica essa semelhança de uma forma particular.

3.2 TÉCNICAS DE PRÉ-PROCESSAMENTO

3.2.1 N-GRAMAS

N-Grama é uma sequencia de palavras na qual o *n* pode assumir valores como, por exemplo, 1, 2 ou 3, sendo nesses casos os n-grama passaria ser chamado de unigrama, bigrama e trigrama respectivamente.

Quando se analisa um bigrama ou trigrama, geralmente objetiva-se analisar a ocorrência desse grupo de palavras naquela determinada ordem, já que dois bigramas ou trigrama contendo as mesmas palavras em ordem diferente são considerados elementos distintos.

Neste trabalho, além dos unigramas também foram usados bigramas com o intuito de usar o LSA acrescido com a informação sobre da ordem das palavras no texto.

3.2.2 REMOÇÃO DE *STOP WORDS*

Stop words são palavras que contém pouca informação semântica, como por exemplo, conjunções, artigos, preposições, pronomes e verbos auxiliares. (CARVALHO, MATOS e ROCIO, 2007)

Estas palavras ocorrem com frequência em qualquer texto e representam uma fração significante em relação ao tamanho total do texto. Sendo assim a remoção destas aumenta consideravelmente a velocidade de qualquer processamento computacional de análise de textual. Segundo RIJSBERGEN (1979) *stop words* são consideradas como ruído e por isso deveria ser removida no pré-processamento de qualquer experimento de análise de texto.

Como exemplo de *stop words* podem ser citadas as palavras: “a”, “as”, “o”, “os”, “de”, “para”, “mas”, “por”, “para”, “se”, “em”, “até”, “e”, “no”, “na”, “num”, “numa”, “ou”, “cada”, “um”, “uma”.

3.2.3 STEMMING

Stemming é o processo de redução de palavras flexionadas ou derivadas para o sua raiz (SOARES, PRATI e MONARD, 2009) através da remoção dos afixos (CARVALHO, MATOS e ROCIO, 2007). Esse processo contribui para a redução da quantidade de palavras distintas, uma vez que as flexões e derivações de diferentes palavras correlacionadas são representadas pela sua raiz. Como por exemplo, as palavras “duvido”, “dúvida”, “duvidamos” e “duvidem” passariam a ser representadas pela raiz “duvid”.

3.3 TÉCNICAS DE PONDERAÇÃO

Técnicas de ponderação (ou pesagem) são técnicas que podem ser usadas para designar um nível de relevância para cada termo de uma matriz (FORONDA, 2005), ou em outras palavras essa técnica faz uma transformação na matriz de modo que aqueles termos que têm maior importância ficarão mais evidenciados.

Essas técnicas são muito usadas na área de indexação de textos, e segundo BERRY e BROWNE (1999) o objetivo de pesar termos na área de indexação é melhorar o desempenho referente à habilidade de recuperar documentos relevantes e deixar de lado a informação irrelevante. Já no contexto desse trabalho, o objetivo é verificar a influência das técnicas de ponderação na acurácia do resultado final.

De acordo com a abrangência do escopo verificado pelo método de ponderação este pode ser classificado em dois tipos: ponderações globais e ponderações locais. Estas técnicas serão mais bem descritas logo a seguir.

3.3.1 PONDERAÇÃO GLOBAL

As técnicas de ponderação global estimam o nível de relevância da palavra em todos os textos do corpus.

A seguir são mostradas algumas das ponderações globais mais utilizadas, com um respectivo exemplo numérico baseado na matriz {X} (Quadro 3-3).

- **Entropia:**

$$entropia(i,j) = 1 + \frac{\sum_{j=1}^{nº\ ensaios} P(i,j) \cdot log_2(P(i,j)}{log_2(nº\ ensaios)}$$

Onde $P(i,j) = \frac{ft(i,j)}{gf(i)}$ é a probabilidade condicional e $gf(i)$ é a freqüência global da palavra i.

Para exemplificar é mostrado um exemplo numérico no Quadro 3-8 logo abaixo, no qual é mostrada a aplicação desta ponderação global a Matriz {X}.

Quadro 3-8 Exemplo numérico da aplicação da Entropia à Matriz {X}.

0,86	0	0,90	0	0	0	0	0	0
0,86	0,91	0	0	0	0	0	0	0
0,91	0,95	0,93	0	0,92	0	0	0	0
0	0,93	0,92	0	0,91	0	0	0	0
0	0,95	0,93	1,73	0	0	0	0	0
0	0,91	0	0	0,88	0	0	0	0
0	0,91	0	0	0,88	0	0	0	0
0	0	0,90	0,86	0	0	0	0	0
0	0,91	0	0	0	0	0	0	0,86
0	0	0	0	0	0,83	0,86	0,89	0
0	0	0	0	0	0	0,86	0,89	0,89
0	0	0	0	0	0	0	0,86	0,86

- **Inverso da frequência do termo entre documentos (IDF)**:

$$idf(i,j) = 1 + log_2(\frac{n^{\underline{o}}\ ensaios}{df(i)})$$

Onde df(i) é o número de ensaios em que a palavra i aparece.

Para exemplificar é mostrado um exemplo numérico no Quadro 3-9 logo abaixo, no qual é mostrada a aplicação desta ponderação global a Matriz {X}.

Quadro 3-9 Exemplo numérico da aplicação do inverso da frequência do termo entre documentos à Matriz {X}.

3,17	0	3,17	0	0	0	0	0	0
3,17	3,17	0	0	0	0	0	0	0
2,17	2,17	2,17	0	2,17	0	0	0	0
0	2,58	2,58	0	2,58	0	0	0	0
0	2,58	2,58	5,17	0	0	0	0	0
0	3,17	0	0	3,17	0	0	0	0
0	3,17	0	0	3,17	0	0	0	0
0	0	3,17	3,17	0	0	0	0	0
0	3,17	0	0	0	0	0	0	3,17
0	0	0	0	0	2,58	2,58	2,58	0
0	0	0	0	0	0	2,58	2,58	2,58
0	0	0	0	0	0	0	3,17	3,17

- **Normal**:

$$normal(i,j) = \frac{1}{\sqrt{\sum_{j=1}^{n^{\underline{o}}\ ensaios}(local(i,j))^2}}$$

Onde local(i, j) representa uma ponderação local qualquer que deve ser usada juntamente com essa ponderação global.

Para exemplificar é mostrado um exemplo numérico no Quadro 3-10 logo abaixo (onde local(i,j) é representado pela ponderação inverso da frequência do termo entre documentos), no qual é mostrada a aplicação desta ponderação global a Matriz {X}.

Quadro 3-10 Exemplo numérico da aplicação da ponderação normal à Matriz {X}.

0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71
0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71
0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5
0,58	0,58	0,58	0,58	0,58	0,58	0,58	0,58	0,58
0,41	0,41	0,41	0,41	0,41	0,41	0,41	0,41	0,41
0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71
0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71
0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71
0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71
0,58	0,58	0,58	0,58	0,58	0,58	0,58	0,58	0,58
0,58	0,58	0,58	0,58	0,58	0,58	0,58	0,58	0,58
0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71	0,71

3.3.2 PONDERAÇÃO LOCAL

As técnicas de ponderação local estimam o nível de relevância da palavra dentro de um único texto.

A seguir são mostradas algumas das ponderações locais mais utilizadas.

- **Termo frequência**: indica a frequência da palavra num texto, onde a palavra que ocorre mais vezes tem maior peso.

$$ft(i,j) = \frac{c(i,j)}{\sum_{k=1}^{I} c(k,j)}$$

Onde $c(i,j)$ é o número de vezes que a palavra i ocorre na resposta j e I é o número total de palavras.

Para exemplificar é mostrado um exemplo numérico no Quadro 3-11 logo abaixo, no qual é mostrada a aplicação desta ponderação local a Matriz {X}.

Quadro 3-11 Exemplo numérico da aplicação da ponderação termo frequência à Matriz {X}.

0,33	0	0,2	0	0	0	0	0	0
0,33	0,14	0	0	0	0	0	0	0
0,33	0,14	0,2	0	0,25	0	0	0	0
0	0,14	0,2	0	0,25	0	0	0	0
0	0,14	0,2	0,67	0	0	0	0	0
0	0,14	0	0	0,25	0	0	0	0
0	0,14	0	0	0,25	0	0	0	0
0	0	0,2	0,33	0	0	0	0	0
0	0,14	0	0	0	0	0	0	0,33
0	0	0	0	0	1	0,5	0,33	0
0	0	0	0	0	0	0,5	0,33	0,33
0	0	0	0	0	0	0	0,33	0,33

- **Binária**: indica apenas se a palavra ocorre ou não no texto, independente da quantidade.

$$bin(i,j) = \begin{cases} 1, & se\ \ ft(i,j) > 0 \\ 0, & se\ ft(i,j) = 0 \end{cases}$$

Onde i é a coluna e j é a linha da matriz e $ft(i,j)$ é o valor do elemento na matriz termo-documento.

Para exemplificar é mostrado um exemplo numérico no Quadro 3-12 logo abaixo, no qual é mostrada a aplicação desta ponderação local a Matriz {X}.

Quadro 3-12 Exemplo numérico da aplicação da ponderação binária à Matriz {X}.

