João Carlos Alves dos Santos

Avaliação Automática de Questões Discursivas Usando LSA

Belém - Pará

05 de Fevereiro de 2016

João Carlos Alves dos Santos

Avaliação Automática de Questões Discursivas Usando LSA

Tese submetida em cumprimento aos requisitos para obtenção do grau de Doutor em Computação Aplicada

Universidade Federal do Pará – UFPa

Instituto de Tecnologia

Programa de Pós-Graduação em Engenharia Elétrica

Orientador: Prof. Dr. Eloi Luiz Favero

Belém - Pará

05 de Fevereiro de 2016

João Carlos Alves dos Santos
Avaliação Automática de Questões Discursivas Usando LSA/ João Carlos Alves dos Santos. – Belém - Pará, 05 de Fevereiro de 2016-
117 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Eloi Luiz Favero

Tese (Doutorado) – Universidade Federal do Pará – UFPa
Instituto de Tecnologia
Programa de Pós-Graduação em Engenharia Elétrica, 05 de Fevereiro de 2016.

1. Avaliação. 2. Automática. I. Elói Luiz Favero. II. Universidade Federal do Pará. III. Instituto de Tecnologia IV. Avaliação Automática de Questões Discursivas Usando LSA

CDU 02:141:005.7

João Carlos Alves dos Santos

Avaliação Automática de Questões Discursivas Usando LSA

Tese submetida em cumprimento aos requisitos para obtenção do grau de Doutor em Computação Aplicada

Trabalho aprovado. Belém - Pará, 05 de Fevereiro de 2016:

Prof. Dr. Eloi Luiz Favero
Orientador

Professor Dr.
Roberto Célio Limão

Professora Dra
Regina Célia Fernandes Cruz

Professor Dr.
Paulo Cerqueira dos Santos

Professor Dr.
Gustavo Augusto Lima de Campos

Belém - Pará
05 de Fevereiro de 2016

Dedico este trabalho a minha esposa Simone.

AGRADECIMENTOS

Agradecer é sentir de verdade a importância de simples atitudes. Algumas foram essenciais:

Professor Dr. Eloi Favero, muito obrigado pela sua paciência e serenidade durante a orientação deste trabalho;

Professor Dr. José Carlos Fernandes de Oliveira, sua contribuição vai além da realização deste trabalho;

Simone, Maria, Arlindo, obrigado por fazerem parte de minha vida.

Agradecimentos especiais ao Centro de Processamento Seletivo [1] da Universidade Federal do Pará (CEPS), por ter fornecido os dados para o início deste trabalho.

[1] <http://www.ceps.ufpa.br/>

"O verdadeiro perigo não é que computadores começarão a pensar como homens, mas que homens começarão a pensar como computadores"

Sydney J. Harris, jornalista e escritor

RESUMO

Este trabalho investiga o uso de um modelo usando **Latent Semantic Analysis**(LSA) na avaliação automática de respostas curtas, com média de 25 a 70 palavras, de questões discursivas. Com o surgimento de ambientes virtuais de aprendizagem, pesquisas sobre correção automática tornaram-se mais relevantes, pois permitem a correção mecânica com baixo custo para questões abertas. Além disso, a correção automática permite um feedback instantâneo e elimina o trabalho de correção manual. Isto possibilita criar turmas virtuais com grande quantidade de alunos (centenas ou milhares). Pesquisas sobre avaliação automática de textos estão sendo desenvolvidas desde a década de 60, mas somente na década atual estão alcançando a acurácia necessária para uso prático em instituições de ensino. Para que os usuários finais tenham confiança, o desafio de pesquisa é desenvolver sistemas de avaliação robustos e com acurácia próxima de avaliadores humanos. Apesar de alguns estudos apontarem nesta direção, existem ainda muitos pontos a serem explorados nas pesquisas. Um ponto é a utilização de bigramas com LSA, mesmo que não contribua muito com a acurácia, contribui com a robustez, que podemos definir como confiabilidade[2], pois considera a ordem das palavras dentro do texto. Buscando aperfeiçoar um modelo LSA na direção de melhorar a acurácia e aumentar a robustez trabalhamos em quatro direções: primeira, incluímos bigramas de palavras no modelo LSA; segunda, combinamos modelos de co-ocorrência de unigrama e bigramas com uso de regressão linear múltipla; terceira, acrescentamos uma etapa de ajustes sobre a pontuação do modelo LSA baseados no número de palavras das respostas avaliadas; quarta, realizamos uma análise da distribuição das pontuações atribuídas pelo modelo LSA contra avaliadores humanos. Para avaliar os resultados comparamos a acurácia do sistema contra a acurácia de avaliadores humanos verificando o quanto o sistema se aproxima de um avaliador humano. Utilizamos um modelo LSA com cinco etapas: 1) pré- processamento, 2) ponderação, 3) decomposição a valores singulares, 4) classificação e 5) ajustes do modelo. Para cada etapa explorou-se estratégias alternativas que influenciaram na acurácia final. Nos experimentos obtivemos uma acurácia de 84,94% numa avaliação comparativa contra especialistas humanos, onde a correlação da acurácia entre especialistas humanos foi de 84,93%. No domínio estudado, a tecnologia de avaliação automática teve resultados próximos aos dos avaliadores humanos mostrando que esta alcançando um grau de maturidade para ser utilizada em sistemas de avaliação automática em ambientes virtuais de aprendizagem.

Palavras-chaves: Avaliação Automática, LSA, Unigrama, Bigrama.

[2] Dificulta um aluno enganar o sistema com repetição de palavras, por exemplo.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LSA	Latent Semantic Analises
PLN	Processamento de Linguagem Natural
SVD	Singular Value Decomposition
ENEM	Exame Nacional do Ensino Médio
SIGAA	Sistema Integrado de Gestão e Atividade acadêmica
PEG	Project Essay Grader
UFPa	Universidade Federal do Pará
USA	Estados Unidos da América
NLTK	Natural Language Toolkit
RSLP Stemmer	Removedor de Sufixos da Lingua Portuguesa

LISTA DE SÍMBOLOS

λ	Letra grega lambda
$\in$	Pertence
σ	Letra grega sigma
$\mathbb{R}$	Conjunto dos números reais
$\mathbb{R}^n$	Espaço euclidiano n-dimensional
$N(A)$	Núcleo da matriz A
$N(A)^{\perp}$	Complemento ortogonal de N(A)
$\|\cdot\|_2$	Norma quadrática

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Durante sua escolarização, o aluno passa naturalmente por um processo avaliativo de ensino aprendizagem contínuo, cumulativo e sistemático. Mesmo diante das concepções pedagógicas mais modernas, a aplicação de avaliações com questões do tipo discursivas tem grande relevância, pois avaliam a capacidade de leitura, interpretação e escrita de um aluno. Entretanto a tarefa de correção manual de respostas a questões discursivas, para um número considerável de alunos, demanda muito tempo do professor. Marin, Nieto e Rodriguez (2008), He, Hui e Quan (2009), Zen, Iskandar e Linang (2011), Islan e Hoque (2012).

Pesquisadores, Burstein, Chodorow e Leacock (2010), citam algumas vantagens na implementação de sistemas informatizados para correção de respostas a questões discursivas:

- uma redução de custos em mão-de-obra e material de expediente;
- o professor fica isento de qualquer parcialidade no momento da correção;
- com um sistema confiável os resultados tornam-se repetíveis;
- os alunos terão um feedback imediato dos resultados das avaliações, o que possibilita planejamento de seus próximos passos;
- o professor terá um ranking continuo da turma, tornando o aluno ciente da sua avaliação em relação a seus colegas;
- a liberação da carga de trabalho do professor com a correção manual dessas respostas;
- finalmente, o professor terá um continuo acompanhamento da performance da turma, com a identificação de situações extremas, onde ele poderá focar seus esforços ou redirecionar o conteúdo programático.

Decorrente dessas vantagens, pesquisadores, Marin, Nieto e Rodriguez (2008), Refaat et al. (2012) e Magliano e Graesser (2012), revelam ainda que um sistema de avaliação automática pode ser adaptado as características de cada estudante em cada momento.

Pesquisadores, Burstein, Chodorow e Leacock (2010), também citam desvantagens na implementação de sistemas informatizados para automática de respostas a questões discursivas:

- a falta de confiança por conta da precisão do sistema;
- a falta da interação humana;
- um sistema pode se tornar vulnerável a possíveis fraudes;
- a necessidade de se dispor de um conjunto de resposta já avaliadas para treinamento.

Pesquisas sobre o desenvolvimento de tecnologias para avaliação automática de textos escritos já vem ocorrendo desde a década de 60. Page (1966), Hearst (2000), Noorbehbahani e Kardan (2011). Na próxima seção será feita uma breve revisão da literatura citando alguns sistemas informatizados para avaliação automática de textos.

1.2 REVISÃO DA LITERATURA

O sistema informatizado pioneiro na avaliação automática de textos foi o PEG (Project Essay Grader) por Page (1966). Este sistema pontuava um texto pela análise do estilo da escrita e não pelo conteúdo. Os resultados não foram satisfatórios e o sistema foi alvo de muitas críticas da comunidade acadêmica.

O desenvolvimento de novos sistemas informatizados se deu somente a partir da década de 90 devido ao surgimento de técnicas de processamento de linguagem natural (PLN) como em Marinrez et al. (2005) e Burstein, Chodorow e Leacock (2010). A Tabela 1 apresenta um resumo comparativo de alguns sistemas informatizados para avaliação automática de textos que surgiram com a aplicação das técnicas de PLN: como resultado os sistemas atingem uma eficiência superando os 80% para classificação de respostas curtas com no máximo 180 palavras.

Ano	Sistema	Abordagem	Desempenho	Corpus
1998	E-rater	Técnicas estatísticas e Processamento de Linguagem Natural	85%	75000 ensaios GMAT
1998	Larkey	Categorização de textos	80%	respostas curtas sobre questões de estudos sociais, física e direito
2000	IntelliMetric	Combina inteligência artificial, processamento de linguagem natural e técnicas estatísticas	83%	594 textos de estudos sociais escritos por estudantes de 11 anos
2004	C-Rater	Processamento de linguagem natural	85%	pequenas respostas localizadas no final de capítulos de livros escolares
2009	Online Arabic	linguística computacional e técnicas estatísticas	60%	pequenas respostas na língua árabe

Tabela 1 – Sistemas informatizados para avaliação automática de textos

Estes resultados geraram otimismo e grandes expectativas em pesquisadores da área de PLN. Entretanto, sistemas que usam apenas técnicas de PLN foram superados por sistemas que permitem o acesso à semântica das palavras.

Deerwester et al. (1990) publicaram LSA como uma técnica que combina o modelo de espaço vetorial com modelo científico matemático decomposição a valores singulares (SVD). Esta abordagem revela a estrutura semântica subjacente ao texto, e assim a correção de ensaios baseada em modelos de LSA em alguns experimentos produziu resultados mais próximos dos resultados avaliadores humanos. O processo LSA possibilita a recuperação da informação textual um texto a partir de outros textos semanticamente associados a ele, como afirma Navega (2004).

LSA está centrada na técnica SVD, que é um modelo científico, pois cria uma representação conceitual através da fatoração de uma matriz como um produto de outras três matrizes. Na literatura encontramos alguns trabalhos sobre ajustes de parâmetros para modelos LSA para avaliação automática de textos livres. Estes trabalhos abordam o tema de forma

parcial ou justificam que os ajustes dependem do domínio particular de aplicação como em Nakov, Popova e Mateev (2001), Lifchitz (2009) e Jorge-Botana et al. (2010). No Capítulo 3 é apresentada uma revisão sistemática sobre a aplicação de modelos LSA para avaliação automática em ambientes educacionais tendo como base os trabalhos de Haley et al. (2007) e Noorbehbahani e Kardan (2011).

1.3 MAIS INVESTIGAÇÃO

Na revisão sistemática apresentada no Capítulo 3 não foi encontrado nenhum resultado que permita sem esforço de pesquisa a construção de sistemas para avaliação automática de respostas a questões discursivas baseados em um modelo LSA. Falta investigação na calibração de parâmetros de tal modo que um sistema possa ser utilizado em mais de um domínio de aplicação, por exemplo, para correção de questões de biologia e geografia; e, principalmente, falta investigação sobre alcançar uma acurácia que torne a avaliação confiável para uso prático.

Qual é o índice de acurácia que torna o método de avaliação confiável? Pesquisadores, Haley et al. (2007), sustentam que para o sistema ser confiável deve ter sua acurácia comparável com a de avaliadores humanos. Neste caso toma-se um conjunto de resposta discursivas e submete-se a correção por dois avaliadores humanos independentes; depois examina-se a distancia dos valores atribuídos pelos avaliadores humanos, por exemplo, suponha que seja 12%. Neste caso um sistema avaliador automático é confiável se pode substituir um avaliador humano, isto é se a sua acurácia comparada com a de humanos alcance 88%.

A motivação para o desenvolvimento de um sistema de avaliação automática baseado em LSA é o fato de que em um modelo LSA o programador pode interagir diretamente nos resultados através da calibração de parâmetros buscando uma melhor aproximação com avaliadores humanos, e principalmente, não existe nenhum estudo em língua portuguesa para desenvolvimento de um sistema de avaliação automática usando LSA visando sua integração em um sistema virtual de aprendizagem.

1.4 PROPÓSITO DA PESQUISA

No sentido de contribuição, esta pesquisa foca no desenvolvimento de uma ferramenta de avaliação automática de respostas a questões discursivas com a abordagem LSA cujo propósito é alcançar uma acurácia confiável para uso prático na avaliação automática de questões discursivas em ambientes virtuais de aprendizagem.

Neste trabalho, foi pesquisado o ajuste de parâmetros de um modelo de avaliação automática baseado em LSA para alcançar uma acurácia próxima a de avaliadores humanos. O modelo deverá ser treinado e testado em pelo menos dois conjuntos de respostas a duas questões discursivas (de domínios diferentes) de um processo seletivo para admissão em um curso superior, assim validando o modelo em dois domínios diferentes.

1.5 OBJETIVOS ESPECÍFICOS

1. Criar um corpus[1] digitalizado constituído por respostas a questão discursivas de provas do vestibular para certificação do sistema baseado em LSA que estamos propondo, considerando pelo menos dois domínios;
2. Fazer um levantamento bibliográfico sobre sistemas de avaliação automática, relatando abordagens utilizadas e melhores resultados obtidos;
3. Preparar uma revisão da literatura sobre o modelos LSA aplicados em domínios educacionais descrevendo os conceitos da abordagem e metodologias da calibração de parâmetros;
4. Refazer experimentos com n-gramas para servir como base-line de performance;
5. Investigar e ajustar parâmetros de LSA visando superar a acurácia de n-gramas;
6. Iniciar estudos para verificar a viabilidade da implementação do sistema de avaliação automática proposto no Sistema Integrado de Gestão de Atividades Acadêmicas – SIGAA – da UFPa
7. Criar um corpus de redações, de provas do Exame Nacional do Ensino Médio (ENEM), e iniciar o estudo de avaliação automática deste corpus.

1.6 METODOLOGIA

Dentro do propósito da pesquisa, o corpus foi constituído por respostas para duas questões discursivas, uma conceitual e outra argumentativa, e que foram avaliadas previamente por avaliadores humanos.

A questão conceitual foi de Biologia que propunha a elaboração de três conceitos de uma dada taxonomia da Citologia. A questão argumentativa foi de Geografia que propunha a elaboração de uma argumentação em defesa de dado ponto de vista formado a respeito da

[1] Um conjunto de textos escritos ou falados numa língua, disponível para análise

Geografia Humana e Econômica da Região.

As questões de Biologia e de Geografia eram parte do boletim de questões do Processo Seletivo Seriado 2008 da Universidade Federal do Pará.

Para composição do corpus, de um universo de 1000 folhas de respostas disponibilizadas foram selecionadas as duas questões com mais folhas de respostas preenchidas: uma questão de Biologia com 130 folhas de respostas e uma questão de Geografia com 229 folhas de respostas. Do total de 26 questões o aluno era obrigado a responder apenas um terço delas, por isso, não tivemos 1000 respostas por cada questão.

Durante o processo de digitalização das respostas foram feitas apenas correções de ortografia sem alterar a concordância gramatical do texto original.

As respostas foram confrontadas com uma resposta de referência. Para a questão de Biologia a resposta base foi um texto, dado por um especialista humano, com todos os conceitos corretos para as escolhas possíveis nas respostas. Para a questão de Geografia a resposta de referência foi montada pela concatenação das respostas dos alunos que obtiveram a maior pontuação por avaliadores humanos.

O modelo aqui proposto representa o corpus de respostas por uma matriz, onde o número de linhas representa o vocabulário das palavras e cada coluna é uma representação vetorial de uma resposta.

O modelo estima a pontuação de cada resposta por um algoritmo em seis passos:

1. Preprocessamento: contamos os unigramas e bigramas de cada uma das resposta para construção da matriz inicial;
2. Pesagem das entradas: uma função peso é aplicada na matriz inicial e expressa a importância de cada palavra em cada resposta e no corpus como um todo;
3. SVD:
    a) Cálculo do SVD: a matriz inicial é fatorada num produto de três outras matrizes;
    b) Redução para o espaço semântico: empiricamente escolhemos a dimensão do espaço semântico;
4. Classificação: cada resposta é comparada com a resposta de referência;

5. Ajustes:

    a) Fator de penalização: baseado no valor médio e desvio padrão do número de palavras de cada resposta;

    b) Re-classificação: após aplicação do fator de penalização, cada resposta é novamente comparada com a resposta de referência;

6. Acurácia:

    a) Cálculo do erro: cálculo da média aritmética dos erros cometidos em cada comparação;

    b) Cálculo da Acurácia:

$$acuracia = \frac{6 - erro\ medio}{6} \times 100$$

O algoritmo é executado algumas vezes, os passos de 1 a 6 são repetidos, os parâmetros são alterados e são guardadas as melhores configurações encontradas. Mais de 60.000 execuções foram realizadas.