1	0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0
1	1	1	0	1	0	0	0	0
0	1	1	0	1	0	0	0	0
0	1	1	1	0	0	0	0	0
0	1	0	0	1	0	0	0	0
0	1	0	0	1	0	0	0	0
0	0	1	1	0	0	0	0	0
0	1	0	0	0	0	0	0	1
0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	1

- **Logaritmo**: aplica-se a função logarítmica a cada elemento da matriz termo documento.

$$logaritmo(i,j) = log_2(ft(i,j) + 1)$$

Onde i é a coluna e j é a linha da matriz e $ft(i,j)$ é o valor do elemento na matriz termo-documento.

Para exemplificar é mostrado um exemplo numérico no Quadro 3-13 logo abaixo, no qual é mostrada a aplicação desta ponderação local a Matriz {X}.

Quadro 3-13 Exemplo numérico da aplicação do logaritmo à Matriz {X}.

0,42	0	0,26	0	0	0	0	0	0
0,42	0,19	0	0	0	0	0	0	0
0,42	0,19	0,26	0	0,32	0	0	0	0
0	0,19	0,26	0	0,32	0	0	0	0
0	0,19	0,26	1,47	0	0	0	0	0
0	0,19	0	0	0,32	0	0	0	0
0	0,19	0	0	0,32	0	0	0	0
0	0	0,26	0,42	0	0	0	0	0
0	0,19	0	0	0	0	0	0	0,42
0	0	0	0	0	1	0,58	0,42	0
0	0	0	0	0	0	0,58	0,42	0,42
0	0	0	0	0	0	0	0,42	0,42

- **Norma euclidiana / soma dos componentes**:

$$b(i,j) = \frac{\|t_j\|_2}{\sum_{k=1}^{l} c(k,j)}$$

Onde i é a coluna e j é a linha da matriz e questão; $c(i,j)$ é o número de vezes que a palavra i ocorre na resposta j; e t_j é o $j-$ésimo vetor coluna.

Para exemplificar é mostrado um exemplo numérico no Quadro 3-14 logo abaixo, no qual é mostrada a aplicação desta ponderação local a Matriz {X}.

Quadro 3-14 Exemplo numérico da aplicação da ponderação norma euclidiana / soma dos componentes à Matriz {X}.

0,58	0	0,45	0	0	0	0	0	0
0,58	0,38	0	0	0	0	0	0	0
0,58	0,38	0,45	0	0,5	0	0	0	0
0	0,38	0,45	0	0,5	0	0	0	0
0	0,38	0,45	1,49	0	0	0	0	0
0	0,38	0	0	0,5	0	0	0	0
0	0,38	0	0	0,5	0	0	0	0
0	0	0,45	0,75	0	0	0	0	0
0	0,38	0	0	0	0	0	0	0,58
0	0	0	0	0	1	0,71	0,58	0
0	0	0	0	0	0	0,71	0,58	0,58
0	0	0	0	0	0	0	0,58	0,58

3.4 MEDIDAS DE SIMILARIDADE

Uma medida de similaridade é uma função que computa o grau de similaridade entre dois objetos que varia de 0 a 1. No contexto deste trabalho a similaridade é usada para comparar duas colunas da matriz termo-documento (vetores que representam uma resposta discursiva) sendo que o resultado desta comparação servira de base para a nota do aluno.

A seguir são mostradas algumas das medidas de similaridade mais utilizadas, sendo que se considera como $t_j = (c_{1j}, \cdots, c_{Ij})$ e $t_k = (c_{1k}, \cdots, c_{Ik})$ a representação vetorial de duas respostas distintas, e j e k o são os índices dessas respostas e I o número total de palavras.

- **Cosseno**: a similaridade é equivalente ao cosseno do angulo formado entre os dois vetores (duas colunas da matriz termo-documento).

$$\frac{t_j \cdot t_k}{\|t_j\|_2 \|t_k\|_2}$$

  Onde $t_j \cdot t_k$ é o produto interno usual entre t_j e t_k.

- **Correlação de Pearson:** Em essência, a de correlação de Pearson encontra a razão entre a covariância e o desvio padrão de ambos os vetores.

$$\frac{cov(t_j, t_k)}{\sqrt{var(t_j) var(t_k)}}$$

Onde cov e var representam covariância e variância entre t_j e t_k, respectivamente.

- **Distância ρ de Spearman:** Distância similar à métrica euclidiana, porém usa permutações.

$$1 - \frac{6\sum_{i=1}^{I} d_i^2}{I^3 - I}$$

Onde di é a distância entre cada valor correspondente de t_k e t_j.

- **Distância de Minkowski:** trata-se de um modelo genérico de distância entre vetores, cujo um dos casos particulares é a distância euclidiana.

$$\sqrt[q]{\sum_{i=1}^{I} d_i^q}$$

Para exemplificar cada uma dessas medidas serão mostrados, no Quadro 3-16, exemplos numéricos para o cálculo da similaridade entre a 1ª e a 4ª coluna da matriz apresentadas no Quadro 3-15, essas colunas são mostradas logo a seguir.

Quadro 3-15 Primeira coluna e quarta coluna da nova matriz {X} para k=2.

0,19	.....	0,26	.....
0,23	.....	0,28	.....
0,46	.....	0,59	.....
0,39	.....	0,5	.....
0,46	.....	0,63	.....
0,26	.....	0,32	.....
0,26	.....	0,32	.....
0,21	.....	0,30	.....
0,17	.....	0,14	.....
-0,02	.....	-0,18	.....
0	.....	-0,21	.....
0	.....	-0,15	.....

Quadro 3-16 Resultado da aplicação das medidas de similaridade

Medidas de Similaridade	Valores
Cosseno	0,967
Correlação de Pearson	0,987
Distância ρ de Spearman	0,985
Distância de Minkowski	0,585

4. METODOLOGIA

4.1 ABORDAGEM PROPOSTA

O problema a ser estudado neste trabalho é a investigação dos parâmetros no uso prático de LSA como uma abordagem de correção automática de questões discursivas. Para isto é utilizada uma abordagem empírica que será validada através de um estudo de caso que será mostrado a seguir.

Nesse estudo de caso, foi desenvolvido um processo metodológico composto de cinco etapas: 1) pré-processamento de texto; 2) ponderação (ou pesagem); 3) cálculo da SVD; 4) cálculo da similaridade; e 5) Aplicação do fator de correção. Sendo que são feitos testes alternando os parâmetros de cada etapa, de modo que seja possível verificar as acurácias de todas as combinações de parâmetros possíveis, a fim de se chegar a uma combinação ótima para cada estudo de caso.

Na fase pré-processamento de textos verificou-se a possibilidade de usar a combinação de bigramas e unigramas como a remoção de *stop words* e o uso de *stemming*. Essas técnicas são aplicadas a todos os textos que servem de entrada para o início de processo, que são as respostas dos alunos e a resposta base do professor.

Na fase de pesagem foram usados esquemas de ponderações locais e globais, bem como as combinações entre ambos os tipos. As ponderações globais utilizadas foram a entropia, inverso da frequência do termo entre documentos e a normal. E por fim, as ponderações locais utilizadas foram o termo-frequência, binária, logaritmo e norma euclidiana/soma dos componentes.

Na fase do cálculo da SVD foi testada a variação da dimensionalidade do espaço semântico (k), a qual pode variar de 1 até o número de elementos do corpus a ser analisado.

Na etapa de cálculo da similaridade foram usados o cosseno, a correlação de Pearson, a distância ρ de Spearman e a distância de Minkowski.

A fim de evitar o mesmo problema ocorrido em DESSUS et al. (2000), onde repostas com poucas palavras ganhavam notas altas, desenvolveu-se a etapa de aplicação do fator de correção. A qual utiliza a seguinte regra: caso um aluno tenha escrito um número de palavras inferior à média de palavras da amostra em estudo menos o desvio padrão da mesma, este aluno terá a nota corrigida proporcionalmente ao número de palavras escrita por esta, de acordo com a fórmula abaixo.

$$nota\ corrigida = nota\ original * \frac{numero\_palavras\_aluno}{média\_palavras\_geral}$$

Este processo resulta nas notas calculadas para todos os alunos, as quais mostram o quanto a resposta de cada aluno é similar a resposta base dada pelo professor. Essas notas podem então ser comparadas às notas atribuídas pelos avaliadores humanos, obtendo-se assim a acurácia.

Este procedimento envolve cálculos bastante trabalhosos e deve ser repetido de diversas vezes a fim de verificar todas as combinações possíveis, surgindo então a necessidade de se criar uma ferramenta que automatize este processo de busca dos parâmetros ótimos. Esta será mostrada logo a seguir.

4.2 FERRAMENTA PARA TESTAR A ABORDAGEM

A fim de automatizar o processo de testes descrito na sessão anterior foi desenvolvida uma ferramenta. A primeira dificuldade encontrada para realização de tal procedimento foi a ausência de um banco de dados de respostas para análise. Dessa forma, decidiu-se criar um módulo de entrada de dados dentro da ferramenta para povoar o banco de dados de questões. Além desse módulo, foram criados mais dois, que são o módulo de pré-processamento e o módulo de ajustes de parâmetros. A organização desses módulos, bem como a descrição resumo do funcionamento da ferramenta, pode ser visualizada na Figura 4-1.

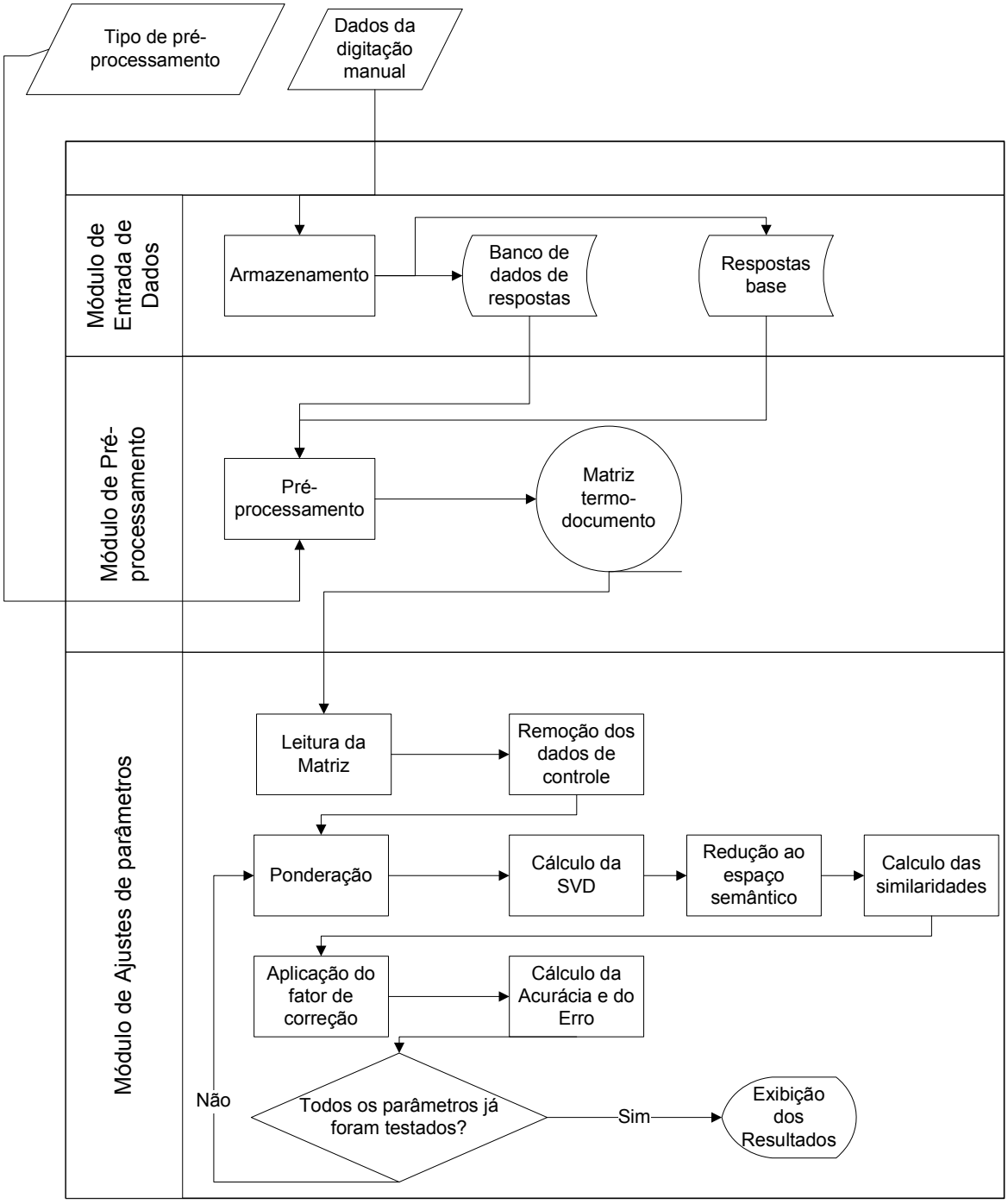


Figura 4-1 Ilustração de funcionamento da ferramenta desenvolvida.

O módulo de pré-processamento usa como dado de entrada o banco de dados que contém as respostas a serem a avaliadas e a resposta correta que servirá de base para a avaliação. Neste módulo, as respostas são submetidas às técnicas de pré-processamento descritas no Capítulo 3 (bigramas ou unigramas, *steamming* e remoção de *stop words*,) onde o usuário pode selecionar quais técnicas vão ser utilizadas. No fim do pré-processamento o

sistema retorna um arquivo texto contendo a matriz termo-documento referente às respostas do aluno juntamente com a resposta base considerando os tipos de pré-processamento selecionados e também informações extras referentes à frequência de uma determinada palavra em todos os textos. Sendo que essas informações são necessárias para os cálculos de algumas ponderações e são acrescentadas como duas colunas e uma linha extras no fim da tabela termo-documento, conforme pode ser visto na Figura 4-2.

MATRIZ TERMO DOCUMENTO	INFORMAÇÃO EXTRA PARA PONDERAÇÃO
INFORMAÇÃO EXTRA PARA PONDERAÇÃO	

Figura 4-2 Ilustração do arquivo texto resultante da etapa de pré-processamento.

Para remoção de *stop words* é utilizada uma lista contendo 299 palavras da língua portuguesa. Para a remoção dos afixos das palavras utiliza-se o algoritmo de *stemming* para a língua portuguesa desenvolvido por ORENGO e HUYCK (2001).

Um dos problemas encontrados durante o desenvolvimento deste módulo da ferramenta foi o processamento dos bigramas, que fazia com que o tempo de execução aumentasse consideravelmente e os resultados não eram satisfatórios. A fim de resolver este problema convencionou-se que só seriam considerados os bigramas que estivessem presentes em pelo menos dois textos. Com isso houve uma melhora no tempo de execução e no resultado final.

O procedimento de ajuste de parâmetros tem como entrada o arquivo texto gerado pelo módulo de pré-processamento que contém a matriz termo documento a ser analisada e outro arquivo contendo as notas dos avaliadores humanos para cada resposta. Após o processamento são geradas como saída, as notas das respostas dos alunos e a melhor configuração de parâmetros que gerou a melhor acurácia entre a avaliação automática a avaliação dos avaliadores humanos, sendo que essa melhor configuração é composta pelos

seguintes parâmetros: melhor ponderação global, melhor ponderação local, melhor valor do espaço semântico (*k*) e melhor similaridade.

Em resumo, o procedimento testa o seguinte conjunto de parâmetros:

- **Espaço semântico:** {1,..., número de amostras}
- **Ponderações locais:** {Termo frequência, Binária, Logaritmo, Norma euclidiana / soma dos componentes}.
- **Ponderações globais:** {Entropia, Inverso da frequência do termo entre documentos, Normal}
- **Similaridades:** {Cosseno, Correlação de Pearson, Distância ρ de Spearman, Distância de Minkowski}

Para esse conjunto de parâmetros é executados o seguinte procedimento descrito na Figura 4-2:

1. Submete a matriz obtida a uma transformação preliminar (ponderação);
2. Calcula a SVD da matriz;
3. Reduz para o espaço semântico;
4. Mede a similaridade entre a resposta base e as demais respostas;
5. Calcula a média e o desvio padrão do número de palavras por texto;
6. Aplica o fator de correção;
7. Calcula a acurácia e o erro cometido.

Figura 4-3 *Procedimento de Ajuste de Parâmetros.*

O procedimento acima é executado repetidas vezes para todas as combinações de parâmetros até que se chegue a combinação que obteve a melhor acurácia.

4.3 ESTUDO DE CASO

O estudo teve como cenário a terceira fase do Processo Seletivo Seriado 2008 da Universidade Federal do Pará (UFPA), o qual se utilizava de questões discursivas para avaliar os seus candidatos. O Centro de Processos Seletivos da Universidade Federal do Pará disponibilizou um conjunto de 1000 respostas selecionadas aleatoriamente dos alunos que fizeram esse concurso de diversas disciplinas.

Nesse concurso as provas de cada disciplina possuíam três questões, no entanto o candidato deveria escolher apenas uma, a qual recebia uma nota de 0 a 6 e representaria a nota total do aluno para aquela disciplina. Deste modo, optou-se por escolher aquelas questões que tiveram maior popularidade para com os alunos, a fim de se ter um banco de dados maior. Esta abordagem resultou na escolha de uma questão de biologia e outra de geografia, que culminou no aproveitamento de 130 avaliações da disciplina de Biologia e 229 avaliações de geografia. As referidas questões serão apresentadas logo a seguir.

A questão de Biologia é de natureza discursivo-conceitual que propunha a elaboração de três conceitos de uma dada taxionomia da Citologia com o seguinte enunciado:

Os tecidos – grupos de células de mesma origem e semelhantes entre si em estrutura e função – são originados nos seres humanos a partir dos três folhetos embrionários. Cite três tipos de tecidos humanos com suas respectivas funções(UFPA, 2008).

No Quadro 4-1, se pode visualizar a grade de correção fornecida para o avaliador da prova de biologia, onde o aluno ganhava um ponto para cada tecido citado corretamente e também recebia mais um ponto, para cada tecido, caso as funções deste tivessem sido citadas corretamente.