Como o avaliador humano atribuiu uma pontuação inteira entre 0 e 6 foi necessário categorizar as pontuações do modelo LSA: particionamos o intervalo $[0, 1]$ em sete partes iguais associando a cada parte os inteiros $0, 1, 2, 3, 4, 5$ e 6, respectivamente. Comparamos a pontuação do modelo LSA com a pontuação dada por avaliadores humanos para calcular a acurácia: o valor absoluto da diferença entre as pontuações é o erro percentual; somando-se todos os erros percentuais e dividindo-se pelo número de respostas obtemos o erro médio. Obtido o erro médio calculamos a acurácia pela fórmula do item 6.b) do algoritmo acima.

1.7 PRINCIPAIS RESULTADOS

Executamos inúmeras iterações do processo como um todo, guardando as melhores configurações de cada experimento realizado. Como resultado dos procedimentos, obteve-se 84,94% como melhor índice de acurácia; com ajustes finos de seus parâmetros o modelo LSA pode atingir uma acurácia minimamente aceitável para uso prático em ambientes virtuais de aprendizagem.

1.8 ORGANIZAÇÃO DO TRABALHO

O trabalho esta organizado da seguinte forma: no Capítulo 1 temos a introdução, no Capítulo 2 temos a Matemática de LSA, no Capítulo 3 temos uma revisão sistemática da literatura de modelos LSA aplicados em ambientes educacionais, no Capítulo 4 temos o modelo LSA proposto, finalmente, no Capítulo 5 temos uma aplicação prática do modelo LSA na avaliação automática de respostas escritas.

2 A MATEMÁTICA DE LSA

2.1 UMA DEFINIÇÃO DE LSA

Análise Semântica Latente (Latent Semantic Analise, LSA) é uma metodologia para recuperação de informação de dados textuais através do computador que foi patenteada em Junho de 1989 por Deerwester et al. (1990). A informação é tratada em um domínio estatístico presumindo que existe uma estrutura semântica latente no uso das palavras dos textos, que entendemos como uma ocorrência linguística, que tem um sentido completo, dotada de certas formalidades que permite estabelecer uma comunicação entre as partes. Esta abordagem baseia-se no fato de que as estimativas desta estrutura semântica são utilizadas para representar e recuperar informações. Uma provável definição de Análise Semântica Latente pode ser dada da seguinte maneira:

Definição 2.1. *Análise Semântica Latente(LSA) é um modelo estatístico - matemático para estimar a similaridade do significado de palavras e partes de textos baseado na co-ocorrência de palavras.*

A estimativa da similaridade decorre do fato de que LSA pressupõe a existência de uma estrutura semântica obscurecida parcialmente pelas flexões gramaticais das palavras. Esta estrutura, frequentemente chamada de espaço semântico, analisa o conteúdo semântico das palavras. Na realidade, LSA nada mais é do que um modelo espaço vetorial: a análise semântica é realizada após a representação vetorial dos textos sendo a similaridade obtida através de operações entre esses vetores. Este é o diferencial de um modelo LSA. Na próxima seção será feita uma breve Descrição de LSA.

2.2 BREVE DESCRIÇÃO DA TÉCNICA LSA

Em ambientes computacionais os conteúdos dos textos normalmente são tratados através de indexações. LSA usa a representação vetorial para aquisição de conhecimentos contidos em textos. A representação é chamada vetorial pelo fato de que cada texto é representado por um vetor cujas coordenadas são as frequências das palavras que compõe o próprio texto.

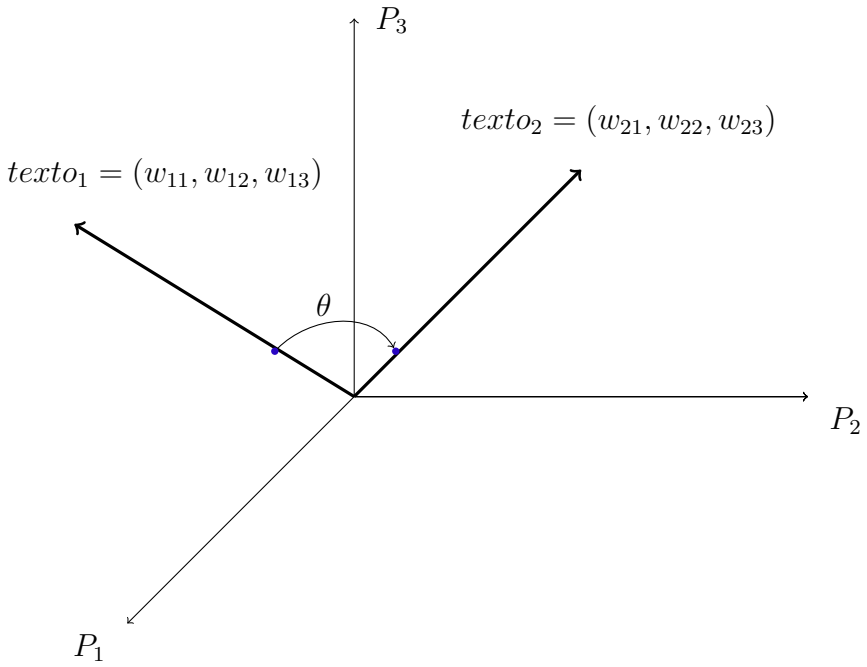


Figura 1 – Representação vetorial de textos

A Figura 1, acima mostra a representação vetorial de dois textos, $texto_1$ e $texto_2$, indexados pelas palavras P_1, P_2 e P_3 com suas respectivos frequências, em geral já submetidas a uma transformação ponderação. Os pesos indicam o grau de especificidade relativamente a cada palavra e são usados para medir a similaridade entre textos. Se θ é o ângulo entre os vetores que representam $texto_1$ e $texto_2$, então $\cos\theta$ é uma estimativa da similaridade entre eles.

Suponha que o corpus é constituído por n documentos e um vocabulário com m palavras. Então construímos uma matriz A de ordem $m \times n$, chamada de matriz termo-documento, onde cada entrada (i, j) dessa matriz significa a quantidade de vezes que a palavra i, com $1 \leq i \leq m$, aparece no documento j, com $1 \leq j \leq n$. Em resumo, cada vetor linha da matriz está associado com uma palavra do vocabulário e cada vetor coluna associado com um documento do corpus. Assim é possível medir a similaridade entre as palavras ou entre documentos através de operações entre seus vetores correspondentes. A forma mais frequentemente utilizada para medir a similaridade é o cosseno do ângulo.

Após construída, a matriz A é submetida a uma transformação preliminar na qual cada frequência é ponderada por uma função peso que estima a importância da palavra no texto em que ela está contida e seu grau de influência no corpus.

Em seguida, aplica-se o Teorema de Decomposição a Valores Singulares(SVD) na matriz A que é fatorada como sendo um produto de três matrizes da seguinte maneira: $A = USV^t$, onde U e V são matrizes ortogonais quadradas de ordens m e n, respectivamente, e

$S = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$ sendo D uma matriz diagonal cujos elementos diagonais são os r valores singulares de A, onde r é o posto da matriz A[1]. Esta decomposição revela a arquitetura das correlações entre as palavras nos textos.

Um problema que ocorre após a etapa SVD é o da dispersão das informações, pois algumas entradas da matriz A podem ser nulas significando que os dados observados não foram suficientes. Uma forma de suavizar este problema é aproximar a matriz A por uma outra matriz A_k do mesmo tipo $m \times n$ e com posto k, em geral, bem menor do que o posto r da matriz A. Para isto escolhemos as k primeiras linhas e colunas das matrizes U, D e V. Nas primeiras colunas dessas matrizes estão os autovetores associados aos valores singulares de maior magnitude. Veremos na Seção 2.4 que a matriz A_k é a melhor aproximação por mínimos quadrados da matriz A.

O método LSA transfere a discussão do espaço inicial para um espaço de dimensão bem menor que é o espaço gerado pelas colunas da matriz A_k, chamado espaço semântico ou latente. Esta redução da dimensão do espaço é etapa mais importante em todo método LSA. Um espaço com dimensão elevada dificulta o tratamento computacional de dados possibilitando omissão de informações e que comparações lexicais sejam consideradas falsas.

Outra razão pela qual a redução da dimensionalidade é importante é o fato de que ocorre uma melhora as relações entre os textos através da identificação de estruturas semânticas ocultas nas relações entre palavras e textos. Na maioria das aplicações, a dimensão do espaço latente é muito menor do que a dimensão do espaço inicial. Não existe nenhuma teoria nem método que forneça a melhor dimensão. Na literatura encontramos, em muitos casos, que a melhor dimensionalidade é intrínseca ao domínio de aplicação e deve ser determinada empiricamente. Entretanto em Wild et al. (2005) encontramos algumas sugestões para esta escolha, e mais recentemente, no trabalho de Fernandes, Artifice e Fonseca (2012) encontramos uma fórmula para a dimensão do espaço semântico. É no espaço semântico que se determina um valor para a similaridade entre textos.

Para finalizar, na maioria dos trabalhos pesquisados foi encontrada a seguinte arquitetura básica para um modelo LSA:

[1] Posto de uma matriz é a dimensão do espaço das linhas (ou colunas) da própria matriz

1. Pré-processamento do texto de entrada;
2. Uso de esquemas de ponderação;
3. Cálculo da SVD
4. Escolha da dimensionalidade;
5. Aplicar medida de similaridade.

Nesta arquitetura observa-se que a principal etapa é o cálculo da SVD, pois a partir deste cálculo que o espaço semântico é gerado. Na próxima seção será apresentada uma demostração do método matemático de análise numérica Decomposição a Valores Singulares(SVD).

2.3 UMA DEMONSTRAÇÃO DO TEOREMA DA SVD

O Método SVD, Lay (2012), afirma que toda matriz real A do tipo $m \times n$ pode ser fatorada como um produto de três matrizes:

$$A = USV^t,$$

onde U é matriz ortogonal do tipo $m \times m$, V é também ortogonal do tipo $n \times n$ e S é uma matriz retangular do mesmo tipo $m \times n$ da matriz A.

Antes da demonstração do Teorema do Método SVD vejamos alguns preliminares.

Definição 2.2. *Seja A uma matriz quadrada de ordem n sobre o corpo $\mathbb{R}$ dos números reais, $\lambda \in \mathbb{R}$ um escalar e v um vetor não nulo pertencente a $\mathbb{R}^n$. Dizemos que λ é um autovalor de A e que v é um autovetor de A associado ao autovalor λ se, e somente se, $Av = \lambda v$.*

Para todo matriz A do tipo $m \times n$ sobre o corpo $\mathbb{R}$, as matrizes A^tA e AA^t estão bem definidas e são matrizes quadradas de ordens n e m, respectivamente. É fácil verificar que as matrizes A^tA e AA^t tem o mesmo posto da matriz A e que são ambas simétricas, sendo portanto diagonalizáveis. Desta forma, podemos construir bases ortonormais de $\mathbb{R}^n$ e $\mathbb{R}^m$ constituídas por autovetores de A^tA e AA^t, respectivamente.

Seja $\{v_1, \cdots, v_n\}$ uma base ortonormal de $\mathbb{R}^n$ constituída por autovetores de A^tA. Supondo $A^tAv_i = \lambda_i v_i$, temos:

$$\begin{aligned} \|Av_i\|^2 &= v_i^t A^t A v_i \\ &= v_i^t \lambda_i v_i \\ &= \lambda_i \|v_i\|^2 \\ &= \lambda_i. \end{aligned}$$

Desta forma os autovalores $\lambda_1, \cdots, \lambda_n$ são números reais não-negativos; assim, podemos ordena-los de forma que

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0.$$

Definição 2.3. Os valores singulares de A são as raízes quadradas dos autovalores de A^tA.

Denotaremos os valores singulares de A por $\sigma_1, \cdots, \sigma_n$:

$$\sigma_1 = \sqrt{\lambda_1} \geq \cdots \geq \sigma_n = \sqrt{\lambda_n}.$$

Teorema 2.1 (Teorema da decomposição a Valores Singulares)**.** *Seja A uma matriz real não-nula do tipo $m \times n$ de posto r. Sejam $\sigma_1 \geq \cdots \geq \sigma_r > 0$ os valores singulares de A. Então existem matrizes ortogonais U e V de ordens m e n, respectivamente, e uma matriz S do tipo $m \times n$ e da forma*

$$S = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix},$$

onde D é a matriz diagonal do tipo $r \times r$ dada por $D = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix}$, tais que

$$A = USV^t.$$

As colunas de U e V,

$$U = [u_1 \ u_2 \ \cdots \ u_m] \text{ e } V = [v_1 \ v_2 \ \cdots \ v_n]$$

são denominadas vetores singulares à esquerda e à direita de A, respectivamente.

Antes de apresentarmos uma demonstração do Teorema 2.1 vejamos o

Lema 2.1. *Suponha que $\{v_1, \cdots, v_n\}$ é uma base ortonormal de $\mathbb{R}^n$ constituída por autovetores de A^tA e que A tem r valores singulares não nulos. Então, o posto de A é igual a r e $\{Av_1, \cdots, Av_r\}$ é uma base ortogonal do espaço das colunas da matriz A.*

Demonstração do Lema 2.1: Suponha que cada autovetor v_i esta associado ao autovalor λ_i, $i = 1, \cdots, n$. Como $\{v_1, \cdots, v_n\}$ é uma base ortonormal de $\mathbb{R}^n$, segue que o conjunto $\{Av_1, \cdots, Av_n\}$ é ortogonal pois, para $i \neq j$, temos

$$< Av_i, Av_j >= v_i^t A^t A v_j = v_i^t \lambda_j v_j = \lambda_j v_i^t v_j = 0.$$

Como só existem r valores singulares não nulos de A, temos que

$$\begin{cases} Av_i \neq 0, & 1 \leq i \leq r \\ Av_i = 0, & r+1 \leq i \leq n \end{cases}.$$

Se y é um elemento qualquer do espaço das colunas de A, então $y = Ax$ para algum $x \in \mathbb{R}^n$. Assim,

$$\begin{aligned} y &= \alpha_1 Av_1 + \cdots + \alpha_r Av_r + \alpha_{r+1} Av_{r+1} + \cdots + \alpha_n Av_n \\ &= \alpha_1 Av_1 + \cdots + \alpha_r Av_r. \end{aligned}$$

Desta forma o conjunto $\{Av_1, \cdots, Av_r\}$ gera o espaço das colunas de A e como é linearmente independente, pois é ortogonal, segue que é uma base para o espaço das colunas de A. Como o posto de A é igual a dimensão do espaço das colunas segue que o posto de A é igual a r. ■

Demonstração do Teorema 2.1: Seja $\{v_1, \cdots, v_n\}$ uma base ortonormal de $\mathbb{R}^n$ constituída por autovetores de A^tA. Denotaremos por $N(A)$ o núcleo da matriz A. Sem perda de generalidade, podemos supor que $V_1 = \{v_1, \cdots, v_r\}$ é uma base ortonormal de $N(A)^\perp$ de A e daí $V_2 = \{v_{r+1}, \cdots, v_n\}$ é uma base ortonormal de $N(A)$. Formemos a matriz

$$V = [V_1\, V_2] = [v_1, \cdots, v_r, v_{r+1}, \cdots, v_n].$$

Pelo Lema 2.1, o conjunto $\{Av_1, \cdots, Av_r\}$ é uma base do espaço das colunas de A; definamos os vetores $u_1, \cdots, u_r$ por

$$u_i = \frac{Av_i}{\|Av_i\|}, \ : i = 1, \cdots, r.$$

Desta forma, o conjunto $U_1 = \{u_1, \cdots, u_r\}$ é uma base ortonormal do espaço das colunas de A e podemos estende-lo até formarmos uma base ortonormal de $\mathbb{R}^m$:

$$\{u_1, \cdots, u_r, f_1, \cdots, f_{m-r}\}.$$

É fácil ver que $U_2 = \{f_1, \cdots, f_{m-r}\}$ é uma base ortonormal do complemento ortogonal do espaço das colunas de A.

Formemos a matriz

$$U = [U_1\, U_2] = [u_1 \ \cdots \ u_r, f_1, \cdots, f_{m-r}].$$

É claro que V e U são matrizes ortogonais. Notemos que o produto da matriz A pela matriz V é dado por

$$AV = A\,[V_1\, V_2] = [AV_1\, AV_2] = [AV_1\, 0].$$

Vamos mostrar que o produto AV é igual ao produto da matriz U pela matriz S. Observemos primeiro que o produto da matriz U_1 pela matriz D é dado por

$$\begin{aligned} U_1 D &= [u_1 \cdots u_r] \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix} \\ &= \left[\frac{Av_1}{\sigma_1} \cdots \frac{Av_r}{\sigma_r}\right] \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix} \\ &= [Av_1 \cdots Av_r] \\ &= AV_1. \end{aligned}$$

O produto da matriz U pela matriz S é então dado por

$$\begin{aligned} US &= [U_1\, U_2] \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \\ &= [U_1 D\, 0] \\ &= [AV_1\, 0]\,. \end{aligned}$$

Portanto

$$AV = US.$$

Como V é ortogonal podemos escrever

$$A = USV^t. \quad \blacksquare$$

A matriz A poderia ser obtida, mais simplesmente, se considerarmos as matrizes

$$U = [u_1 \cdots u_r]\,,\ D = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix},\ V = [v_1 \cdots v_r]\,.$$

O Teorema de Decomposição a Valores Singulares é um dos resultados mais importantes em Álgebra Linear Computacional. O método SVD é a base dos métodos para resolução de problemas de mínimos quadrados. Será visto na próxima seção como o método SVD pode ser aplicado na resolução do problema de aproximar uma matriz por outra matriz de posto menor.