Quadro 4-1 Grade de correção para questão de biologia(UFPA, 2008)

Tecidos	Funções
Epitelial ou Glandular	Revestimento interno (trocas, absorção de substâncias), ou externo (proteção, perda de água,) proteção (mecânica), percepção de estímulos, percepção substâncias.
Glandular	Produção de substâncias.
Conjuntivo	Preenchimento, suporte, nutrição dos epitélios, proteção contra infecção, transporte de substância, armazenamento e produção de substâncias, cicatrização de tecidos lesados.
Conjuntivo frouxo	Preenchimento, suporte, nutrição defesa.
Adiposo	Preenchimento e reserva energética.
Reticular	Sustentação.
Cartilaginoso	Sustentação, proteção ou revestimento das articulações.
Ósseo	Proteção, sustentação, armazenamento de cálcio.
Hematopoiético mieloide	Produção de glóbulos vermelhos e plaquetas.
Hematopoiético linfóide	Produção de glóbulos brancos.
Tecido sangüíneo	Transporte de gases, nutrientes.
Tecido linfático	Defesa.

Tecido Muscular	Movimento, contração.
Tecido Nervoso	Regulação e integração interna e coordenação corporal, homeostase, raciocínio, memória, irritabilidade, condução de impulsos nervosos.

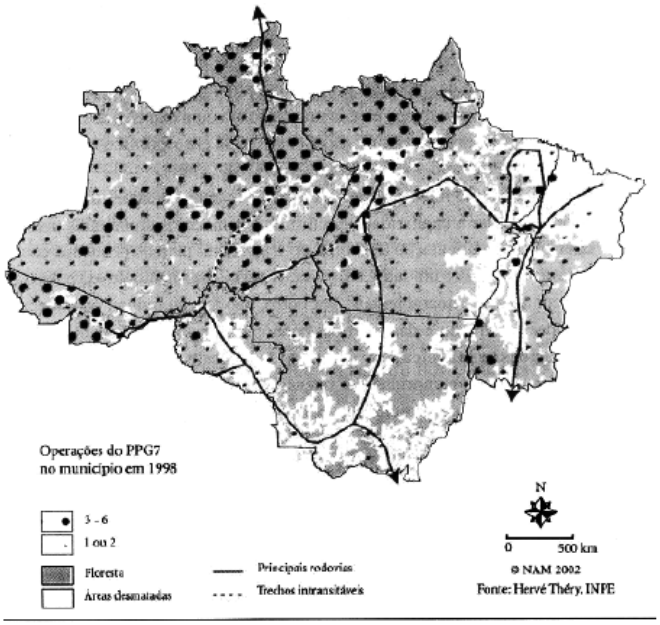
Para a criação da resposta base desta questão foi consultado um especialista, o qual criou texto contendo todos os itens descritos na grade de correção.

Já a questão de geografia era uma questão de natureza discursivo-argumentativa, pois propunha a elaboração de argumentação em defesa de dado ponto de vista formado a respeito de uma constatação que implica conhecimentos da Geografia Humana e Econômica da Região conforme se pode perceber no enunciado do Quadro 4-2. Para essa questão, o aluno receberia uma pontuação baseada nas grades de correção apresentadas nos Quadros 4-3 e 4-4, onde o primeiro Quadro seria a pontuação para uma resposta totalmente correta e o segundo Quadro seria a avaliação para uma resposta parcialmente correta.

Quadro 4-2 Questão de geografia usada no Estudo de Caso (UFPA, 2008)

Sobre o desmatamento na Amazônia, leia o texto e o mapa abaixo:

"De fato, pelas imagens de satélite a possibilidade de imprecisão em torno do desmatamento é grande e os pesquisadores trabalham com aproximações e com cenários projetados. Nestes termos, esse fenômeno, mais as queimadas e a exploração madeireira são das realidades que mais se observa quando se está em trabalho de campo, quer no interior, quer na periferia das cidades. E, para além de uma sociedade em geral insensível quanto à importância dos recursos naturais, dentre os quais os florestais, tem-se um Estado que age, porém, em descompasso com a celeridade dos processos produtivos. O mesmo também se apresenta sempre enfraquecido quanto à questão da garantia dos direitos ambientais definidos constitucionalmente e em leis específicas, o que termina sustentando a impunidade nessa área". (SIMONIAN, L.Tendências recentes quanto à sustentabilidade no uso dos recursos naturais pelas populações tradicionais amazônidas. In: ARAGON, L. E. (ORG.). População e Meio Ambiente na Pan-Amazônia.



Impacto territorial do PPG7

Considerando as informações acima e seus conhecimentos sobre a realidade amazônica:

(A) Identifique as áreas com maior impacto de desmatamento.

(B) Explique o processo de intensificação do desmatamento na Amazônia, valendo-se de dois fatores que estão diretamente relacionados com esse processo.

Quadro 4-3 Grade de correção para questão de geografia para respostas totalmente corretas (UFPA, 2008)

Desempenho	Pontuação
Identifica as áreas considerando os agrupamentos, que podem ser: arco do desmatamento ou arco do povoamento consolidado, ou Amazônia oriental e meridional ou leste e sul da Amazônia. OU Identifica as áreas no interior dos territórios dos Estados membros na sua totalidade.	2,0
Explica que as políticas territoriais implementadas pelo governo federal a partir dos anos sessenta priorizaram a instalação de um modelo baseado na exploração em larga escala dos recursos naturais, por meio da abertura de eixos rodoviários, a exemplo da Belém-Brasília e da Transamazônica, o que facilitou o avanço de frentes de expansão, tais como o extrativismo madeireiro, a mineração e a agropecuária, atividades essas responsáveis por grande parte do desmatamento na região; e refere o processo de povoamento decorrente das migrações com a proliferação de vilarejos, povoados e cidades, o que também contribui para a intensificação do desmatamento.	4,0

Quadro 4-4 Grade de correção para questão de geografia para respostas parcialmente corretas (UFPA, 2008)

Desempenho	Pontuação
Identifica parcialmente as áreas no interior dos territórios dos Estados membros, como, por exemplo, sudeste e sul do Pará, ou norte e noroeste do Tocantins.	1,0
Respostas explicativas que apresentarem apenas um fator, como, por exemplo, a atividade madeireira e sua relação com o desmatamento.	2,0

Devido à natureza discursivo-argumentativa da questão, a grade de correção é mais subjetiva, pois sugere que itens deveriam estar presentes em uma resposta, para que esta recebesse a pontuação máxima. Desse modo, houve a necessidade de se criar uma resposta base para servir de referência. Foram selecionadas aquelas respostas que tiveram a maior pontuação dos avaliadores, com o objetivo de aumentar o vocabulário da resposta base e melhorar a acurácia dos experimentos. A resposta base foi formada pela concatenação de cinco respostas que conseguiram atingir a nota 5, de um total de 229, em um escala de 0 a 6, de maneira semelhante ao que foi feito por SANTOS *et al.* (2007).

Outro detalhe a se ressaltar é que esta questão era subdivida em dois itens, no entanto, para avaliação automática a subdivisão não foi considerada, sendo o texto sempre analisado como um bloco único de texto, até porque só se tinha a pontuação do avaliador para a questão toda ao invés de ter a pontuação individual para cada subitem desta.

Após a definição das questões a serem trabalhadas, com suas respectivas respostas bases, foi feita a digitação manual das questões usando o módulo de entrada de dados da ferramenta desenvolvida neste trabalho. Durante este processo de digitação das respostas foram feitas apenas correções ortográficas por meio de um corretor ortográfico e não foi efetuada nenhum tipo de correção no nível de concordância gramatical.

Terminado o processo de digitação, iniciou-se a aplicação da metodologia proposta deste trabalho, usando a ferramenta desenvolvida cujos resultados serão apresentados no próximo Capítulo.

5. RESULTADOS

Neste capítulo são discutidos os resultados obtidos no estudo de caso. Os resultados são primeiramente agrupados pela combinação dos tipos de pré-processamento onde será mostrada a configuração de parâmetros que obteve o melhor resultado para cada combinação. Após isso, serão mostrados os gráficos do desempenho de cada tipo de parâmetro envolvido, considerando a combinação ótima de parâmetros para cada combinação das técnicas de pré-processamento.

5.1 PRÉ-PROCESSAMENTO

Neste estudo de caso, o processo de avaliação foi executado separadamente para cada combinação de tipo de pré-processamento. Os resultados dessas execuções para a disciplina de biologia podem ser vistos nos Quadros 5-1 e 5-2, enquanto os resultados para a disciplina de geografia são apresentados nos Quadros 5-3 e 5-4. Em ambos os casos o resultado com melhor acurácia está destacado.

Quadro 5-1 Resultados da avaliação automática para disciplina de biologia considerando unigramas.

	Sem pré-processamento	**Sem Stop Words**	**Sem Stop Words+Stemming**
Melhor Acurácia (%)	82,03	84,24	83,33
Redução do Espaço Semântico	7	4	4
Ponderação Local	Termo Frequência	Termo Frequência	Logaritmo
Ponderação Global	Inverso da frequência do termo entre documentos	Inverso da frequência do termo entre documentos	Inverso da frequência do termo entre documentos
Similaridade	Correlação de Pearson	Correlação de Pearson	Correlação de Pearson

Quadro 5-2 Resultados da avaliação automática para disciplina de biologia considerando bigramas.

	Sem pré-processamento	**Sem Stop Words**	**Sem Stop Words+Stemming**
Melhor Acurácia (%)	82,85	84,77	84,23
Redução do Espaço Semântico	6	6	8
Ponderação Local	Termo Frequência	Norma euclidiana / Soma dos componentes	Termo Frequência
Ponderação Global*	-	-	-
Similaridade	Correlação de Pearson	Cosseno	Cosseno

*Para esta combinação de pré-processamento nenhuma técnica contribuiu para um aumento de acurácia.

Quadro 5-3 Resultados da avaliação automática para disciplina de geografia considerando unigramas.

	Sem pré-processamento	**Sem Stop Words**	**Sem Stop Words+Stemming**
Melhor Acurácia (%)	86,89	86,56	86,35
Redução do Espaço Semântico	3	7	8
Ponderação Local	Binária	Binária	Binária
Ponderação Global*	-	-	-
Similaridade	Correlação de Pearson	Cosseno	Cosseno

*Para esta combinação de pré-processamento nenhuma técnica contribuiu para um aumento de acurácia.

Quadro 5-4 Resultados da avaliação automática para disciplina de geografia considerando bigramas.

	Sem pré-processamento	**Sem Stop Words**	**Sem Stop Words+Stemming**
Melhor Acurácia (%)	86,56	85,83	85,97
Redução do Espaço Semântico	10	8	8
Ponderação Local	Sem ponderação	Sem ponderação	Binária
Ponderação Global*	-	-	-
Similaridade	Distância de Minkowski	Distância de Minkowski	Distância de Minkowski

*Para esta combinação de pré-processamento nenhuma técnica contribuiu para um aumento de acurácia.

Analisando os Quadros acima se percebe que para a questão de Biologia o melhor resultado foi obtido usando-se a técnica de remoção de *stop words*, tanto para bigramas quanto para unigramas. Enquanto que para as questões de geografia o melhor resultado foi obtido sem o uso de técnicas de pré-processamento em bigramas e unigramas.

5.2 DIMENSÃO DO ESPAÇO SEMÂNTICO (K)

Para avaliar a influência da variação da dimensão do espaço semântico, variaram-se os valores desse parâmetro para todos os valores possíveis e mantiveram-se fixos todos os outros parâmetros que influenciam no processo de avaliação automática. De maneira que o algoritmo de ajuste de parâmetro foi testado para todos os todos os valores possíveis.

A análise dessa influência para as disciplinas de biologia e geografia pode ser vista nas Figuras 5-1 e 5-2 que apresentam o comportamento da acurácia para a variação de k para as notas da disciplina de Biologia e Geografia, respectivamente. Nelas, percebe-se que tanto para

a disciplina de biologia, quanto para a disciplina de geografia a acurácia máxima é obtida para valores pequenos de *k* (geralmente um valor menor que 10).

Percebe-se ainda na Figura 5-1 que a combinação de unigramas e *stemming* é a técnica de pré-processamento que sofre menos influência de uma escolha errada do valor de *k*. O mesmo também é percebido na Figura 5-2 para todas as combinações que envolvem unigramas.

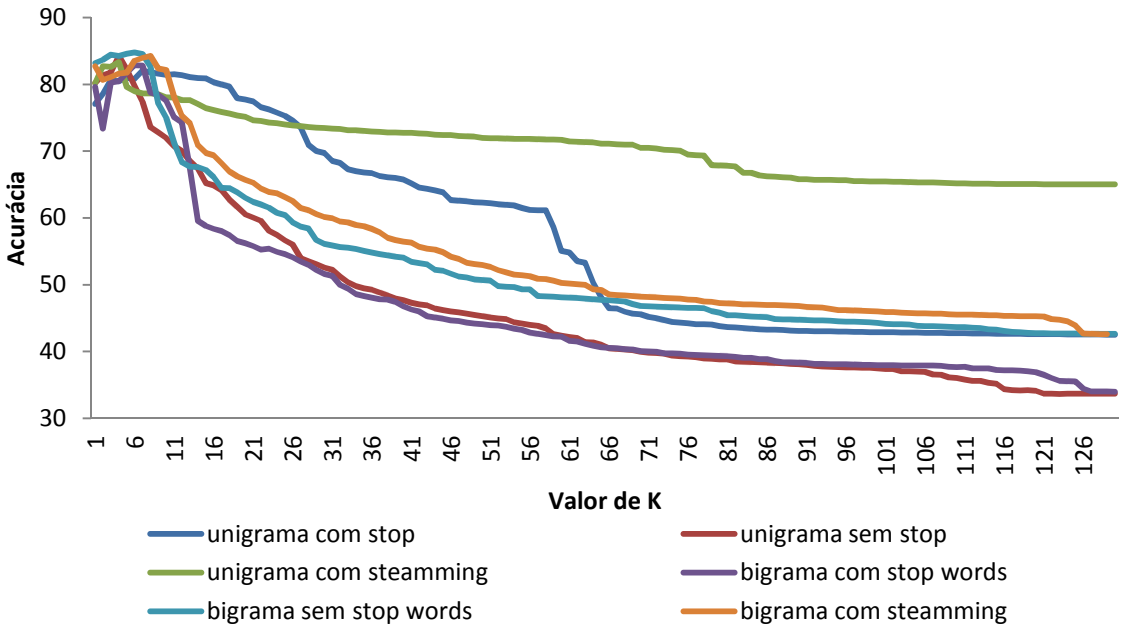


Figura 5-1 *Comportamento da acurácia para a variação de k para as notas da disciplina de Biologia*

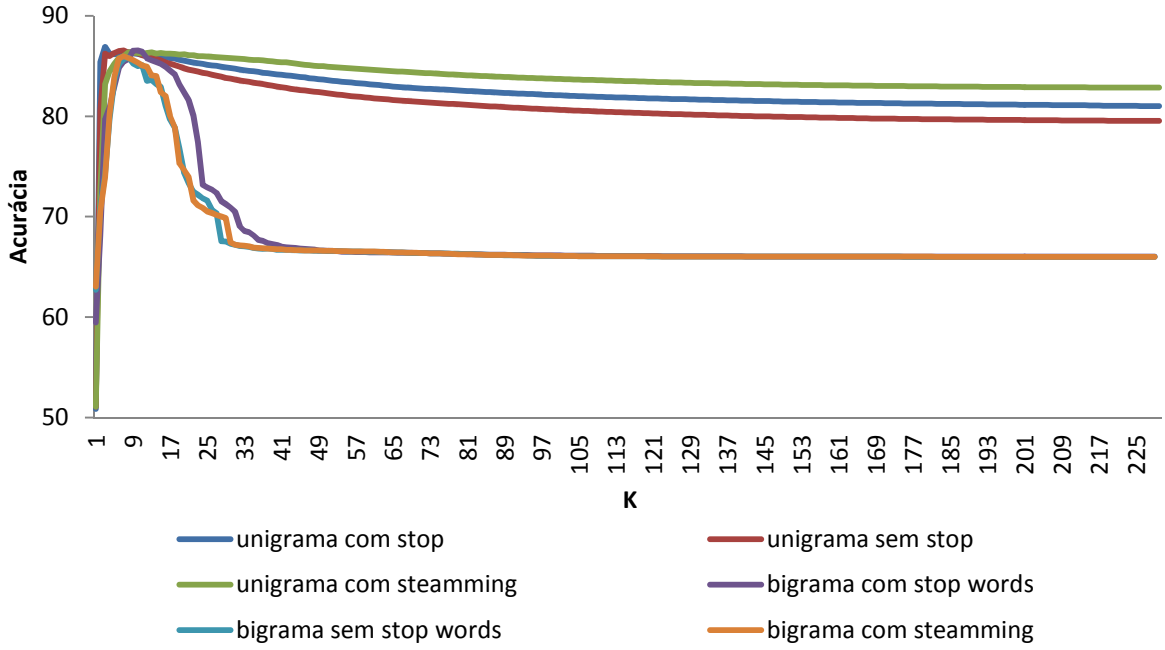


Figura 5-2 *Comportamento da acurácia para a variação de k para as notas da disciplina de Geografia*

5.3 PONDERAÇÃO LOCAL

Para avaliar a influência da variação da ponderação local, verificou-se o valor da acurácia para cada tipo de ponderação local e mantendo fixos todos os outros parâmetros que influenciam no processo de avaliação automática, de maneira similar ao que foi feito na subseção anterior. Como resultado obteve os gráficos da Figura 5-3 e 5-4 para as disciplinas de biologia e geografia, respectivamente.

A Figura 5-3 apresenta o comportamento da acurácia para a variação dos tipos de ponderação local para as notas da disciplina de Biologia. Ao analisá-la, verifica-se que para as notas de biologia, a ponderação local que gerou os melhores resultados na maioria dos casos foi a ponderação Termo Frequência, entretanto o ponderação do Logarítmica obteve um resultado ligeiramente melhor na combinação de unigramas e *stemming* e a ponderação da norma euclidiana / soma dos componentes apresentou um resultado melhor na combinação de bigramas com remoção de *stop words*.