2.4 MELHOR APROXIMAÇÃO EM MÉDIA QUADRÁTICA DE MATRIZES DE POSTO FIXADO

Seja A uma matriz real do tipo $m \times n$. A norma quadrática de A é definida por

$$\| A \|_2 = \max_{x \in \mathbb{R}^n, \|x\|=1} \| Ax \|_2 .$$

Suponhamos que o postoda matriz A seja igual a r. Pelo Teorema de Decomposição a Valores Singulares, a matriz A pode ser escrita da seguinte forma:

$$A = UDV^t = [u_1 \ \cdots \ u_r] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} [v_1 \ \cdots \ v_r]^t .$$

Ou ainda

$$A = \sigma_1 u_1 v_1^t + \cdots + \sigma_r u_r v_r^t = \sum_{i=1}^{r} \sigma_i u_i v_i^t.$$

Nesta decomposição, podemos escolher as k primeiras colunas da matriz A e formar uma matriz:.

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^t.$$

Como A tem posto r e $k < r$, então a matriz A_k tem posto k. Vamos mostrar que A_k é a matriz que melhor se aproxima da matriz A relativamente à norma $\| \ . \ \|_2$.

Pelo Teorema de Decomposição a Valores Singulares, a matriz $A - A_k$ pode ser escrita na forma:

$$A - A_k = U \begin{bmatrix} \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \sigma_{k+1} & & \\ & & & & \ddots & \\ & & & & & \sigma_r \end{bmatrix} & 0 \\ 0 & 0 \end{bmatrix} V^t.$$

Teorema 2.2. *A norma quadrática da matriz $A - A_k$ é dada por*

$$\|A - A_k\|_2 = \sigma_{k+1}.$$

Demonstração: Pela definição de $\| \ . \ \|_2$, temos:

$$\| \begin{bmatrix} \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \sigma_{k+1} & & \\ & & & & \ddots & \\ & & & & & \sigma_r \end{bmatrix} & 0 \\ 0 & 0 \end{bmatrix} \|_2 = \max_{\|x\|=1} \| \begin{bmatrix} \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \sigma_{k+1} & & \\ & & & & \ddots & \\ & & & & & \sigma_r \end{bmatrix} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \|_2$$

$$= \max_{\|x\|_2=1} \| \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sigma_{k+1}x_{k+1} \\ \vdots \\ \sigma_r x_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} \|_2$$

$$= \max_{\|x\|_2=1} \sqrt{(\sigma_{k+1}x_{k+1})^2 + \cdots + (\sigma_r x_r)^2}$$

$$= \sigma_{k+1} \max_{\|x\|_2=1} \sqrt{x_{k+1}^2 + \cdots + (\frac{\sigma_r}{\sigma_{k+1}}x_r)^2}$$

$$\leq \sigma_{k+1} \max_{\|x\|_2=1} \| x \|_2$$

$$= \sigma_{k+1}.$$

Na realidade $\| A - A_k \|_2 = \sigma_{k+1}$, pois o máximo é atingido considerando o vetor $x = (0, \cdots, 0, x_{k+1} = 1, 0, \cdots, 0)$. ■

Teorema 2.3. *A_k é a matriz com posto k que mais se aproxima da matriz A relativamente à norma $\| \, . \, \|_2$.*

Demonstração: Seja B uma matriz real do tipo $m \times n$ e de posto k. Vamos mostrar que

$$\|A - B\|_2 \geq \sigma_{k+1}.$$

Basta mostra que existe um vetor unitário $z \in \mathbb{R}^n$ tal que $\|(A - B)z\|_2 \geq \sigma_{k+1}$. Como B tem posto k então existem $n - k$ vetores ortonormais em $\mathbb{R}^n$, digamos $\{b_1, \cdots, b_{n-k}\}$, que geram o espaço nulo da matriz B. Como $k < r \leq n$ então podemos escolher $k + 1$ vetores $v_1, \cdots, v_{k+1}$ de tal modo que

$$[b_1, \cdots, b_{n-k}] \cap [v_1, \cdots, v_{k+1}] \neq \{0\}.$$

Esta escolha é possível visto que o conjunto $\{b_1, \cdots, b_{n-k}, v_1, \cdots, v_{k+1}\}$ é linearmente dependente em $\mathbb{R}^n$. Assim, podemos escolher um vetor unitário z nesta interseção. Para este vetor z temos que $Bz = 0$, pois z pertence ao espaço nulo da matriz B, e como z também pertence ao espaço gerado pelos vetores $\{v_1, \cdots, v_{k+1}\}$ segue que:

$$Az = \left(\sum_{i=1}^{r} \sigma_i u_i v_i^t \right) z = \sum_{i=1}^{k+1} \sigma_i u_i (v_i^t z).$$

Desta forma,

$$\| A - B \|_2 = \max_{\|y\|=1} \| (A-B)y \|_2 \geq \| (A-B)z \|_2 = \| Az \|_2 .$$

Entretanto,

$$\| Az \|_2 = \| \sum_{i=1}^{k+1} \sigma_i u_i (v_i^t z) \|_2 \geq \sigma_{k+1} \| \sum_{i=1}^{k+1} u_i (v_i^t z) \|_2 .$$

Como $v_j^t z = 0$ para $j \geq k+2$ e $\| z \|_2 = 1$, temos:

$$\| \sum_{i=1}^{k+1} u_i (v_i^t z) \|_2 = \| \begin{bmatrix} u_1 & \cdots & u_{k+1} & u_{k+2} & \cdots & u_m \end{bmatrix} \begin{bmatrix} v_1^t \\ \vdots \\ v_{k+1}^t \\ 0 \\ \vdots \\ 0 \end{bmatrix} z \|_2 = \| UV^t z \|_2 = \| z \|_2 = 1.$$

Portanto,

$$\|A - B\|_2 \geq \sigma_{k+1}. \quad \blacksquare$$

O principal efeito de aproximarmos a matriz $A = UDV^t$ pela matriz A_k é a redução da dimensão do espaço, pois o posto de A_k é bem menor do que o posto da matriz A. Um outro efeito é o fato de que escolhendo os k maiores valores da matriz D, que são os valores singulares de A com maior magnitude, os demais valores de D podem ser considerados todos nulos, e assim, as linhas e colunas correspondentes das matrizes U e V^t contém apenas zeros. Este é o X da questão para formação do espaço semântico e cálculo de uma estimativa para similaridades entre palavras e/ou documentos.

2.5 COMO LSA FAZ O CÁLCULO DA SIMILARIDADE ENTRE TEXTOS

Considere a matriz fatorada $A = UDV^t$. Após a escolha dos k maiores valores da matriz D e eliminando-se as linhas e colunas nulas das matrizes U e V^t, podemos perceber que a matriz A_k é o produto dessas matrizes modificadas, ou seja,

$$A_k = U_k D_k V_k^t,$$

onde as matrizes U_k, D_k e V_k tem ordens $m \times k$, $k \times k$ e $n \times k$, respectivamente. É fácil ver que a matriz A_k tem ordem $m \times n$ e posto igual a k. Os vetores colunas da matriz A_k representam os vetores colunas da matriz A no espaço semântico. Entretanto, LSA não usa explicitamente a matriz A_k para calculo de similaridades.

Para calcularmos a similaridade entre palavras, que são os vetores linha da matriz A_k, basta considerarmos a matriz

$$U_k D_k,$$

que é uma matriz de ordem $m \times k$, onde cada linha i, com $1 \leq i \leq m$, corresponde a uma palavra do vocabulário.

Para calcularmos a similaridade entre textos, que são os vetores coluna da matriz A_k, basta considerarmos a matriz

$$D_k V_k^t,$$

que é uma matriz de ordem $k \times n$, onde cada coluna j, com $1 \leq j \leq n$, corresponde a um texto do corpus. Vejamos uma justificativa para este fato.

Denotemos por $(A_k)_1, \cdots, (A_k)_i, \cdots, (A_k)_m$ os vetores linha e por
$(A_k)^1, \cdots, (A_k)^j, \cdots, (A_k)^n$ os vetores coluna da matriz A_k.

Vamos calcular a similaridade entre os textos representados pelas colunas $(A_k)^i$ e $(A_k)^j$. Se θ é o ângulo entre os vetores $(A_k)^i$ e $(A_k)^j$, então a similaridade é dada por

$$\cos\theta = \frac{< (A_k)^i, (A_k)^j >}{\|(A_k)^i\| \|(A_k)^j\|}.$$

Sem perda de generalidade, podemos supor que $\|(A_k)^i\| = \|(A_k)^j\| = 1$; assim, a similaridade é simplesmente o produto interno $< (A_k)^i, (A_k)^j >$ entre os vetores $(A_k)^i$ e $(A_k)^j$.

Consideremos a matriz simétrica

$$A_k^t A_k = \begin{bmatrix} (A_k)_1 \\ \vdots \\ (A_k)_i \\ \vdots \\ (A_k)_n \end{bmatrix} \begin{bmatrix} (A_k)^1 & \cdots & (A_k)^j & \cdots & (A_k)^n \end{bmatrix}$$

É fácil verificar que $(A_k^t A_k)_{i,j} =< (A_k)^i, (A_k)^j >$. Entretanto, como U_k é uma matriz ortogonal temos que

$$\begin{aligned} A_k^t A_k &= (U_k D_k V_k^t)^t (U_k D_k V_k^t) \\ &= (V_k D_k U_k^t)(U_k D_k V_k^t) \\ &= (V_k D_k)(U_k^t U_k)(D_k V_k^t) \\ &= (D_k V_k^t)^t (D_k V_k^t) \end{aligned}$$

Isto justifica o fato do uso da matriz $D_k V_k^t$ ao invés da matriz A_k para o cálculo da similaridade entre os vetores coluna da matriz A_k. A justifica o fato de usarmos a matriz

$U_k D_k$ ao invés da matriz A_k para o cálculo da similaridade entre os vetores linha da matriz A_k é análogo.

Foi visto na seção 2.5 que LSA pode ser aplicado para estimar a similaridade entre dois textos. No Capítulo 4 será proposto um modelo LSA para avaliação automática de questões discursivas. Antes, no Capítulo 3 será feita uma revisão sistemática de modelos LSA aplicados em ambientes educacionais.

3 UMA REVISÃO SISTEMÁTICA DE MODELOS LSA APLICADOS EM AMBIENTES EDUCACIONAIS

Desde que foram criados, Deerwester et al. (1990), modelos LSA vem sendo utilizados como base para sistemas de avaliação automática com aplicações em ambientes educacionais. Muitos trabalhos utilizaram modelos LSA com resultados satisfatórios em ambientes específicos; desta forma será feita uma analise dos resultados destes trabalhos relacionados, verificando suas metodologias e parâmetros utilizados. Esta analise se dará por meio de uma revisão bibliográfica sistemática como sugerido no trabalho de Russell et al. (2009). De acordo com Kitchenham (2004), uma revisão sistemática envolve algumas atividades em etapas distintas: Identificação da pesquisa, Estudos primários, Identificação da necessidade de Revisão, Protocolo de Revisão e Análise da revisão da literatura.

3.1 IDENTIFICAÇÃO DA PESQUISA

Esta pesquisa pode ser identificada como parte do campo de pesquisa denominado "Avaliação Assistida por Computador" (Computer Assisted Assessment) como sugerido por Whittington e Hunt (1999). Dentro desta linha, foi proposto o desenvolvimento de um sistema para avaliação automática de respostas a questões discursivas.

3.2 ESTUDOS PRIMÁRIOS

Pesquisas em "Avaliação Assistida por Computador" já vem ocorrendo desde a década de 60. O projeto pioneiro foi o Project Essay Grader (PEG) em um trabalho de Page (1966). O sistema PEG fazia análise do estilo de escrita e pontuava o texto pela sua qualidade e não pelo seu conteúdo[1]. Os primeiros resultados não foram satisfatórios em termos de precisão e o projeto foi alvo de muitas críticas por parte comunidade acadêmica. Esses resultados fizeram com que a pesquisa nessa área ficasse quase estagnada até a década de 90. A partir da década de 90 com o avanço de técnicas de processamento de linguagem natural e recuperação de informação pesquisas voltaram a ser desenvolvidas. Neste período surgiram sistemas de avaliação automática de textos com a utilização de várias técnicas, como por exemplo, medida de qualidade da escrita, técnicas de processamento de linguagem natural

[1] O PEG considerava a coesão e a coerência e não considerava as informações contidas no texto

e técnicas estatísticas.

Em 1988 o trabalho de Landauer et al. (1988) publicou Latent Semantic Analysis (LSA) que é uma técnica estatístico matemática para extração e inferência de relações de contexto em passagens do discurso. Estas relações tornam possível uma comparação de dois trechos de texto usando uma medida de similaridade. O trabalho de Deerwester et al. (1990) explica a nova teoria e afirma que a abordagem LSA diminui as deficiências da abordagem termo-frequência[2]

3.3 IDENTIFICAÇÃO DA NECESSIDADE DE REVISÃO

A abordagem LSA tem sido utilizada para avaliação automática de textos produzindo resultados próximos de avaliadores humanos. O que se pretende é reunir e avaliar trabalhos que usam a abordagem LSA aplicados em ambientes educacionais. Na literatura pesquisada ainda não foram encontrados resultados que permitam sem esforço de pesquisa a construção para uso prático de sistemas para avaliação automática de respostas a questões discursivas baseados em modelos LSA, em especial para a língua portuguesa.

3.4 PROTOCOLO DE REVISÃO

O protocolo que serviu como guia para esta revisão sistemática foi constituído por um conjunto de atividades visando melhorar a estruturação da pesquisa:

i) especificar os objetivos;

ii) definir as questões de pesquisa;

iii) definir os critérios de seleção dos artigos;

iv) compilar os dados.

Esta revisão sistemática tem o propósito de alcançar os seguintes objetivos:

1. Identificar os artigos e trabalhos científicos relacionados com avaliação automática de respostas a questões discursivas baseados em modelos LSA;

2. Analisar as abordagens e metodologias utilizadas pelos trabalhos identificados no item anterior;

[2] Segundo Baeza-Yates e Ribeiro-Neto (2013) esta abordagem representa um documento por meio do conjunto completo de palavras que os compõe.

3. Analisar os resultados obtidos.

A questão de pesquisa estabelecida foi a seguinte:

"Qual o estado da arte no uso da abordagem LSA para avaliação automática de questões discursivas ?"

Além desta questão principal, outras questões secundárias foram identificadas:

"Quando, quem e onde foi realizada cada pesquisa ?

"Que as áreas do conhecimento foram utilizadas no corpus de cada pesquisa ?

"Quais os detalhes técnicos de LSA utilizados por cada pesquisador ?

"Quais as melhores acurácias alcançadas ?

Outra etapa deste protocolo de revisão é definir os critérios para seleção dos artigos. Foram estabelecidos os seguintes critérios:

- Consultas via web (em bases abertas para a comunidade acadêmica);
- Artigos em inglês e espanhol;
- Disponibilidade de obter na íntegra os artigos por meios de buscas na internet.

Para realização das buscas foi utilizado apenas o método automático com procedimento de consultas às bases internacionais. Com base na questão de pesquisa principal identificamos as seguintes palavras-chave:

Em inglês: automatic, evaluation, latent, semantic, analysis, essay, questions;

Em espannhol: automática, evaluación, latente, semántica, análisis, ensayo, temas,

Executadas as pesquisas, a seleção dos artigos foi realizada após leitura do resumo (abstract): foram relacionados principalmente trabalhos que incluem esforços de pesquisa em LSA com sobreposição de aplicações educacionais. É claro que a pesquisa não ficou restrita ao escopo das interseções de LSA com aplicações educacionais. Foram também relacionados os trabalhos mais citados na pesquisa; alguns destes trabalhos foram os primeiros trabalhos explicando a teoria básica de LSA. Finalmente, foram relacionados trabalhos que apresentaram novas informações sobre LSA. Esta pesquisa resultou em

57 trabalhos cujos títulos e resumos estão no Apêndice A. Vale ressaltar que destes 57 trabalhos apenas 18 tratam de experimentos cujo foco é avaliação automática de ensaios escritos por estudantes numa comparação com avaliadores humanos. A última fase deste protocolo é compilar os dados extraídos dos trabalhos relacionados. Os resultados obtidos foram dados quantitativos e qualitativos e serão analisados na próxima seção.

3.5 ANÁLISE DA REVISÃO DA LITERATURA

Após a seleção de todos trabalhos realacionados, a Figura 2 apresenta os números de publicações por ano destes trabalhos desde 1988 até 2015.