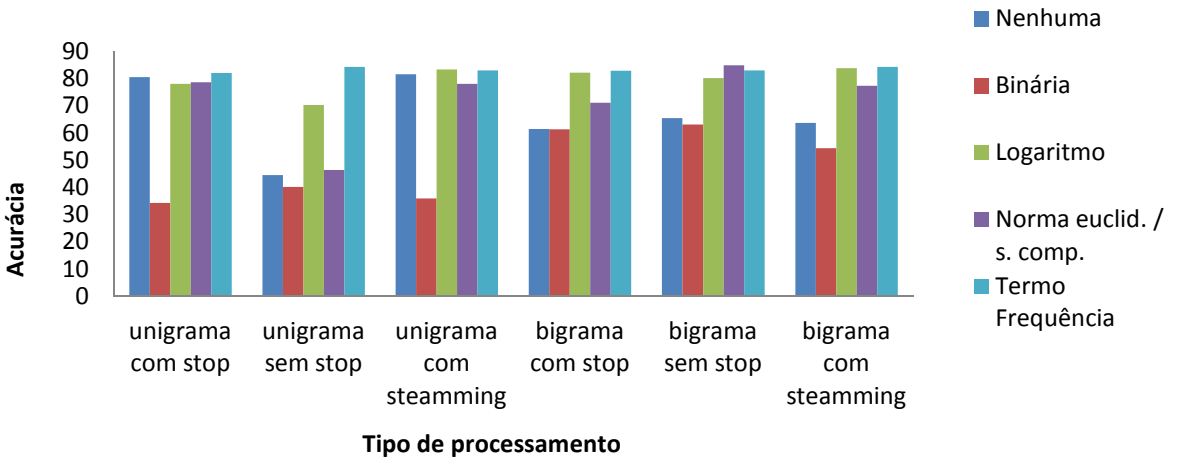


Figura 5-3 *Comportamento da acurácia para a variação dos tipos de ponderação local para as notas da disciplina de Biologia*

Já na Figura 5-4 que apresenta o comportamento da acurácia para a variação dos tipos de ponderação local para as notas da disciplina de Geografia, verifica-se que para as notas de geografia, a ponderação local binária gerou acurácias melhores, quando combinada com unigramas. Para bigramas, os melhores resultados foram obtidos sem o uso de ponderação local.

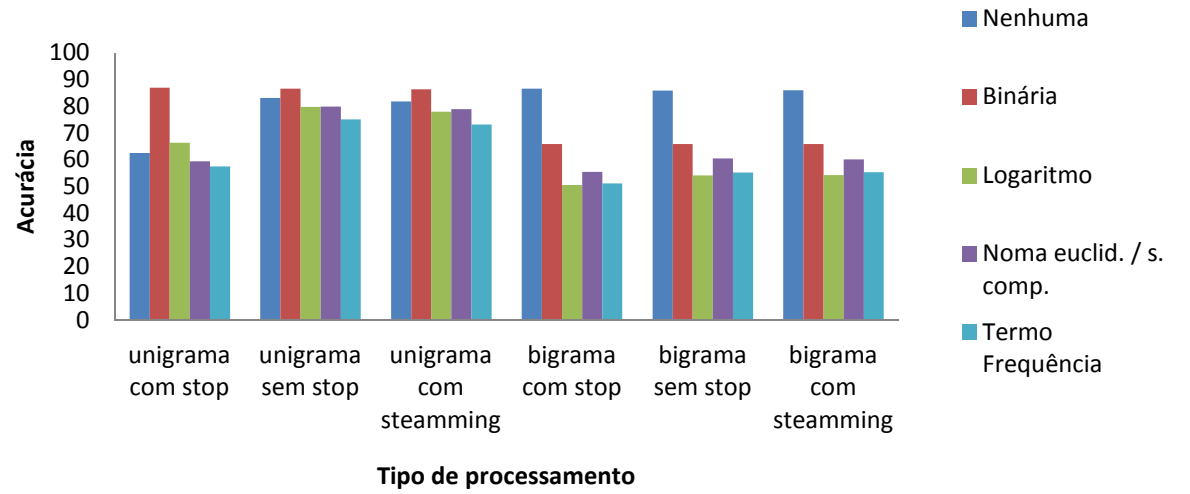


Figura 5-4 *Comportamento da acurácia para a variação dos tipos de ponderação local para as notas da disciplina de Geografia*

5.4 PONDERAÇÃO GLOBAL

Para avaliar a influência da variação da ponderação global, verificou-se o valor da acurácia para cada tipo de ponderação global e mantiveram fixos todos os outros parâmetros que influenciam no processo de avaliação automática, de maneira similar ao que foi feito nas subseções anteriores.

As Figuras 5-5 e 5-6 apresentam o comportamento da acurácia para a variação dos tipos de ponderação global para as notas da disciplina de Biologia e da disciplina de Geografia, respectivamente. Nelas, percebe-se que na maioria dos casos as ponderações globais não tiveram tanta influencia, já que em na grande maioria dos casos quase todas as ponderações globais geraram o mesmo valor de acurácia da ausência de ponderação global. As exceções foram o uso do inverso da frequência do termo entre documentos e a normal termo-frequência. O inverso da frequência do termo entre documentos gerou um pequeno aumento da acurácia nas notas de biologia quando usados juntamente com unigramas e gerou uma diminuição da acurácia nas notas de geografia; e a normal termo-frequência gerou uma grande diminuição da acurácia em combinação com algumas técnicas de pré-processamento.

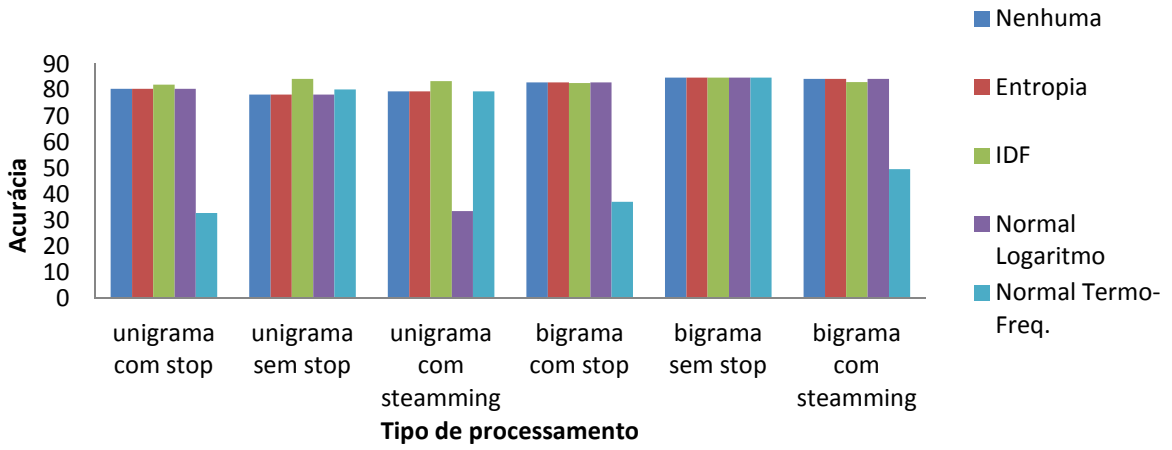


Figura 5-5 *Comportamento da acurácia para a variação dos tipos de ponderação global para as notas da disciplina de Biologia.*

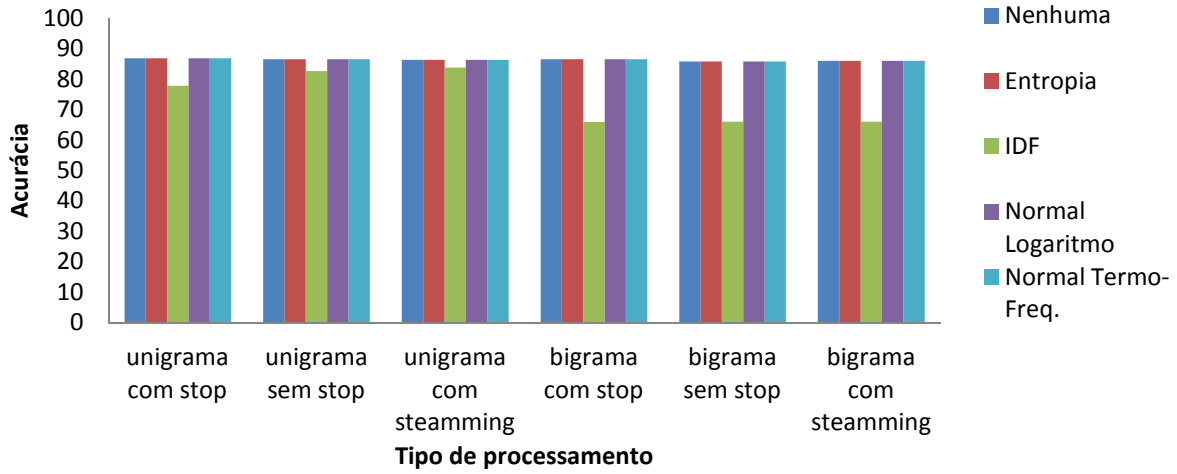


Figura 5-6 *Comportamento da acurácia para a variação dos tipos de ponderação global para as notas da disciplina de Geografia.*

5.5 SIMILARIDADE

Para avaliar a influência da variação da medida de similaridade, verificou-se o valor da acurácia para cada tipo de similaridade e mantiveram fixos todos os outros parâmetros que influenciam no processo de avaliação automática, de maneira similar ao que foi feito nas subseções anteriores.