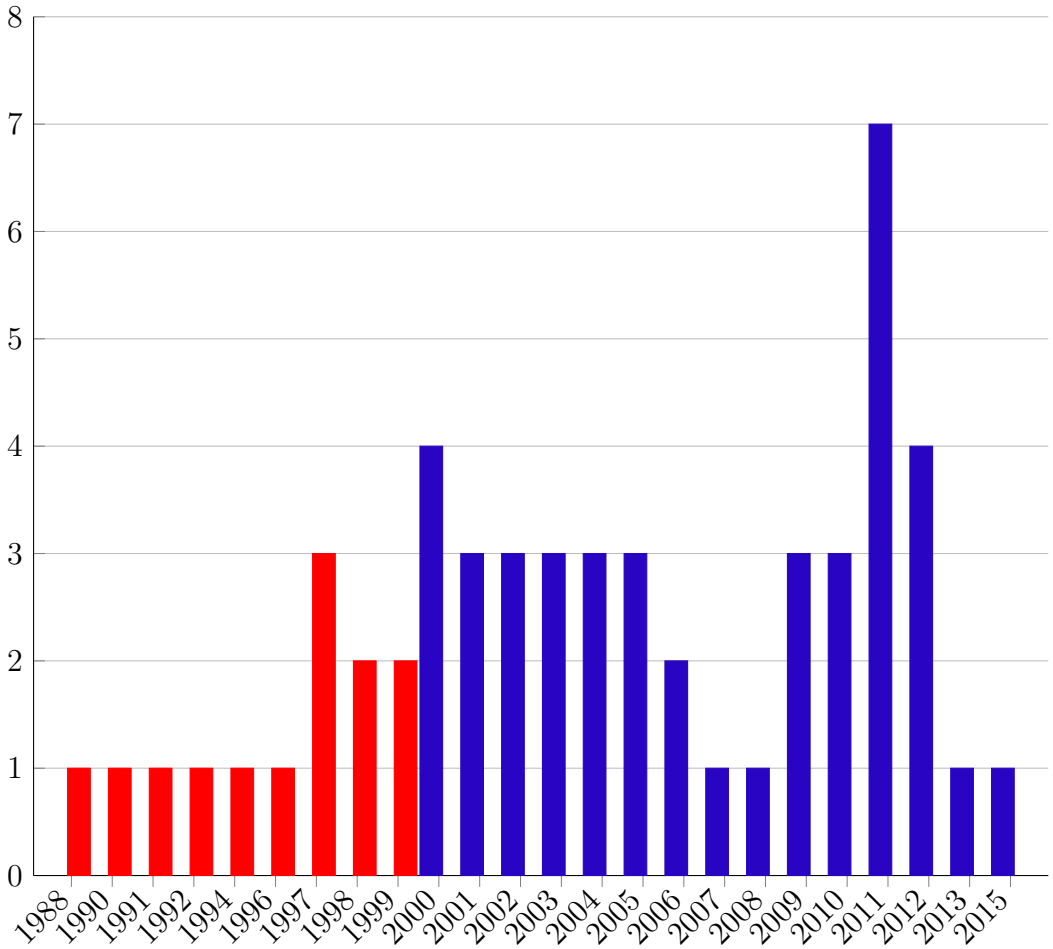


Figura 2 – Publicações anuais

A Figura 2 nos mostra que tivemos uma média de aproximadamente 02 (dois) trabalhos publicados anualmente. O trabalho de Landauer et al. (1988) com o título *Information Retrieval using a Singular Value Decomposition Latent Semantic Struture* apresenta a técnica LSA. Até o ano de 1995 todos os trabalhos tem foco principal na recuperação de informações de textos, sendo que o trabalho de Dumais (1991) inova com o uso de funções de peso. O trabalho de Rehder et al. (1998) menciona pela primeira vez que LSA pode

comparar um ensaio escrito por um aluno com um texto de referência. O trabalho de Foltz, Laham e Landauer (1999) apresenta o sistema The Intelligent Essay Assessor que é um software que pontua a qualidade do conteúdo de ensaios escritos por alunos. O trabalho de Wiemer-Hastings (2000) diz que não existe um consenso sobre a importância da informação sintática na representação do conhecimento e propõe a adição de informação sintática para LSA. O trabalho de Steinhart (2001) é uma tese de doutorado que descreve a evolução e testes para implementação do Summary Street, um sistema de tutoria inteligente que usa LSA para apoiar atividades de escrita e revisão. Os trabalhos de Nakov, Popova e Mateev (2001) e Hu et al. (2003), respectivamente, discutem o papel dos parâmetros, função peso e dimensionalidade na performance de um modelo LSA. Já o trabalho de Kanejiya, Kumar e Prasad (2003) apresenta o sistema SELSA que considera a palavra e sua vizinhança sintática. O trabalho de Marinrez et al. (2005) combina LSA com técnicas de procedimento de linguagem natural para avaliação de textos livres. O trabalho de Wild et al. (2005) mostra como cada parâmetro LSA influencia na pontuação automática de um ensaio e Diana et al. (2005) apresenta uma avaliação comparativa entre um algoritmo baseado na abordagem n-gramas e LSA. O trabalho de He, Hui e Quan (2009) propõe uma abordagem que combina LSA com n-gramas. O trabalho de Jorge-Botana et al. (2010) discute a busca de melhores parâmetros para melhorar a eficácia de um modelo LSA em comparação com especialistas humanos. O ano de 2011 foi o período com o maior número de publicações num total de 7(sete). Destacamos alguns destes trabalhos: Fernandes, Artifice e Fonseca (2012) propõe uma fórmula para estimar a dimensão do espaço semântico de LSA; Klein, Kyrilov e Tokman (2011) descreve a concepção, implementação e avaliação de um sistema de avaliação automática, com base em LSA, para pontuação de questões abertas em ciência da computação; Olmos et al. (2011) combina LSA com o modelo linguístico Rouge-N; Ramachandran e Gehringer (2011) usa um modelo LSA para classificar comentários de revisões através da qualidade e do tom de cada um deles; Zen, Iskandar e Linang (2011) faz uma adaptação de LSA para classificar scrits de linguagens de programação. Em 2012 tivemos quatro trabalhos publicados e todos merecem destaque: o trabalho de Chali e Hasan (2012) usa um modelo LSA para avaliar respostas escritas por alunos de um curso de Terapia Ocupacional; o trabalho de Layfield (2012) afirma que quanto maior o número de textos melhor o desempenho de um modelo LSA; o trabalho de Islan e Hoque (2012) apresenta o sistema GLSA (LSA generalizada), onde considera n-gramas na matriz termo-documento e o trabalho Refaat et al. (2012) avalia automaticamente respostas de textos livres em língua árabe. Em 2013, o trabalho de Leon et al. (2013) avalia sumários escritos por estudantes espanhóis usando um método baseado em uma equação de regressão que incorpora semelhança semântica e comprimento do vetor. Em 2015, o trabalho de Jorge-Botana et al. (2015) discute algumas considerações sobre avaliação automática de sumários no ensino à distância.

Quanto aos países onde os trabalhos foram publicados, a Figura 3 mostra que quase 40 % dos trabalhos foram publicados nos EUA, onde a técnica LSA foi apresentada inicialmente. A Espanha aparece com um pouco mais de 10% dos trabalhos publicados, seguido da Inglaterra com 9%.

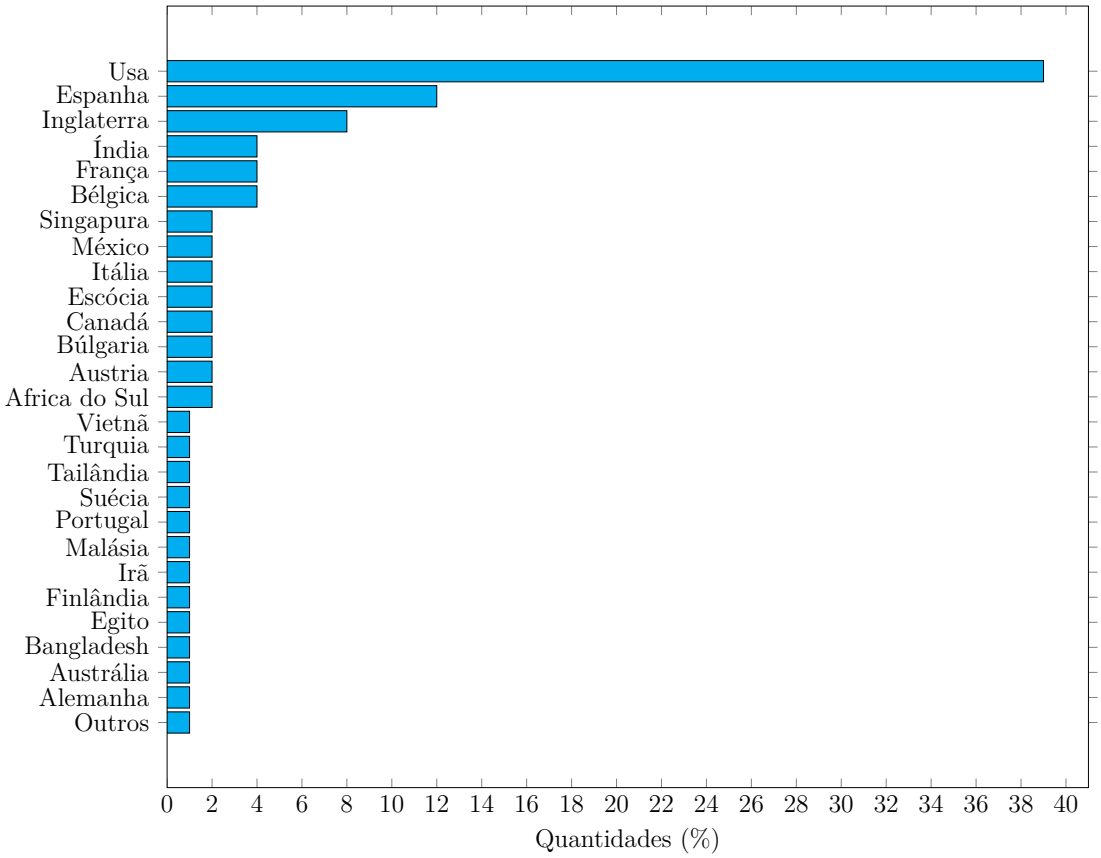


Figura 3 – Países onde os trabalhos foram publicados

A Figura 3 mostra que trabalhos relacionados com LSA foram publicados em todos os continentes. Entretanto, na América latina menos de 2% do número total de publicações: dois trabalhos na Universidad Nacional Autónoma de México. Em língua portuguesa foi publicado apenas o trabalho de Fernandes, Artifice e Fonseca (2012) que propõe uma fórmula para estimar a dimensão do espaço semântico. No Brasil não existe nenhum estudo para implementação de um modelo baseado em LSA para avaliação automática de respostas escritas.

A técnica LSA foi apresentada inicialmente nos EUA no trabalho de Landauer et al. (1988) na University of Colorado; assim é razoável esperar que este grupo de pesquisa seja responsável pelo maior número de publicações. A Tabela 2 mostra os pesquisadores com o maior número de publicações. De 57 trabalhos publicados sobre LSA foi verificado que oito pesquisadores concentraram mais de 70 % do total destes trabalhos.

Pesquisador	Universidade	Qtd de publicações
Thomas K Landauer	University of Colorado	10
Susan T Dumais	University of Indiana	6
Peter W Foltz	New Mexico State University	5
Darrell Laham	University of Colorado	5
Walter Kintsch	University of Colorado	4
Ricardo Olmos	Universidad Autónoma de Madrid	4
Guillermo Jorge-Botana	Universidad Autónoma de Madrid	4
Inmaculada Escudero	Universidad Autónoma de Madrid	4

Tabela 2 – Pesquisadores mais produtivos

Nos últimos quatro anos vem se destacando o grupo de pesquisa coordenado pelo professor Guillermo Jorge-Botana na Universidad Autónoma de Madrid. Este grupo de pesquisa (*Grupo de Interés en el Análises de la Semántica Latente*) produziu o Gallito 2.0, que é uma ferramenta de procedimento de linguagem natural baseada em LSA. O propósito principal desta ferramenta é solucionar diversos problemas de classificação, categorização e avaliação de textos acadêmicos em diversas áreas do conhecimento.

A Figura 4 mostra os principais sub-áreas do conhecimento que serviram como domínios de corpus utilizados pelos pesquisadores.

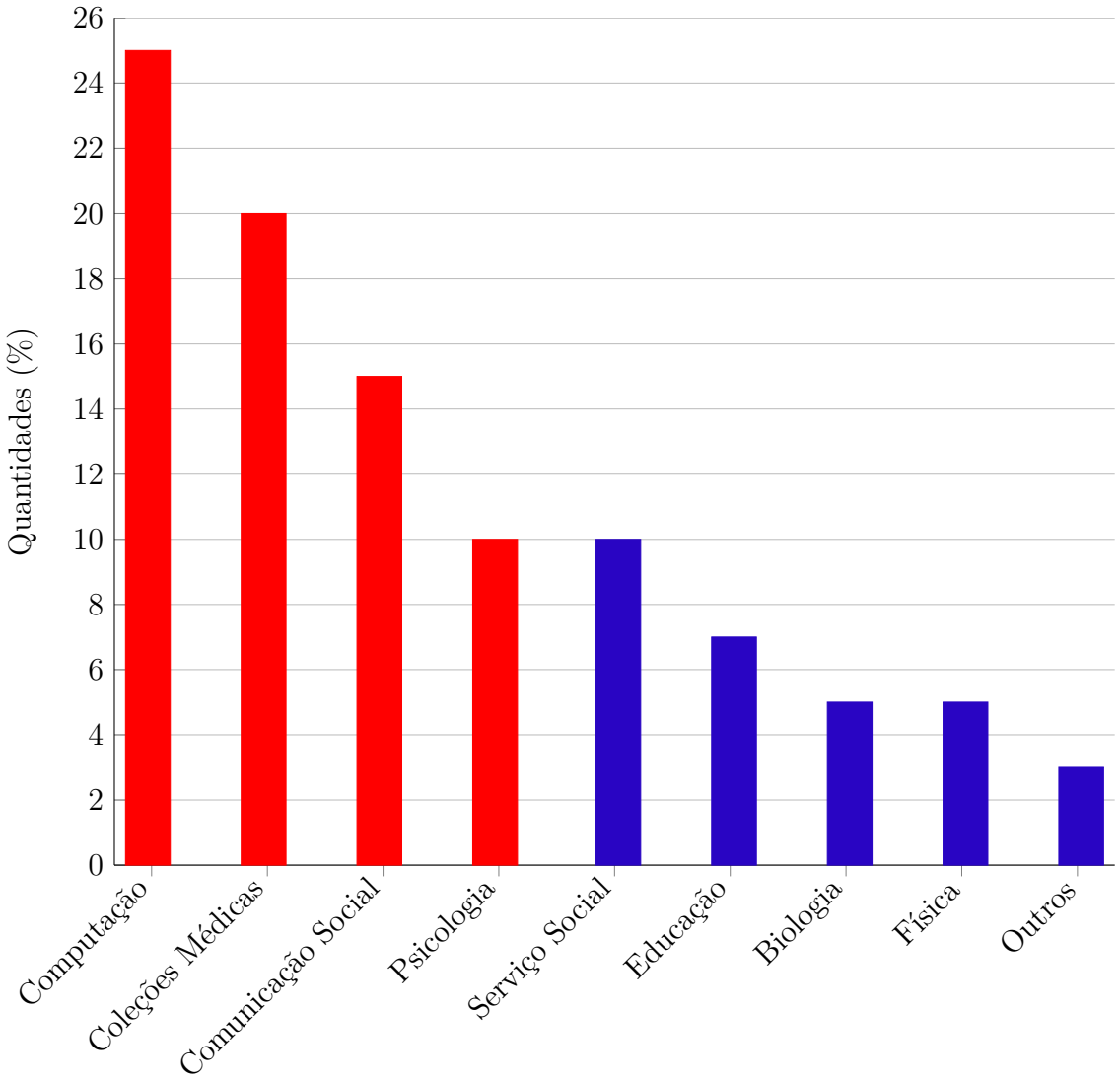


Figura 4 – Principais sub-áreas do conhecimento que foram objetos de pesquisa

A Figura 4 mostra que Ciência da Computação foi o domínio do corpus mais utilizada; entretanto domńios com especificidades como Biologia e Física já são utilizadas como corpus de pesquisa.

Tipicamente, a arquitetura de um sistema LSA compreende as etapas de pré-processamento construção da matriz, pesagem, SVD, redução para o espaço semântico e medidas de similaridades. Na etapa de pré-processamento é frequente o uso de técnica de pré-processamento de textos com objetivo de melhorar os índices de similaridade entre os textos comparados. As principais técnicas de pré-processamento utilizadas são mostradas na Figura 5.

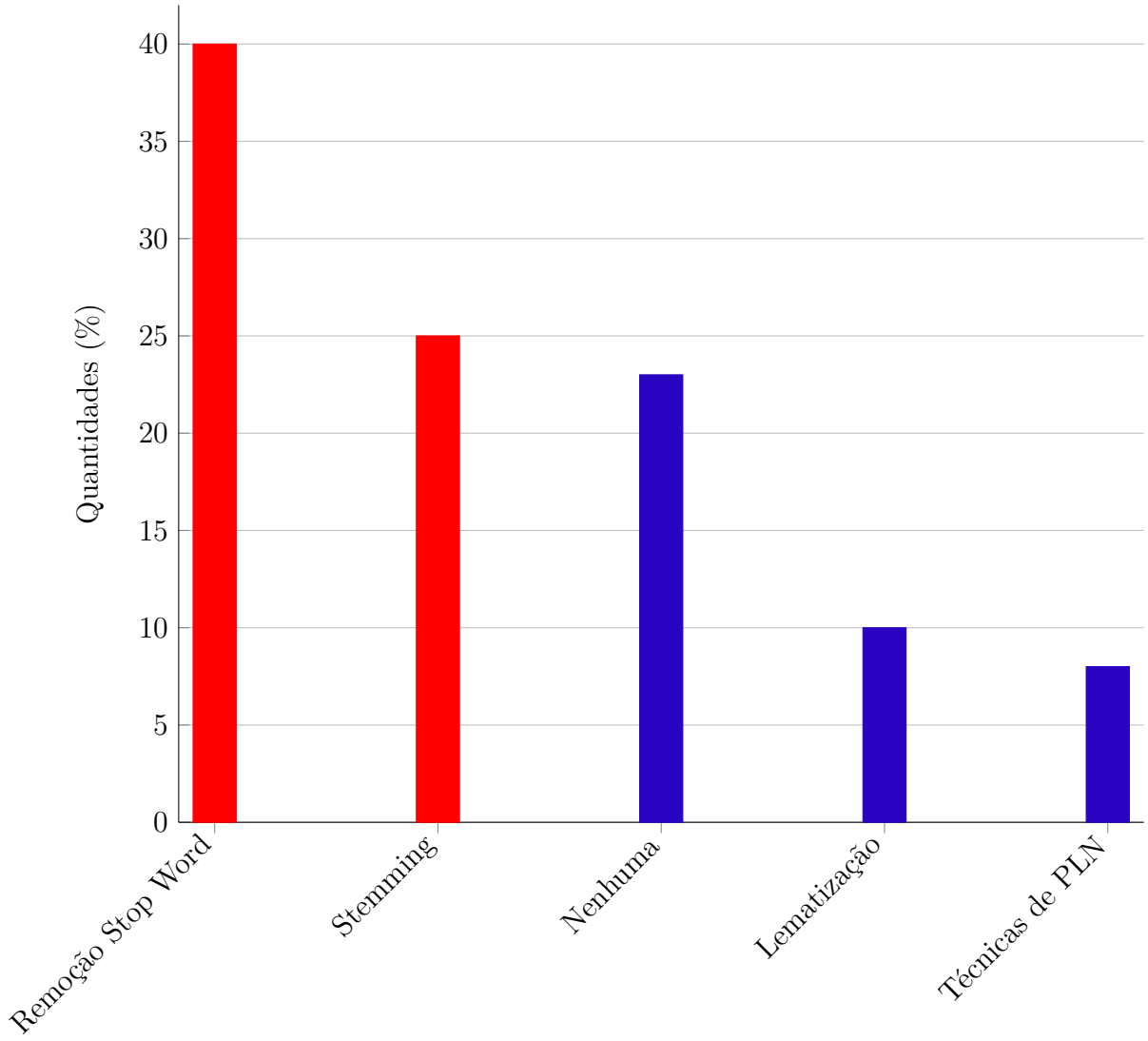


Figura 5 – Técnicas utilizadas durante o pré-processamento

A Figura 5 mostra que remoção de stop word foi a técnica mais utilizada; entretanto vemos também que mais de 20% dos trabalhos não utilizaram nenhuma técnica durante o pré-processamento.

A hipótese de Luhn, descrita no trabalho de Baeza-Yates e Ribeiro-Neto (2013), afirma que a importância de uma palavra em um texto é diretamente proporcional a sua frequência no próprio texto. Este fato é conhecido como fator ou função ponderação. A Figura 6 mostra que mais de 40 % dos trabalhos utilizaram simplesmente *termo-frequência* como função de ponderação.

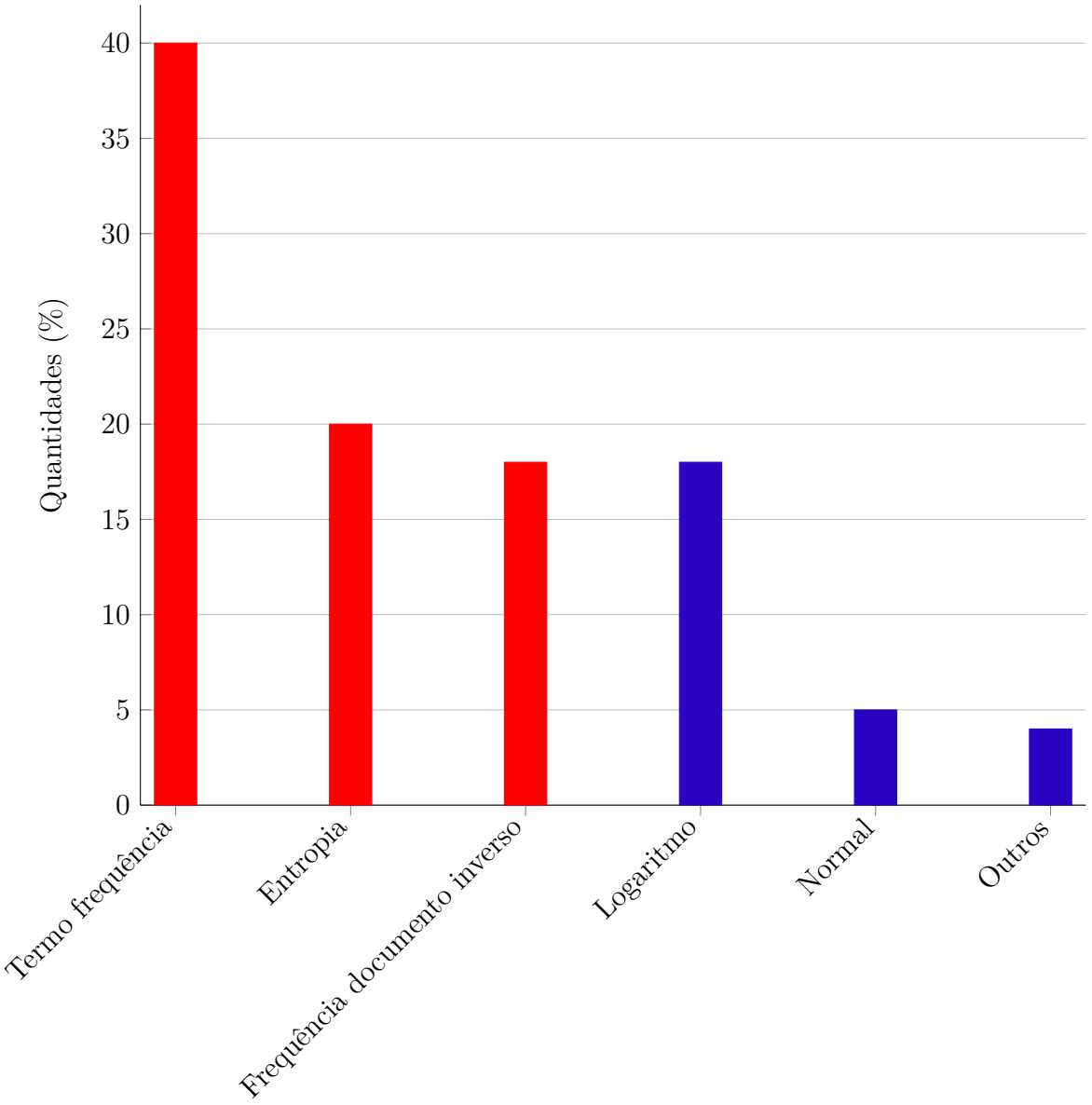


Figura 6 – Funções ponderações utilizadas

A dimensão do espaço semântico é o parâmetro que mais influência na eficácia de um modelo LSA e ainda é uma questão em aberto. Embora o trabalho de Fernandes, Artifice e Fonseca (2012) tenha proposto uma fórmula para estimar esta dimensão, na maioria dos trabalhos esta dimensão foi determinada empiricamente considerando sempre o número de textos que compõem o corpus de pesquisa. A Tabela 3 mostra as dimensionalidades mais utilizadas pelo trabalhos relacionados nesta pesquisa.

dimensionalidade	Quantidade
2	9
10	3
50	5
100	12

Tabela 3 – Dimensionalidades mais utilizadas

Como LSA é um modelo vetorial, o grau de similaridade entre dois textos é dado pela correlação entre os vetores que representam cada um dos textos. A melhor maneira de quantificar esta relação é através do cálculo do cosseno entre os vetores. A Figura 7 mostra que o cosseno foi disparadamente a medida de similaridade mais utilizada pelos trabalhos relacionados nesta revisão.

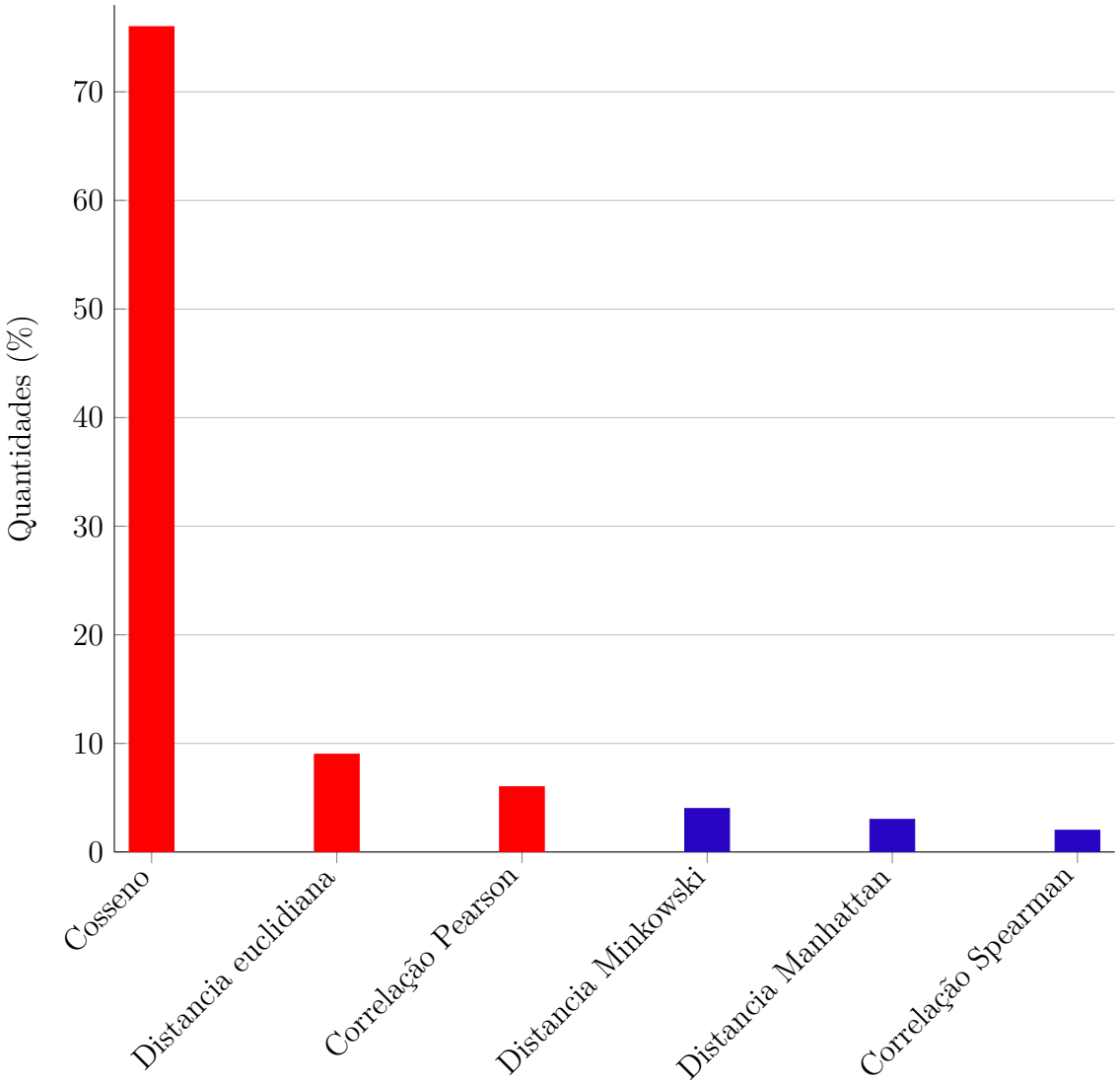


Figura 7 – Medidas de similaridades mais utilizadas

O trabalho de Jorge-Botana et al. (2010) afirma que distância euclidiana é uma medida que combina o cosseno do ângulo e o comprimento do vetor e pode corrigir pontuações de respostas com um número reduzido de palavras. Experimentos foram realizados com a distancia euclidiana para esta finalidade. Entretanto, como os resultados não foram satisfatórios, optou-se pela utilização de um outro fator de penalização para corrigir este problema: a distância entre dois vetores pode ser grande, mesmo que os textos representados por estes vetores tenham uma distribuição similar entre as palavras.

Para finalizar este seção faremos uma breve análise da questão de pesquisa. Na revisão sistemática, aqui apresentada, dos 57 trabalhos que relatam modelos LSA aplicados em ambientes educacionais apenas 18 tratam especificamente de avaliação automática de ensaios escritos por alunos. Numa comparação com avaliação humana, o menor índice de acurácia foi de 50% e o maior de 80% sendo o índice médio de 76,40%. A meta a ser alcançada é que seja possível desenvolver um sistema de avaliação automática baseado em LSA que possa atingir uma acurácia de mais de 90% tendo desempenho similar ao de avaliadores humanos.

Testes revelam que identificar a calibragem ideal é um processo complicado e está ligado diretamente com a finalidade da aplicação do modelo LSA. No próximo capítulo será vista a metodologia do processo de calibração do modelo LSA proposto

4 O MODELO LSA PROPOSTO

Neste capítulo será visto como foi desenvolvida a metodologia e todo o processo de calibração de um modelo LSA para uso prático em avaliação automática de respostas escritas. O processo metodológico foi desenvolvido utilizando uma abordagem empírica da seguinte maneira:

1. Digitalização das respostas;
2. Representação matricial do corpus;
3. Pré-processamento das respostas;
4. Aplicação da função ponderação;
5. Cálculo da SVD e escolha da dimensionalidade;
6. Classificação das respostas;
7. Calibração do modelo LSA;
8. Cálculo da acurácia.

4.1 DIGITALIZAÇÃO DAS RESPOSTAS

As respostas selecionadas passaram por um processo de digitalização manual onde foram feitas correções com um corretor ortográfico automático, mas não foi feito qualquer tipo de correção de concordância gramatical.

A Figura 8 mostra uma resposta escrita para a questão de Biologia.

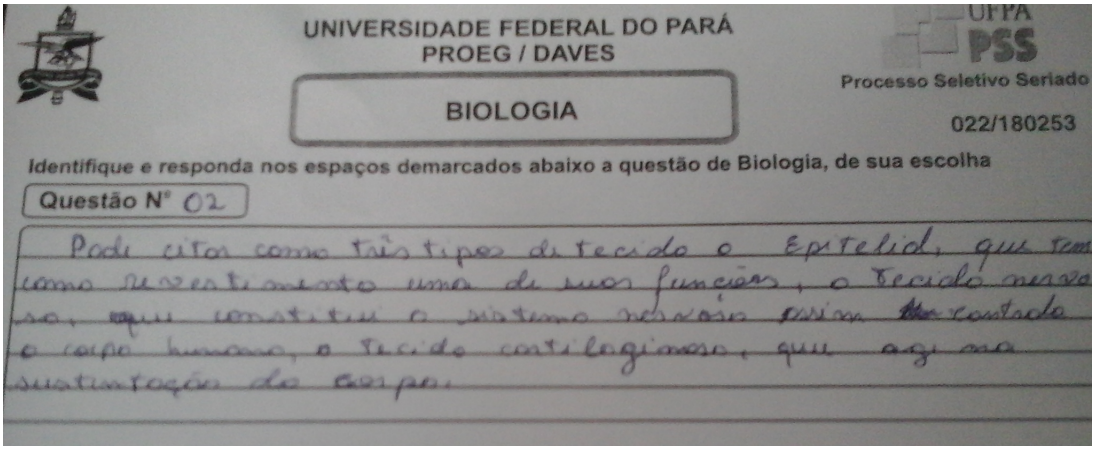

UNIVERSIDADE FEDERAL DO PARÁ
PROEG / DAVES
UFPA PSS
Processo Seletivo Seriado
BIOLOGIA
022/180253
Identifique e responda nos espaços demarcados abaixo a questão de Biologia, de sua escolha
Questão Nº 02
Pode citar como três tipos de tecido o Epitelial, que tem como revestimento uma de suas funções, o Tecido nervoso, que constitui o sistema nervoso assim controla o corpo humano, o Tecido cartilaginoso, que age na sustentação do corpo.

Figura 8 – Resposta escrita para a questão de Biologia

Abaixo temos a digitalização manual para esta resposta:

> Pode citar como três tipos de tecido o epitelial, que tem como revestimento uma de suas funções, o tecido nervoso, que constitui o sistema nervoso assim controla o corpo humano, o tecido cartilaginoso, que age na sustentação do corpo.

Pode ser observado que durante o processo de digitalização da resposta escrita apresentada na Figura 8 não foi necessária nenhuma correção ortográfica e foram desconsiderados os erros de concordância gramatical e adequação de termos.

A Figura 9 mostra uma resposta escrita para a questão de Geografia.