As Figuras 5-7 e 5-8 mostram o comportamento da acurácia para a variação dos tipos de similaridade para as notas da disciplina de Biologia e da disciplina de Geografia, respectivamente. Ao analisá-las, percebe-se que na maioria dos casos, as medidas cosseno e correlação de Pearson geram resultados de acurácias melhores, principalmente paras notas de biologia. Já em geografia, essas medidas continuam gerando resultados melhores apenas para unigramas, pois em bigramas a distância Minkowski gera resultados ligeiramente melhores em comparação a essas duas distâncias.

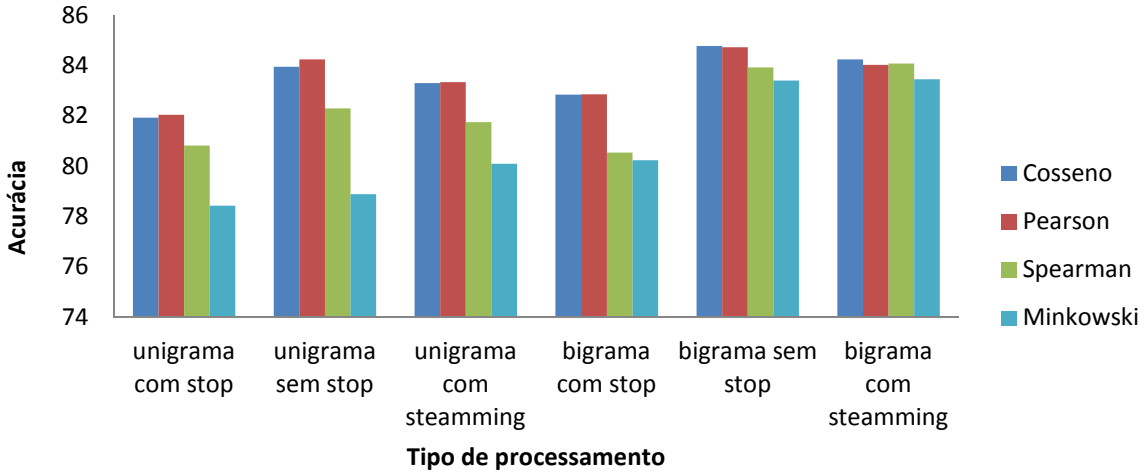


Figura 5-7 *Comportamento da acurácia para a variação dos tipos de similaridade para as notas da disciplina de Biologia.*

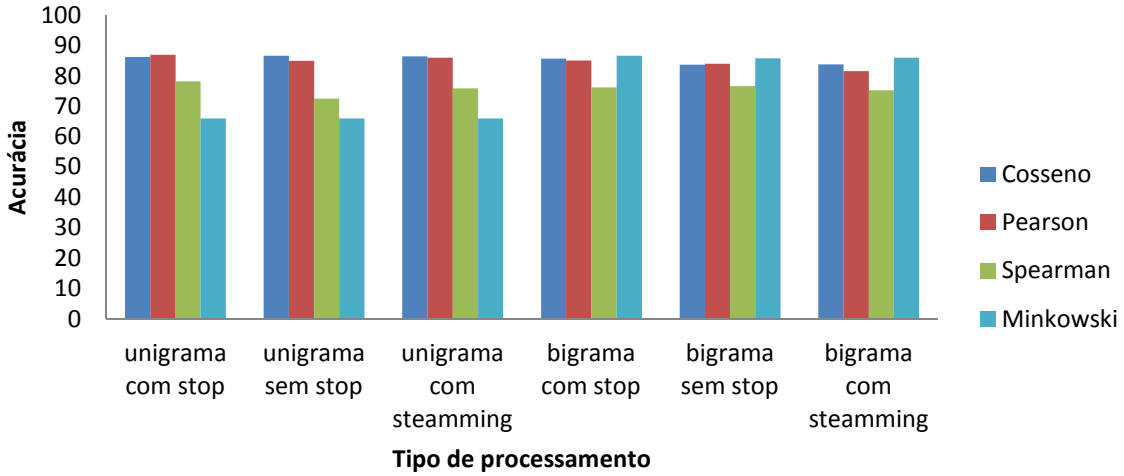


Figura 5-8 *Comportamento da acurácia para a variação dos tipos de similaridade para as notas da disciplina de Geografia.*

6. CONCLUSÕES

Este trabalho apresentou um estudo de caso de aplicação da técnica LSA e uma ferramenta que a implementasse, com o intuito de possibilitar a correção automática de questões discursivas com uma acurácia aceitável em comparação à avaliação humana.

A fim de validar a proposta, esta ferramenta foi testada em um estudo de caso realizado nas provas de biologia e geografia do processo seletivo seriado da Universidade Federal do Pará ocorrido em 2008.

Deste estudo de caso destacam-se como melhores resultados da ferramenta uma acurácia de 84,77% em comparação com as notas dadas por avaliadores humanos para a disciplina de biologia, e a acurácia de 86,89% para a disciplina de geografia. Para biologia, esses resultados foram alcançados usando a combinação de bigramas e a remoção de *stop words* como técnica de pré-processamento; a redução do espaço semântico para 6; a Norma euclidiana / Soma dos componentes como ponderação local; a entropia como ponderação global; e o cosseno como similaridade. Já para geografia, o melhor resultado foi alcançado utilizando unigramas juntamente com a remoção de *stop words*; a redução do espaço semântico para 3; a ponderação local binária; a ponderação global da entropia; e a correlação de Pearson como similaridade.

Com os melhores resultados próximos de 85% considera-se que o objetivo geral do trabalho foi atingido, já que foi desenvolvido com sucesso um estudo de caso para avaliar a viabilidade de utilizar a técnica LSA para correção automática de questões com respostas discursivas do processo seletivo da UFPA com a ajuda da ferramenta desenvolvida que permitiu a execução de diversos experimentos para o ajuste de parâmetros da técnica LSA. Com esses resultados, acredita-se que a tecnologia já pode ser utilizada em para a avaliação automática e em ambientes virtuais de ensino.

Também vale a pena ressaltar que os seguintes objetivos específicos também foram atingidos:

- Fazer um levantamento bibliográfico sobre avaliação automática
- Revisão da literatura sobre o método LSA.
- Composição de uma base de questões de processos seletivos da UFPA.
- Investigação e ajuste de parâmetros para uso prático da técnica usando a ferramenta desenvolvida.

- Realização dos experimentos com a técnica, investigando quais parâmetros trazem os melhores resultados.

Apesar dos bons resultados, vale a pena ressaltar que o objetivo desta ferramenta não é substituir o professor, mas sim agilizar o trabalho deste. Pois o ideal é que esta ferramenta seja sempre supervisionada por especialista, que na maioria das vezes será o próprio professor.

Outro fato a ser destacado é que os resultados não podem ser generalizados, considerando que tanto o desempenho como a calibração de parâmetros do LSA, que é a principal técnica usada na ferramenta, varia de acordo com o domínio do problema. De modo que cada domínio de problema deve ser estudado separadamente.

6.1 TRABALHOS FUTUROS

Para trabalhos futuros, pretende-se verificar aplicação dessa mesma metodologia para outras disciplinas ou até mesmo redações; integrar esta metodologia em um ambiente virtual de aprendizagem; estender o uso da técnica de n-gramas ao invés de limitar-se apenas ao uso de unigramas e bigramas; e aumentar o número de respostas da base de teste.

6.2 PUBLICAÇÕES

- XXXII Congresso da Sociedade Brasileira de Computação (CSBC) - DEsafIE! - I Workshop de Desafios da Computação Aplicada à Educação: **Aceito.**
- Annals of Mathematics and Artificial Intelligence: **Submetido.**

Bibliografia

BERRY, M. W.; BROWNE, M. **Understanding Search Engines, mathematical modeling and text retrieval**. Philadelphia: Society for Industrial and Applied Mathematics, 1999.

BERRY, M. W.; DUMAIS, S. T.; O'BRIEN, G. W.; BERRY, M. W.; DUMAIS, S. T.; GAVIN. Using Linear Algebra for Intelligent Information Retrieval. **SIAM Review**, v. 37, p. 573-595, 1995.

BLACKBOARD. Blackboard - Technology and Solutions Built for Education, 2012. Disponivel em: <http://www.blackboard.com/>. Acesso em: 13 Junho 2012.

BURSTEIN, J.; CHODOROW, M. **Automated essay scoring for nonnative English speakers**. Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics. 1999. p. 68-75.

BURSTEIN, J.; KUKICH, K.; WOLFF, S.; LU, C.; CHODOROW, M.; BRADEN-HARDER, L.; HARRIS, M. D. **Automated scoring using a hybrid feature identification technique**. Proceedings of the 17th international conference on Computational linguistics - Volume 1. Montreal, Quebec, Canada: Association for Computational Linguistics. 1998. p. 206-210.