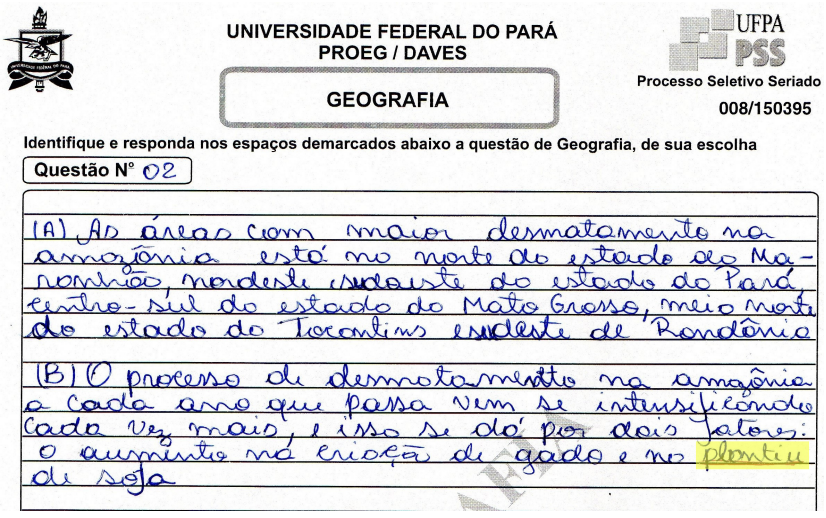

UNIVERSIDADE FEDERAL DO PARÁ
PROEG / DAVES
UFPA PSS
Processo Seletivo Seriado
GEOGRAFIA
008/150395
Identifique e responda nos espaços demarcados abaixo a questão de Geografia, de sua escolha
Questão Nº 02
(A) As áreas com maior desmatamento na amazônia está no norte do estado do Maranhão, nordeste sudoeste do estado do Pará centro-sul do estado do Mato Grosso, meio norte do estado do Tocantins sudeste de Rondônia
(B) O processo de desmatamento na amazônia a cada ano que passa vem se intensificando cada vez mais, e isso se dá por dois fatores: o aumento na criação de gado e no plantio de soja

Figura 9 – Resposta escrita para a questão de Geografia

Abaixo temos a digitalização manual para esta resposta:

> As áreas com maior desmatamento está no norte do estado do Maranhão, nordeste, sudoeste do estado do Pará, centro-sul do estado do Mato Grosso, meio norte do estado do Tocantins e sudeste de Rondônia. O processo de desmatamento na Amazônia a cada ano que passa vem se intensificando cada vez mais, e isso se dá por dois fatores: o aumento na criação de gado e no **plantio** de soja.

Pode ser observado que durante o processo de digitalização da resposta escrita apresentada na Figura 9 foi necessário fazer uma correção ortográfica: plantiu por plantio. A correção se faz necessária pois, em caso contrário, seria acrescentada uma nova linha na matriz que representa o corpus, o que implica em um aumento de entradas nulas da matriz, que pode ser traduzido como provável perda de informações, além do aumento no custo computacional do modelo. Foram desconsiderados os erros de concordância gramatical e adequação de termos.

As respostas após digitalizadas foram armazenadas em um sistema de gerenciamento de banco de dados. É a partir deste banco de dados que obtemos uma representação matricial do corpus da pesquisa.

4.2 REPRESENTAÇÃO MATRICIAL DO CORPUS

A representação matricial do corpus é o ponto de partida qualquer modelo LSA. Suponha que o corpus seja constituído por apenas três respostas dadas para a questão de Biologia, que chamaremos de **Resposta 1**, **Resposta 2**, **Resposta 3**, respectivamente:

> **Resposta 1**: Pode citar como três tipos de tecido o epitelial, que tem como revestimento uma de suas funções, o tecido nervoso, que constitui o sistema nervoso assim controla o corpo humano, o tecido cartilaginoso, que age na sustentação do corpo.
>
> **Resposta 2**: Tecido muscular esquelético responsável pela sustentação dos Tecido epitelial proteção e revestimento Tecido muscular cardíaco bombeamento de fluxo sanguíneo, circulação sanguínea.
>
> **Resposta 3**: Tecido Adiposo Absorção de impactos mecânicos. Tecido ósseo sustentação Tecido epitelial Revestimento do corpo e controle de temperatura.

Para que se possa obter uma representação matricial deste corpus a primeira coisa que deve ser feita é construir o vocabulário deste corpus. Pode-se verificar sem dificuldades que o vocabulário deste corpus é dado por:

$$V = \{age, assim, Absorção, Adiposo, bombeamento, cartilaginoso, citar, como, controle, constitui, controla, corpo, cardíaco, circulação, de, do, dos, e, epitelial, esquelético, funções, fluxo, humano, impactos, mecânicos, muscular, na, nervoso, o, ósseo, Pode, pela, proteção, que, revestimento, responsável, sustentação, sanguínea, sanguíneo, sistema, suas, tecido, tem, tipos, três, temperatura, uma\}$$

Foi verificado que o vocabulário do corpus é constituído por 47 palavras, sendo que 13 são stopwords: *como, de, do, dos, e, na, o, pode, pela, que, suas, tem, uma* . Aplicando-se a técnica de **remoção de stopwords** para as três respostas em questão, a matriz que representará o corpus será do tipo 34×3 e sua construção é feita da seguinte maneira: cada entrada (i, j) da matriz é a frequência com que a palavra i, $1 \leq i \leq 34$ aparece na resposta j, $1 \leq j \leq 3$. Assim, a matriz A que representa o corpus é dada por:

$$A = \begin{bmatrix} & \text{Resposta 1} & \text{Resposta 2} & \text{Resposta 3} \\ age & 1 & 0 & 0 \\ assim & 1 & 0 & 0 \\ Absorção & 0 & 0 & 1 \\ Adiposo & 0 & 0 & 1 \\ bombeamento & 0 & 0 & 1 \\ cartilaginoso & 1 & 0 & 0 \\ citar & 1 & 0 & 0 \\ controle & 0 & 0 & 1 \\ constitui & 1 & 0 & 0 \\ controla & 1 & 0 & 0 \\ corpo & 2 & 0 & 1 \\ cardíaco & 0 & 1 & 0 \\ circulação & 0 & 1 & 0 \\ epitelial & 1 & 1 & 1 \\ esquelético & 0 & 1 & 0 \\ funções & 1 & 0 & 0 \\ fluxo & 0 & 1 & 0 \\ humano & 1 & 0 & 0 \\ impactos & 0 & 0 & 1 \\ mecânicos & 0 & 0 & 1 \\ muscular & 0 & 2 & 0 \\ nervoso & 2 & 0 & 0 \\ ósseo & 0 & 0 & 1 \\ proteção & 0 & 1 & 0 \\ revestimento & 1 & 1 & 1 \\ responsável & 0 & 1 & 0 \\ sustentação & 1 & 1 & 1 \\ sanguínea & 0 & 1 & 0 \\ sanguíneo & 0 & 1 & 0 \\ sistema & 1 & 0 & 0 \\ tecido & 3 & 3 & 3 \\ tipos & 1 & 0 & 0 \\ três & 1 & 0 & 0 \\ temperatura & 0 & 0 & 1 \end{bmatrix}$$

A matriz que representa o corpus é frequentemente chamada **matriz termo-documento**. Diversos modelos utilizaram tão somente a matriz-termo documento, estes modelos podem ser chamados de LSA puros. Entretanto, outros modelos fizeram uso de aplicação de

alguma técnica como procedimento para pré-processamento do corpus. Foram utilizadas algumas dessas técnicas neste trabalho.

4.3 PRÉ-PROCESSAMENTO DAS RESPOSTAS

Durante o processo de representação matricial do corpus foi feito o pré-processamento das respostas através da combinação técnicas de $n-$gramas, remoção de stop words e aplicação de um processo de stemming. Essas técnicas foram aplicadas em todas as respostas escritas que constituem o corpus da pesquisa. Nas três próximas subseções será feita uma breve incursão sobre estas técnicas.

4.3.1 n-gramas

Um *n-grama* é uma subsequência de n elementos de uma dada sequência. Frequentemente 1-grama é chamado de unigrama, 2-gramas de bigrama, 3-gramas de trigrama.

No modelo proposto o banco de dados é composto por respostas escritas; foram considerados unigramas e bigramas de todas estas respostas. O uso de unigramas é mais frequente em modelos LSA, o que caracteriza LSA como um *saco de palavras*: não podemos trocar a ordem das palavras do texto sem modificar os resultados. O uso de bigramas foi com o propósito de tornar o sistema mais robusto, o que será visto com mais detalhes no próximo capítulo.

4.3.2 Remoção de stop words

A técnica de *remoção de stop words* tem como objetivo principal eliminar termos que não são semanticamente representativos nos textos. As stop words tipicamente são artigos, preposições, conjunções pronomes e advérbios que ao mesmo tempo são termos pouco frequentes em textos e que carregam pouca informação semântica. Alega-se que stop words pouco contribuem para o contexto ou informações do texto e que podem ser removidas durante o processamento, o que reduz consideravelmente a velocidade de qualquer processamento computacional de análise de textos [Lo e Ounis (2005)]. O programa usa uma lista de stop word do português com 299 palavras.

4.3.3 Stemming

O uso de técnicas para normalização de variações linguísticas são frequentes em processos computacionais para análise de textos. O processo de *stemming (ou radicalização)* combina as diferentes formas de uma palavra em uma representação comum que é o radical ou *stem*

da palavra [Orengo e Huyck (2001)]. As palavras *ácido* e *acidez* tem significados diferentes a partir da mesma raiz:

$$\text{stemming(ácido)=stemming(acidez)=acid.}$$

Nos experimentos foi utilizado o módulo RSLPStemmer do NLTK/Python.

Tanto a aplicação de stemming como a retirada de stopwords proporcionam uma compactação das respostas pela remoção de partes menos significativas, entretanto estes processos podem levar a perda de informações. Em geral esta perda de informações não é significativa na análise dos resultados obtidos.

No modelo LSA proposto durante o pré-processamento a metodologia adotada foi combinar as técnicas de $n-$gramas, remoção de stop words e stemming para representação matricial do corpus. Isto resultou na possibilidade de seis tipos de pré-processamentos para cada um dos conjuntos de respostas:

- unigramas com stop words;
- unigramas sem stop words;
- unigramas sem stop words + stemming;
- bigramas com stop words;
- bigramas sem stop words;
- bigramas sem stop words + stemming;

A implementação da etapa de pré-processamento em linguagem python foi feita da seguinte maneira:

```
1. arquivo = "geo.xml"
2. qtd = 230
3. respostas=processaTodasRespostas(arquivo,qtd,0,removeStopW = False,
   pStemmer = False, tokeniza = False, metodoNltk = False, bigramas = False,
   trigramas = False)
4. vocabulario = uniaoSublistas(respostas)
```

Na linha 1, a variável *arquivo* recebe o arquivo constituído pelas respostas a serem processadas; na linha 2, a variável *qtd* recebe a quantidade de respostas da variável *arquivo*; na linha 3, temos a variável *respostas* que recebe a função *processaTodasRespostas* com nove parâmetros: as varáveis *arquivo* e *qtd*, a opção *0* significa não comparar a resposta de referência com ela mesma durante a etapa de classificação, a opção *removeStopW* é para remoção (TRUE) ou não (FALSE) das stop words, a opção *pStemmer* é para aplicação ou não do algoritmo de stemming, a opção *tokeniza* faz a tokenização que decompõe cada resposta em cada palavra que a compõe, as opções são métodos do NLTK para formação de bigramas e trigramas das respostas; na linha 4. a variável *vocabulario* recebe a união da quantidade de respostas para formar o vocabulário do modelo LSA proposto para cada um dos dois conjuntos de respostas.

É claro que para que o pré-processamento fosse realizado foi necessário a importação de alguns módulos da linguagem python:

```
1. import nltk
2. from nltk.corpus import stopwords
3. from nltk.tokenize import sent_tokenize
4. from nltk.stem import RSLPStemmer as stemmer
5. from nltk import bigrams
6. from nltk import trigrams
```

Na linha 1, foi importado o módulo NLTK (Natural Language ToolKit); na linha 2, foi importado o pacote do NLTK que se refere as stop words do português; na linha 3, foi importado o módulo do KLTK que possibilita a tokenização das respostas; na linha 4, foi importado o algoritmo RSLPStemmer; na linha 5, foi importada a função que compõe bigramas do NLTK; na linha 6, foi importada a função que compõe trigramas do NLTK.

A partir do vocabulário foi construída a matriz termo-documento com o código abaixo:

```
#Construção da matriz inicial para unigramas

1. l = []
2. listaPalavra = []
3. for word in vocabulario:
4.       for i in range(0, len(respostas)):
5.              l.append(ContaFreqPalavra(respostas[i], word))
6.       listaPalavra.append(l)
7.       l = []
8.
```

Na linha 1, a variável *l* recebe uma lista vazia; na linha 2, a variável *listaPalavra* é também uma lista vazia; na linha 3, temos um laço de repetição que percorre todas as palavras do vocabulário; na linha 4, temos novamente um laço de repetição que vai desde *0* até o comprimento da variável *respostas*; na linha 5, temos duas funções: a função *ContaFreqPalavra* conta a frequência de uma determinada palavra em uma determinada resposta e a função *append* que anexa a frequência da palavra na variável *l*; na linha 6, a função *append* anexa a variável *l* na variável *listaPalavra*; na linha 7, a variável *l* torna-se novamente uma lista vazia.

A variável *listaPalavra* fornece uma lista onde cada elemento dessa lista é também uma lista sendo que cada elemento dessa sub-lista foi uma determinada palavra e suas frequências em todas as respostas. Entretanto, foi necessário apenas das frequências numéricas.

```
import numpy as np

9.  listaFreq = []
10. matrizListaFreq = []
11. for j in range(0, len(listaPalavra)):
12.     for i in range(0, len(listaPalavra[j])):
13.         listaFreq.append(listaPalavra[j][i][1])
14.     matrizListaFreq.append(listaFreq)
15.     listaFreq = []
16.
17. matriz = np.matrix(matrizListaFreq)
```

Na linha 9, a variável *listaFreq* recebe uma lista vazia; na linha 10, a variável *matrizListaFreq* é também uma lista vazia; na linha 11, temos um laço de repetição que vai desde *0* até o comprimento da variável *listaPalavra*; na linha 12, temos novamente um laço de repetição que vai desde *0* até o comprimento de uma sub-lista da variável *listaPalavra*; na linha 13, a função *append* anexa as frequências numéricas na variável *listaFreq*; na linha 14, a função *append* anexa a variável *listaFreq* na variável *matrizListaFreq*; na linha 15, a variável *listaFreq* torna-se novamente uma lista vazia; na linha 17, o módulo *numpy* transforma a variável *matrizListaFreq* em uma outra variável *matriz*, que é a representação matricial do corpus.

4.4 APLICAÇÃO DA FUNÇÃO PONDERAÇÃO

Após construída, a matriz termo-documento é submetida a uma transformação preliminar, chamada função ponderação, definida como um produto de um número que representa uma ponderação local por outro que representa uma ponderação global.

A ponderação local estima a importância da palavra i na resposta j, enquanto que a ponderação global estima o grau de influência da palavra i na coleção de respostas como um todo. O trabalho de Lifchitz (2009) afirma que esta transformação preliminar tem grande impacto nos resultados de LSA.

Após aplicação da função ponderação cada entrada da matriz torna-se:

$$a(i, j) = L(i, j) \cdot G(i),$$

onde $L(i, j)$ denota a ponderação local e $G(i)$ denota a ponderação global.

Na Tabela 4 foram relacionadas as ponderações locais e globais utilizadas no modelo LSA proposto.

Tabela 4 – Ponderações locais e globais

Ponderações locais			
	tf	$L(i,j) = tf(i,j)$	$tf(i,j)$ é o número de vezes que a palavra i ocorre na resposta j
	logaritmo	$\log_2(tf(i,j)+1)$	
Ponderações globais			
	trivial	$G(i) = 1$	
	normal	$G(i) = \frac{1}{\sqrt{\sum_j L(i,j)^2}}$	
	GfIdf	$G(i) = \frac{gf(i)}{df(i)}$	$gf(i)$ é a frequência global da palavra i e $df(i)$ é o número de respostas em que a palavra i aparece
	Idf	$1 + \log_2(\frac{n^o\ respostas}{df(i)})$	
	entropia	$1 + \frac{\sum_j p(i,j) \cdot \log_2 p(i,j)}{\log_2 n^o\ respostas}$	$p(i,j) = \frac{tf(i,j)}{gf(i)}$ é a probabilidade condicional
	entropia real	$-\sum_j p(i,j) \cdot \log_2 p(i,j)$	

Algumas notações foram retiradas do trabalho de Nakov, Popova e Mateev (2001): **tf** é a ponderação local trivial termo frequência que nada mais é do que a frequência com que a palavra i aparece na resposta j; a ponderação local **logaritmo** é usada para diminuir números elevados; a ponderação global **normal** representa a normalização das linhas; **GfIdf** é a relação da frequência global de uma palavra pelo número de respostas em que ela aparece; e **Idf** é a ponderação global frequência documento inverso que pode ser interpretada como a quantidade de informações do aparecimento da palavra i no conjunto total de respostas.

O código abaixo mostra como a transformação termo-frequência vs frequência documento inverso foi aplicada na matriz que representa o corpus.

```
from sklearn.feature_extraction.text import TfidfTransformer

# esquema de ponderação tf-idf

1. def TfIdf(matriz):
2.     transformer = TfidfTransformer()
3.     tfidf = transformer.fit_transform(matriz)
4.      return tfidf.toarray()
```

Na linha 1, foi definida a função *TfIdf* que recebe como parâmetro a variável *matriz*; na linha 2, a variável *transformer* recebe a transformação termo-frequência vs frequência documento inverso *TfidfTransformer*; na linha 3, a variável *tfidf* é o resultado da aplicação da função *TfidfTransformer* na variável *matriz*; na linha 4, temos o retorno da variável *tfidf* como matriz.

Depois de representar o corpus pela matriz termo-documento e submete-la a um esquema de ponderação, é calculada a SVD da matriz termo-documento para aproxima-la por uma outra matriz em um espaço de dimensão menor.

4.5 CÁLCULO DA SVD E ESCOLHA DA DIMENSIONALIDADE

Quando LSA foi criado buscava-se um modelo que revelasse as relações entre palavras e textos usando uma matriz de ocorrência observada. Já existiam linhas de pesquisas que trabalhavam com modelos de proximidade, entretanto estes modelos eram fracos na captura de relações semânticas entre as palavras dos textos. A abordagem proposta no trabalho de Landauer, Foltz e Laham (1998) foi um modelo de análise fatorial baseado em Decomposição a Valores Singulares (SVD) que tinha como principal diferença examinar problemas de tamanho razoável em espaços multidimensionais com as seguintes características:

1. Ajustes na representação dos conjuntos de textos;
2. Representação explicita de textos e palavras;
3. Tratamento computacional para grandes conjuntos de dados.

O processo de Decomposição a Valores Singulares (SVD) é o cerne de qualquer modelo LSA. No modelo LSA aqui proposto o cálculo da SVD da matriz termo-documento é feito após submeter esta matriz a um esquema de ponderação. Na prática, denotando por A a matriz termo-documento e se A é do tipo $m \times n$, então o Teorema da Decomposição a Valores Singulares garante que A pode ser fatorada como sendo um produto de três outras matrizes: duas matrizes ortogonais U e V de ordens m e n, respectivamente, e uma matriz diagonal S do mesmo tipo $m \times n$ da matriz A. A Figura 10 fornece uma visualização desta decomposição:

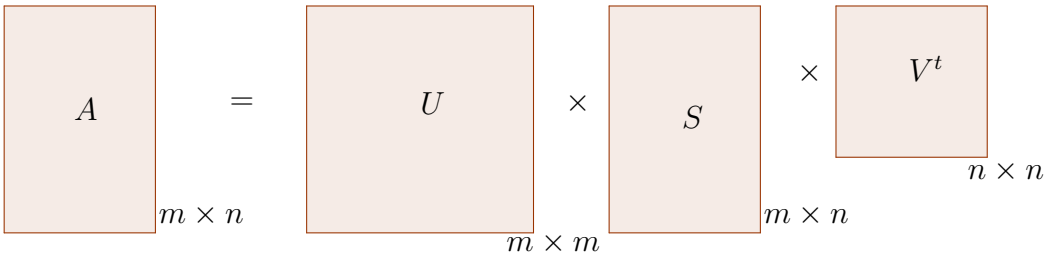


Figura 10 – Decomposição a Valores Singulares da matriz termo-documento A

Considerando a matriz A como sendo uma transformação linear de $\mathbb{R}^n$ em $\mathbb{R}^m$, as colunas da matriz V formam uma base do domínio e as colunas da matriz U formam uma base da imagem desta transformação. Quando expressamos vetores relativamente a estas bases, a transformação dilata ou contrai suas componentes de acordo com a magnitude de seus valores singulares, e possivelmente descarta componentes ou acrescenta zeros caso seja necessário. Como os valores singulares são dispostos em ordem decrescente de importância podemos escolher aqueles com maior magnitude para obter uma aproximação da matriz A. Por exemplo, escolhemos as k primeiras colunas das matrizes U, V formando as matrizes U_k e V_k, respectivamente, e as k primeiras linhas e colunas da matriz S formando a matriz S_k. O produto das matrizes U_k, S_k e V_k gera uma matriz A_k com as mesmas dimensões da matriz A, de menor posto e que é a melhor aproximação em média quadrática da matriz A. A Figura 11 mostra uma visualização de como a matriz A_k é gerada após a escolha da dimensionalidade.

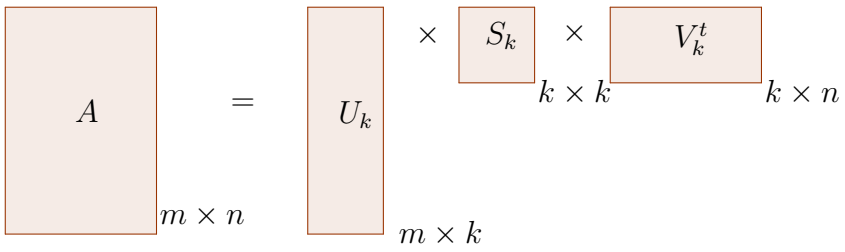


Figura 11 – Escolha da dimensão k

Na redução para o espaço semântico ocorre um truncamento do espaço das colunas da matriz A restrito aos autovetores associados aos valores singulares de maior magnitude. Na prática os vetores que representam os textos são transferidos de um espaço vetorial n-dimensional para outro k-dimensional, onde k é muito menor do que n, na grande maioria dos casos. É listado abaixo algumas razões do porquê devemos reduzir para o espaço semântico:

- Facilidade no tratamento computacional: uma aproximação de A por uma matriz A_k de posto menor necessitaria de um espaço de armazenamento menor;
- Melhorias nas relações entre os textos através da identificação de estruturas semânticas ocultas nas relações entre palavras;
- Trabalhando com valores singulares de maior magnitude temos uma redução na propagação de ruídos e variabilidade no uso de palavras.

A escolha do número k, que é a dimensão do espaço semântico, ainda é uma questão em aberto e possivelmente tem o impacto mais significativo sobre os resultados de LSA como afirma o trabalho de Wild et al. (2005). Na literatura foram encontradas algumas possibilidades para escolher a dimensão do espaço semântico. O trabalho *Automatic Estimation of the LSA Dimension* (Fernandes, Artifice e Fonseca (2012)) propõe a fórmula

$$k = n_T^{\left(\frac{1}{1+\frac{log(n_T)}{10}}\right)},$$

onde n_T é o número de textos, para estimar a dimensão do espaço semântico de um modelo LSA .

Nos experimentos realizados, os melhores resultados foram obtidos com a escolha do número k de forma empírica através de testes, fazendo k variar de 2 até o número de textos a serem avaliados. A cada passo o programa LSA gera um espaço semântico no

qual é realizada uma classificação entre cada uma das respostas e a resposta base através de uma medida de similaridade.

O código abaixo mostra como foi feito o cálculo da SVD e redução para o espaço semântico do modelo LSA proposto.

```
import numpy as np
from scipy import linalg

1. matriz_peso=TfIdf(matriz)
2.
3. # cálculo da SVD
4.
5. U,Sigma,Vt=linalg.svd(matriz_peso)
6.
7. # redução para o espaço semantico
8.
9. S=np.array(Sigma[0:k])
10.
11. Sk=np.diag(S)
12.
13. Vtk=Vt[:,0:k]
14. Vtkt=Vtk.transpose()
15.
16. Ak=np.dot(Sk,Vtkt)
```

Na linha 1, a variável *matriz_peso* recebe a variável *matriz* já submetida a transformação ponderação; na linha 5, temos o cálculo SVD da variável *matriz_peso* através do pacote *linalg.svd* do módulo *scipy*: a variável *matriz_peso* é fatorada como um produto de três outas matrizes representadas pelas variáveis *U*, *Sigma* e *Vt*, respectivamente; na linha 9, a variável *S* é uma lista como os *k* primeiros valores singulares da variável *Sigma* escolhidos empiricamente; na linha 11, o módulo *numpy* transforma a variável *S* em uma matriz diagonal representada pela variável *Sk*; na linha 13, a variável *Vtk* representa uma matriz formada pelas *k* primeiras colunas da matriz representada pela variável *Vt*; na linha 14, a variável *Vtkt* é a transposta da matriz *Vtk*; na linha 16, a variável *Ak* é o produto das matrizes representadas pelas variáveis *Sk* e *Vtkt*, respectivamente.

A matriz representada pela variável *Ak* referida no código acima é a melhor aproximação em norma quadrática da matriz inicial do modelo LSA proposto. Esta matriz tem como principal característica o fato de que a dimensão do espaço das linhas é muito menor do

que o espaço das linhas da matriz inicial, e neste espaço, chamado de espaço semântico, é que é realizada a etapa de classificação das respostas.

4.6 CLASSIFICAÇÃO DAS RESPOSTAS

4.6.1 Medidas de similaridade utilizadas

A grande maioria dos trabalhos da literatura tratam similaridade entre textos como sendo uma distância em um espaço métrico.

Esta classificação é um valor estimado por uma medida de similaridade através de comparações entre os textos. Neste estudo foi utilizada a abordagem espaço vetorial: convertemos cada resposta em vetores e através de cálculos com estes vetores foi possível estipular um valor para a similaridade entre as respostas. As medidas de similaridade utilizadas foram o cosseno do ângulo, o coeficiente de correlação de Pearson e o coeficiente de correlação ρ de Spearman.

Na tabela 5 temos as fórmulas que definem cada uma destas medidas: $t_j = (c_{1j}, \cdots, c_{Ij})$ é a representação vetorial de uma resposta, onde j é o número respostas e I significa o número total de palavras do vocabulário, $cov(t_j, t_k)$ e $var(t_j, t_k)$ representam covariância e variância entre t_j e t_k, respectivamente, e $d_i = c_{ij} - c_{ik}$.

Tabela 5 – Medidas de Similaridade

medidas de similaridade	
cosseno	$\frac{t_j \cdot t_k}{\|t_j\|_2 \, \|t_k\|_2}$
correlação de Pearson	$\frac{cov(t_j, t_k)}{\sqrt{var(t_j) \cdot var(t_k)}}$
correlação de Spearman	$1 - \frac{6 \sum_{i=1}^{I} d_i^2}{I^3 - I}$

4.6.2 Classificação

Na abordagem utilizada buscou-se estimar computacionalmente, através de um valor, o quanto cada resposta é similar a resposta referência. A ideia básica da classificação é a seguinte: a medida de similaridade recebe um par de respostas:

medidadesimilaridade(resposta referência, resposta),

sendo que somente o parâmetro resposta varia. Para cada resposta é retornado um valor pontuação que indica sua similaridade com a resposta referência.

Nos experimentos os melhores resultados foram obtidos para o cosseno do ângulo. Como foi dito anteriormente, codificamos a primeira coluna da matriz Ak como sendo a resposta de referência, assim foi implementado o código para o cálculo da medida de similaridade calculando o cosseno do ângulo entre a primeira coluna da matriz Ak e as demais colunas da própria matriz.

```
 # Calculo da medida de similaridade

1. L=[]
2. for i in range(1,q):
3.         L.insert(0,fabs(cosseno(A2[:,0],A2[:,i])))
4. L.reverse()
```

Na linha 1, a variável L recebe uma lista vazia; na linha 2, temos um laço de repetição que varia desde *1* até o número q de colunas da matriz Ak; na linha 3, a função *insert* insere na variável L o cosseno do ângulo entre a primeira e as demais colunas da matriz Ak; a linha 4, apenas reverte a lista representada pela variável L.

4.7 CALIBRAÇÃO DO MODELO LSA

Durante a etapa de classificação das respostas foi observado grandes diferenças entre as pontuações dos avaliadores humanos e as estimadas computacionalmente pelo modelo LSA para respostas com poucas palavras. Foi implementado um fator de ajuste para corrigir essas discrepâncias. Após aplicação do fator de ajuste o modelo LSA classifica novamente as respostas, calcula o erro cometido e a acurácia da estimativa. Será visto com mais detalhes como foram implementados os métodos para calibração do modelo LSA no próximo capítulo.

4.8 CÁLCULO DA ACURÁCIA

O cálculo da acurácia foi feito da seguinte forma:

1. calcula a soma dos erros cometidos em cada aproximação:

   $$\text{soma dos erros} = \sum \mid \text{avaliacao humana} - \text{valor pontuacao atribuida pelo programa} \mid$$

2. calcula a média dos erros:

   $$\text{media} = \frac{\text{soma dos erros}}{\text{numero de respostas}}$$

3. calculo da acurácia:

   $$\text{acuracia} = \frac{\text{pontuacao maxima - media}}{\text{pontuacao maxima}}$$

No Capítulo 3 foi visto que modelos LSA são adequados para muitas aplicações. O principal objetivo é a implementação de um modelo LSA em um ambiente virtual de aprendizagem para avaliação automática de respostas para questões discursivas e que a acurácia desse modelo contra avaliadores humanos seja a maior possível.

A avaliação do sistema foi feita comparando-se a acurácia do modelo LSA contra a acaurácia entre doios avaliadores humanos e de um sistema baseado apenas em n-gramas. O trabalho de Salton, Wong e Yang (1975) mostra que a técnica de n-gramas pode ser usada para medir a similaridade de dois textos em sistemas de recuperação de informações. Entretanto, um modelo LSA captura relações do uso contextual de palavras, sinônimos, etc., podendo ser uma medida de similaridade mais fina e até mesmo ser um aperfeiçoamento da técnica de n-gramas.

Para certificação desta última afirmação implementou-se um modelo centrado apenas em n-gramas, o qual denominamos de modelo *base-line*. O objetivo é que o modelo n-gramas sirva como *base-line* de performance para mostrar que o modelo LSA é uma versão mais fina e apurada quando comparada com n-gramas.

No modelo base-line implementado foram utilizados separadamente unigramas e bigramas de palavras, como também a combinação de unigramas e bigramas através de um modelo de regressão linear múltipla. É o que será visto na próxima seção.

4.9 IMPLEMENTAÇÃO DE UM MODELO BASE-LINE

Esta subseção sobre n-gramas foi incluída com intuito de fornecer uma visão geral sobre a técnica de n-gramas.

Os n-gramas podem ser utilizados tanto para palavras como para caracteres. No caso de caracteres são úteis para recuperação de palavras. Por exemplo, as palavras *abacaxi* e *abcaxi* são diferentes, mas compartilham certa similaridade.

Uma maneira de calcular a similaridade por n-gramas entre duas sequências ordenadas de caracteres (ou palavras) é contar o número de n-gramas comuns em ambas as sequências, multiplicar este total por 2 e dividir pela soma dos n-gramas que constam em cada uma das duas sequências como visto no trabalho de Emygdio et al. (2003).

Apenas para efeito de exemplificação, a Tabela 6 apresenta os unigramas, bigramas e trigramas dos caracteres das palavras *abacaxi* e *abcaxi* e o respectivo coeficiente de similaridade.

Tabela 6 – Unigramas, bigramas e trigramas dos caracteres das palavras *abacaxi* e *abcaxi*

Palavra	abacaxi	abcaxi	similaridade
unigrama	[a,b,a,c,a,x,i]	[a,b,c,a,x,i]	$\frac{2 \cdot 6}{6+7} = 0.92$
bigrama com espaço	[(32,a),(a,b),(b,a),(a,c), (c,a),(a,x),(x,i),(i,32)]	[(32,a),(a,b),(b,c),(c,a), (a,x),(x,i),(i,32)]	$\frac{2 \cdot 6}{8+7} = 0.80$
bigrama sem espaço	[(a,b),(b,a),(a,c), (c,a),(a,x),(x,i)]	[(a,b),(b,c),(c,a), (a,x),(x,i)]	$\frac{2 \cdot 4}{6+5} = 0.73$
trigrama com espaço	[(32,a,b),(a,b,a),(b,a,c),(a,c,a) (c,a,x),(a,x,i),(x,i,32)]	[(32,a,b),(a,b,c),(b,c,a), (c,a,x),(a,x,i),(x,i,32)]	$\frac{2 \cdot 4}{7+5} = 0.67$
trigrama sem espaço	[(a,b,a),(b,a,c),(a,c,a) (c,a,x),(a,x,i)]	[(a,b,c),(b,c,a), (c,a,x),(a,x,i)]	$\frac{2 \cdot 2}{5+4} = 0.44$

Foi mplementado um modelo de n-gramas, no qual foram utilizados unigramas e bigramas de palavras, para estimar a similaridade das respostas que constituem o corpus da pesquisa.

Medimos a similaridade de duas respostas pelo modelo *base-line* de duas maneiras distintas: por unigramas de palavras contando as palavras comuns nas duas respostas sem levar em conta a ordem das mesmas, e por bigramas de palavras contando as sequências de duas palavras nas duas respostas.

A Tabela 7 apresenta duas respostas escritas para a questão de Biologia.

Tabela 7 – Respostas escritas para a questão de Biologia

Respostas	
tecido epitelial que é responsável pelo revestimento do corpo e proteção contra choques por ter grande elasticidade tecido nervoso é responsável pela coordenação de movimentos e estímulos sensoriais e motores por todo o corpo tecido ósseo é responsável pela sustentação graças a intensa dos ossos o que confere rigidez para manter o corpo ereto	tecido nervoso é formado por um grupo de células chamadas de neurônios, que são responsáveis por coordenação de movimentos, pensamentos e sensações tecido ósseo formado por células ósseas, responsáveis pela sustentação do corpo humano tecido epitelial formado por células epiteliais, responsáveis pelo revestimento do corpo humano

A similaridade entre as respostas da Tabela 7 pode ser estimada por unigramas de palavras. Para isto basta contar a quantidade de palavras de cada uma das respostas. É fácil verificar que as quantidades de palavras destas respostas são 53 e 47, respectivamente, sendo 14 o número de palavras comuns.