BURSTEIN, J.; LEACOCK, C.; SWARTZ, R. **Automated Evaluation of Essays and Short Answers**. Learning & Teaching Development. Loughborough University, UK: [s.n.]. 2000.

BURSTEINL, J.; CHODOROW, M.; LEACOCK, C. **Criterion Online Essay Evaluation:** An Application for Automated Evaluation of Student Essays. Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence. [S.l.]: [s.n.]. 2003. p. 3-10.

CALDAS, V. M.; FAVERO, E. L. **Uma Proposta de Avaliação Automática de Mapas Conceituais para Ambientes de Ensino a Distância**. XXXV Conferência Latino Americana de Informática. Pelotas: [s.n.]. 2009. p. 1-1.

CALLEAR, D.; JERRAMS-SMITH, J.; SOH, V.; JERRAMS-SMITH, J.; AE, H. P. **CAA of Short Non-MCQ Answers**. Proceedings of the 5th International CAA conference. [S.l.]: [s.n.]. 2001.

CARVALHO, G.; MATOS, D. M. D.; ROCIO, V. **Document retrieval for question answering:** a quantitative evaluation of text preprocessing. Proceedings of the ACM first Ph.D. workshop in CIKM. New York, NY, USA: ACM. 2007. p. 125-130.

CHRISTIE, J. R. **Automated Essay Marking - for both Style and Content**. [S.l.]. 1999.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. Indexing by latent semantic analysis. **Journal of the american society for information science**, v. 41, n. 6, p. 391-407, 1990.

DESSUS, P.; LEMAIRE, B.; VERNIER, A. **Free-text assessment in a virtual campus**. Proc. Third International Conference on Human System Learning(CAPS ' 3). Paris: Europia. 2000. p. 61-76.

FOLTZ, P.; LAHAM, D.; LANDAUER, T. Automated essay scoring: Applications to educational technology. **Proceedings of EdMedia'99**, p. 939-944, 1999.

FORONDA, D. A. H. **Estudo exploratório da Indexação Semântica Latente e das funções "peso"**. Pontifícia Universidade Católica do Rio Grande do Sul. Porto Alegre, p. 97. 2005.

HEARST, M. The Debate on Automated Essay Grading. **IEEE Intelligent Systems**, Piscataway, NJ, USA, v. 15, p. 22-37, setembro 2000.

KANEJIYA, D.; KUMAR, A.; PRASAD, S. **Automatic evaluation of students' answers using syntactically enhanced LSA**. Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing. Stroudsburg, PA, USA: [s.n.]. 2003. p. 53-60.

KARANIKOLAS, N. N. Computer Assisted Assessment (CAA) of Free-Text: Literature Review and the Specification of an Alternative CAA System. **Enabling Technologies, IEEE International Workshops on**, Los Alamitos, CA, USA, v. 0, p. 116-118, 2010. ISSN 1524-4547.

LANDAUER, T.; FOLTZ, P.; LAHAM, D. An Introduction to Latent Semantic Analysis. **Discourse Processes**, 1998. 259-284.

LARKEY, L. S. **Automatic essay grading using text categorization techniques**. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM. 1998. p. 90-95.

LEACOCK, C. Scoring free-responses automatically: A case study of a large-scale assessment. **Versão em Inglês de C. (2004). Automatisch beoordelen van antwoorden op open vragen**, 2004. Disponivel em: <http://www.cs.pitt.edu/~litman/courses/slate/pdf/erater_examens_leacock.pdf>. Acesso em: 1 Junho 2012.

LIFCHITZ, A.; JHEAN-LAROSE, S.; DENHIÈRE, G. Effect of Tuned Parameters on a LSA MCQ Answering Model. **Behavior Research Methods**, v. 41, p. 1201-1209, 2009.

LINO, A. D. P.; FAVERO, E. L.; SILVA, A. S. **Avaliação automática de consultas SQL em ambiente virtual de ensino-aprendizagem**. Conferencia Ibérica de Sistemas y Tecnologías de la Información. Fernando Pessoa: [s.n.]. 20007. p. 89-100.

MASON, O.; GROVE-STEPHENSON, I. **Automated free text marking with paperless school**. 'Proceedings of the 6th International Computer Assisted Assessment Conference. [S.l.]: [s.n.]. 2002.

MCCALLUM, A.; NIGAM, K. **A comparison of event models for Naive Bayes text classification**. AAAI-98 Workshop on Learning for Text Categorization. [S.l.]: AAAI Press. 1998. p. 41-48.

MIKHAILOV, A. **Indextron**. Intelligent Engineering Systems Through Artificial Neural Networks. [S.l.]: [s.n.]. 1998. p. 57-67.

MING, Y.; MIKHAILOV, A.; LAY KUAN, T. **Intelligent essay marking system**. Educational Technology Conference. [S.l.]: [s.n.]. 2000.

MITCHELL, T.; RUSSELL, T.; BROOMHEAD, P.; ALDRIDGE, N. **Towards robust computerised marking of free-text responses**. Loughborough University. [S.l.]. 2002.

MOODLE, 2012. Disponivel em: <http://www.moodle.org.br/>. Acesso em: 13 Junho 2012.

MORGADO, F. F. **Representação de Documentos Através de Nuvens de Termos**. UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação. Rio de Janeiro, p. 133. 2010.

ORENGO, V. M.; HUYCK, C. **A Stemming Algorithm for the Portuguese Language**. Symposium on String Processing and Information Retrieval - SPIRE'2001. Laguna de San Raphael, Chile: [s.n.]. 2001. p. 186-193.

PAGE, E. B. The Imminence of. Grading Essays by Computer. **The Phi Delta Kappan**, v. 47, n. 5, p. 238-243, Janeiro 1966.

PÉREZ, D.; ALFONSECA, E.; RODRÍGUEZ, P.; GLIOZZO, A.; STRAPPARAVA, C.; MAGNINI, B. About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment. **Revista signos**, v. 38, n. 59, p. 325 - 343, 2005.

PÉREZ, D.; GLIOZZO, A.; STRAPPARAVA, C.; ALFONSECA, E.; RODRIGUEZ, P.; MAGNINI, B. **Automatic assessment of students' free-text answers underpinned by the combination of a BLEU-inspired algorithm and latent semantic analysis**. Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2005 - Recent Advances in Artifical Intelligence. Clearwater Beach, FL, United states: American Association for Artificial Intelligence. 2005.

QUINLAN, J. R. **C4.5:** programs for machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

RIJSBERGEN, C. J. V. **Information Retrieval**. 2ª Edição. ed. London: Butter-worths, 1979.

ROSÉ, C. P.; ROQUE, A.; BHEMBE, D.; VANLEHN, K. **A Hybrid Text Classification Approach for Analysis of Student Essays**. In Building Educational Applications Using Natural Language Processing. [S.l.]: [s.n.]. 2003. p. 68-75.

RUDNER, L. M.; LIANG, T. Automated essay scoring using bayes' theorem. **The Journal of Technology, Learning and Assessment**, n. 1, p. 3-21, 2002.

SANTOS, T. L. T.; SILVA, A. S.; FAVERO, E. L.; LINO, A. D. P. **Avaliação automática de questões conceituais discursivas**. IX Argentine Symposium on Artificial Intelligence - ASAI. Mar Del Plata: Sociedad Argentina de Informática. 2007. p. 128-138.

SANTOS, T. L. T.; SILVA, A.; FAVERO, E. L.; LINO, A. D. P. **Avaliação automática de questões conceituais discursivas**. Simposio Argentino de Inteligencia Artificial. Mar del plata: [s.n.]. 2007. p. 128-138.

SOARES, M. V. B.; PRATI, R. C.; MONARD, M. C. **Improvements on the Porter's Stemming Algorithm for Portuguese**. Latin America Transactions, IEEE (Revista IEEE America Latina). [S.l.]: [s.n.]. 2009. p. 472 -477.

SUKKARIEH, J. Z.; PULMAN, S. G.; RAIKES, N. **Auto-marking:** using computational linguistics to score short, free text responses. Proceedings of the 29th Annual Conference of the International Association for Educational Assessment. Manchester, UK: [s.n.]. 2003.

UFPA. Grades de respostas das provas pss 2008. **Site da Universidade Federal do Pará**, 2008. Disponivel em:

<http://www.ceps.ufpa.br/daves/pss2008/Fase%203/GRADES%20DE%20RESPOSTAS%20DAS%20PROVAS%20PSS%202008.pdf>. Acesso em: 17 Julho 2012.

VALENTI, S.; NERI, F.; CUCCHIARELLI, R. An overview of current research on automated essay grading. **Journal of Information Technology Education**, v. 2, p. 319-330, 2003.

VANTAGE LEARNING TECH. **A study of expert scoring and intellimetric scoring accuracy for dimen-sional scoring of grade 11 student writing responses**. Newtown, PA. 2000.

WHITTINGTON, D.; HUNT, H. **Approaches to the Computerized Assessment of Free Text Responses**. Proceedings of the 3rd International Computer Assisted Assessment Conference. Loughborough University: [s.n.]. 1999. p. 207-219.