O coeficiente de similaridade por unigramas de palavras entre as duas respostas é dado por:

$$\frac{2 \cdot 14}{53 + 47} = 0.28$$

A mesma estimativa feita pelo modelo LSA proposto resulta em uma similaridade de 0.6667 entre estas mesmas respostas.

Nesta pequena amostra verifica-se uma relativa melhora no índice de similaridade em favor do modelo LSA.

A implementação de um modelo baseado apenas em n-gramas é apenas para verificar a performance do modelo LSA proposto. A expectativa é que LSA seja um refinamento para n-gramas. No próximo capítulo será feita a aplicação prática tanto do modelo LSA quanto do modelo base-line na avaliação automática de respostas escritas.

5 APLICAÇÃO PRÁTICA DO MODELO LSA NA AVALIAÇÃO AUTOMÁTICA DE RESPOSTAS ESCRITAS

5.1 O CORPUS DA PESQUISA

O propósito da pesquisa é investigar e desenvolver parâmetros para uso prático de um modelo LSA para correção automática de respostas escritas a questões discursivas.

O Corpus da Pesquisa foi constituído por textos escritos, por alunos, como respostas a perguntas produzidas sob comandos de: i) uma questão discursiva conceitual de Biologia, e ii) uma questão discursiva argumentativa de Geografia.

As questões de Biologia e Geografia constam no boletim *EDITAL 016/2007- UFPA 3ª FASE* do Processo Seletivo Seriado de 2008 da Universidade Federal do Pará. Este boletim era constituído por 75 (setenta e cinco) questões analítico-discursivas de 25 disciplinas, com uma média de três questões por disciplina, além de uma redação, sendo que cada candidato deveria escolher e responder apenas 1 (uma) questão de cada disciplina.

As respostas eram escritas em boletins específicos para cada disciplina. De um total de 15.154 boletins de respostas foram selecionados aleatoriamente 1.000 boletins de respostas, já com as pontuações atribuídas pelos especialistas humanos. A pontuação máxima de cada resposta era 6 pontos, sendo que só era permitido para o avaliador humano atribuir as pontuações $0, 1, 2, 3, 4, 5$ ou 6. Cada resposta era avaliada por dois avaliadores humanos, sendo que cada avaliador não conhece a avaliação do outro. A pontuação era atribuída da seguinte maneira: no caso de coincidência era atribuída a pontuação comum; se a diferença entre as pontuações fosse de 1 ponto, então era atribuída a maior pontuação; uma diferença igual ou superior a 2 pontos era considerada uma discrepância, e neste caso a pontuação era atribuída por um terceiro avaliador (após o processo de discrepância).

Como cada candidato só poderia responder uma resposta por didciplina, de um universo de 1.000 respostas foram escolhidas aquelas respostas com maior público: 130 respostas para uma questão de Biologia e 229 respostas para uma questão de Geografia.

As 359 respostas escolhidas passaram por um processo de digitalização manual onde foram feitas correções com um corretor ortográfico automático, mas não foi feito qualquer tipo de correção de concordância gramatical.

Nas três próximas subseções são apresentados os enunciados das questões cujas respostas constituem o corpus desta pesquisa.

5.1.1 Questão de Biologia

A questão de Biologia de natureza discursivo-conceitual propunha a elaboração de três conceitos de uma dada taxonomia da Citologia. Abaixo temos o enunciado da questão:

> *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*

Apenas para efeito ilustrativo na Tabela 8 temos uma amostra de cinco respostas que foram digitalizadas para esta questão com a respectiva pontuação atribuída por avaliadores humanos:

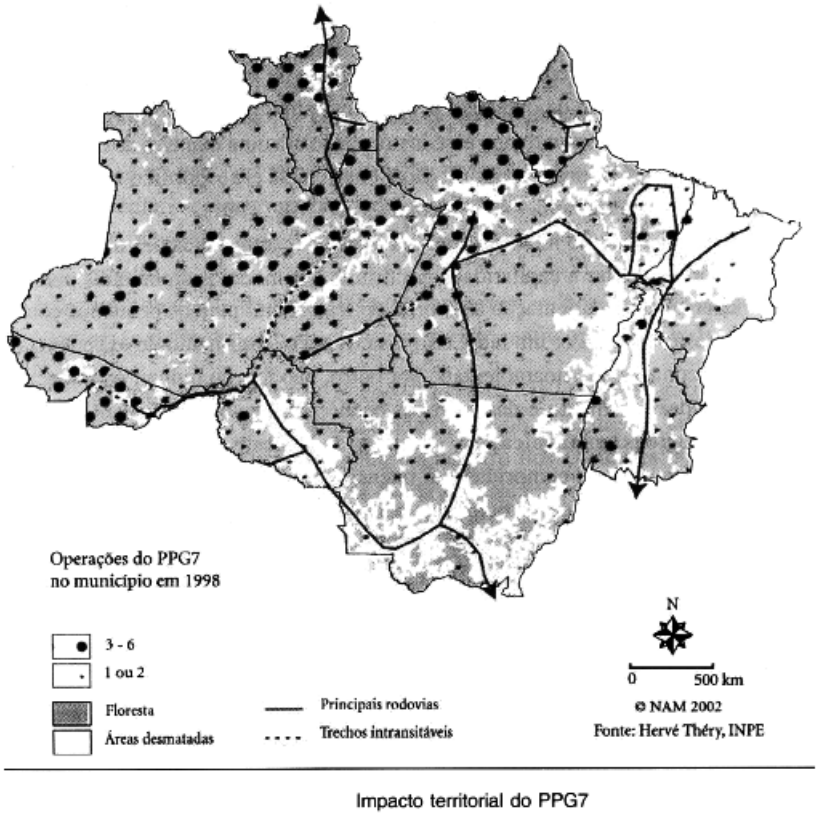
Tabela 8 – Respostas escritas para a questão de Biologia

Tecido muscular esquelético responsável pela sustentação dos Tecido epitelial proteção e revestimento Tecido muscular cardíaco bombeamento de fluxo sanguíneo, circulação sanguínea	6
tecido ósseo sustentação do corpo tecido cartilaginoso proteção da pele tecido muscular responsável pelas articulações	5
tecido Ósseo serve para a estrutura dos ossos tecido Nervoso serve para circulação do sangue Tecido epitelial serve para	4
tecido conjuntivo sua função é a estruturação tecido nervoso sua função tecido epitelial sua função é a sustentação	3
epiderme tecido do revestimento	0

5.1.2 Questão de Geografia

A questão de Geografia de natureza discursivo-argumentativa propunha a elaboração de argumentação em defesa de dado ponto de vista formado a respeito da Geografia Humana e Econômica da Região. Abaixo temos o enunciado da questão:

> *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. Belém: UFPA/NAEA,2007, p. 30).*



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.

Apenas para efeito ilustrativo na Tabela 9 temos uma amostra de três respostas que foram digitalizadas para esta questão com a respectiva pontuação atribuída pelos avaliadores humanos:

Tabela 9 – Respostas escritas para a questão de Geografia

As áreas leste e sul da região amazônica, têm o maior impacto de desmatamento. O processo de desmatamento da amazônia, iniciou se a décadas e intensifica se com o fácil acesso às terras, devido aos cartório fraudulentos, o que leva a grilagem e a partir daí o desmatamento para o desenvolvimento de atividades. Além disso, a infraestrutura (rodovias, incentivos fiscais) oferecida pelo estado, para implantação de projetos também demanda áreas florestais. Na medida em que há o desenvolvimento urbano ao entorno das rodovias e dos projetos, responsável pela atração de atores sociais, culturalmente distintos, e que implantam novas atividades (pecuária, madeireiras, e agricultura de grãos) na região	4
A região oriental da amazônia. Com o aproveitamento das rodovias, intensificam se a chegada de madeireiras e fazendeiros que irão desmatar áreas da amazônia, provocando vários impactos ambientais como: a lixiviação, o aumento do fluxo de CO2 para a atmosfera, contribuindo para o agravamento do efeito estufa.	2
Os estados do pará e Mato-Grosso. Um dos principais motivos do desmatamento é a exploração de minérios que continua intenso com a presença de empresas estrangeiras e um outro motivo é a agricultura que vem desmatando grandes áreas com a expansão da soja.	1

Para verificação do desempenho do modelo LSA proposto as respostas escritas para as duas questões enunciadas acima foram comparadas com uma reposta de referência. Por esta razão foi necessário incluir no corpus da pesquisa duas respostas de referência: uma para as respostas de Biologia e outra para as respostas de Geografia.

5.1.3 Respostas de Referência

Como já foi dito anteriormente o corpus da pesquisa foi constituído por respostas escritas por candidatos de um processo seletivo para ingresso na Universidade Federal do Pará. Logo após a realização das provas, equipes de avaliadores são formadas para correção das provas. Estas equipes são responsáveis pela discussão de uma grade de correção para as respostas de todas as disciplinas.

Na Tabela 10 temos a grade de correção utilizada como referência para as respostas para a **Questão 2** de Biologia. O critério pra atribuição das pontuações de cada uma das respostas era 1 ponto para cada tecido e 1 ponto para cada função.

Tabela 10 – Grade de correção da questão de Biologia

TECIDOS	FUNÇÕES
Epitelial ou glandular Epitelial	Revestimento interno (trocas, absorção de substâncias), ou externo (proteção, perda de água,) proteção (mecânica), percepção de estímulos,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
sanguíneo	Transporte de gases, nutrientes
linfático	Defesa
Muscular	Movimento, contração
Nervoso	Regulação e integração interna e coordenação corporal, homeostase, raciocínio, memória, irritabilidade, condução de impulsos nervosos

Para efeito de comparação, foi considerada como resposta referência para as respostas de Biologia a concatenação das linhas e colunas da Tabela 10.

Na Tabela 11 temos a grade de correção utilizada como referência para a questão de Geografia e o critério pra atribuição das pontuações de cada uma das respostas.

Tabela 11 – Grade de correção da questão de Geografia

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 pt
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 pt
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 pt
Respostas explicativas que apresentarem apenas um fator, como, por exemplo, a atividade madeireira e sua relação com o desmatamento	2,0 pt

Observa-se que a grade de correção para a questão de Geografia não fornece o conteúdo, apenas sugere como a questão pode ser respondida. Para efeito de comparação, como a técnica LSA necessita de conteúdos, foi considerada como resposta referência a concatenação de três respostas que obtiveram a maior pontuação por avaliadores humanos.

5.2 MÉTODOS PARA CALIBRAÇÃO DO MODELO LSA

A calibração do modelo foi realizada durante a execução dos experimentos: foi utilizada uma abordagem onde encontra-se os melhores valores possíveis para cada parâmetro visando obter a melhor acurácia.

No etapa de pré-processamento, as variações foram nas entradas da matriz termo-documento A por:

a) considerar todas as stop words;

b) remover todas stop words;

c) remover todas stop words e aplicar um algoritmo de stemming

Isto possibilitou a construção de seis matrizes termo-documentos distintas para cada conjunto de respostas.

Na literatura pesquisada existe um consenso de que a combinação de unigramas com remoção de palavras e aplicação de um processo de stemming fornece os melhores resultados para LSA. Entretanto, como foram utilizados bigramas de palavras, foi feita a opção de considerar as alternativas a), b) e c) descritas na lista acima durante o pré-processamento das respostas. Além disso, foram considerados apenas aqueles bigramas que aparecem em pelo menos duas respostas, para evitar uma grande quantidade de entradas nulas na matriz termo-documento.

Durante a etapa de pesagem, a matriz termo-documento A é submetida a uma transformação preliminar chamada "função peso"; esta função é definida como um produto de um número que representa uma pesagem local por um outro que representa uma pesagem global. O esquema que forneceu os melhores resultados foi *termo frequência (tf) vs frequência documento inverso (idf)*. Este mesmo esquema foi utilizado nos trabalhos de Dumais (1991), Nakov, Popova e Mateev (2001), Marinrez et al. (2005), Wild et al. (2005), Jorge-Botana et al. (2010),Zen, Iskandar e Linang (2011).

Na etapa cálculo da SVD é que é que é escolhida a dimensão do espaço semântico; esta etapa tem o maior impacto na performance de modelos LSA. Algumas sugestões podem ser encontradas na literatura sobre a escolha da dimensão k do espaço semântico, como nos trabalhos de Wild et al. (2005) e Jorge-Botana et al. (2010). No entanto não existe consenso: o valor da dimensão k é uma função do espaço das colunas da matriz A.

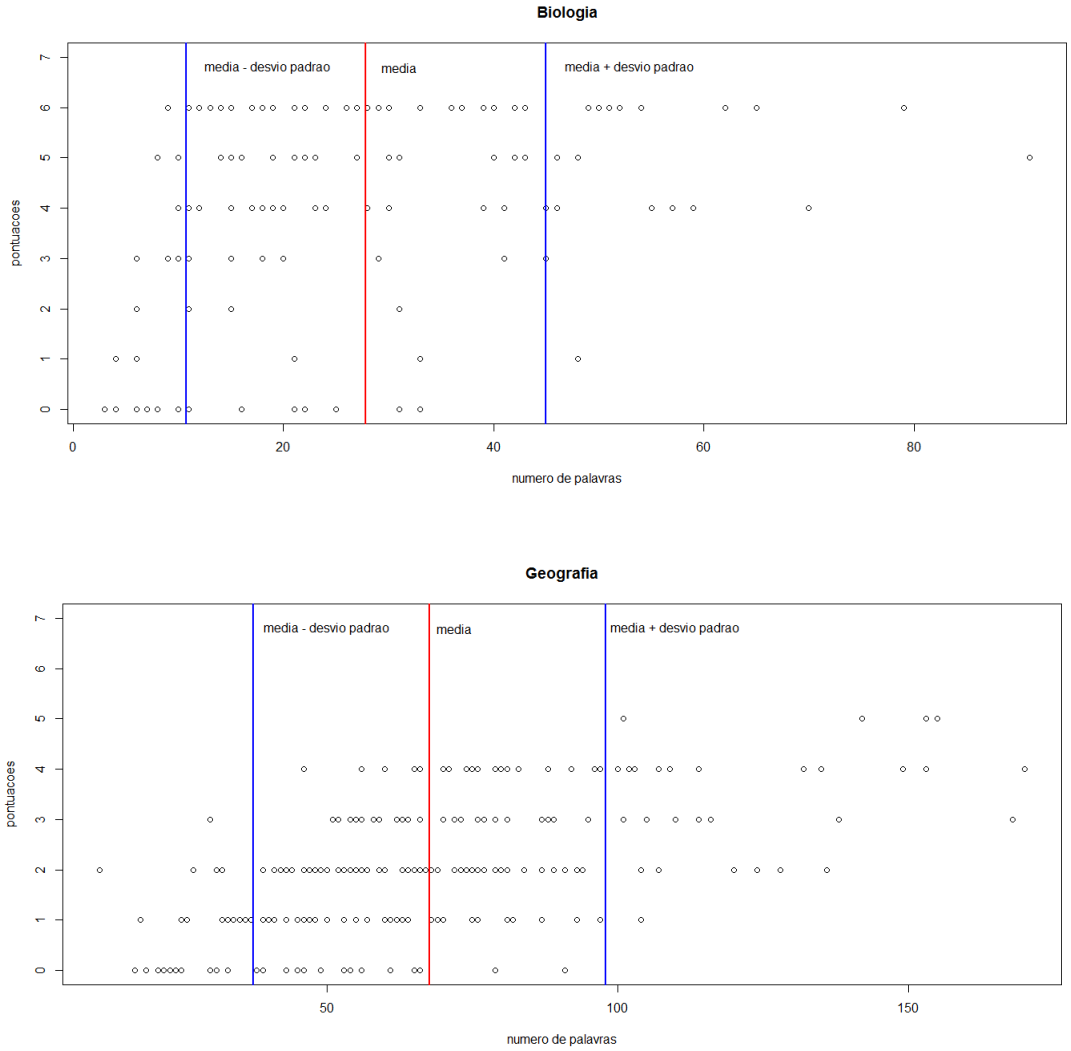
Através de um método de força bruta foi feita a variação da dimensão k de 1 até o número total de respostas, e escolhido aquele valor de k que fornece o melhor índice de acurácia. Os melhores resultados foram obtidos para valores de k entre 2 e 8, provavelmente devido ao fato de que o método SVD ordena os valores singulares, do maior para o de menor magnitude com o respectivo autovetor associado. Os trabalhos de Landauer, Foltz e Laham (1999), Wolfe et al. (1998), Foltz (2009), Magliano e Graesser (2012) consideraram a dimensionalidade como sendo 94, 100, 200 ou 300, respectivamente.

Na etapa de classificação, foi feita uma estimativa da similaridade entre a resposta de referência e as demais respostas. Foi utilizada a abordagem espaço vetorial: cada resposta foi convertida em um vetor, e através de cálculos com estes vetores estimamos um valor de similaridade entre as respostas. Os melhores resultados foram obtidos considerando o cosseno do ângulo entre dois vetores, embora a correlação de Pearson tenha fornecido resultados similares. Os trabalhos de Jorge-Botana et al. (2010), Olmos et al. (2011) utilizaram uma abordagem que combina o cosseno do ângulo com a distância euclidiana . Em alguns experirmentos a distância euclidiana foi utilizada como medida de similaridade, entretanto os resultados não foram satisfatórios.

Para as respostas das questões de Biologia e Geografia, nos primeiros experimentos foi observado grandes distorções entre as classificações atribuídas pelo modelo LSA e por avaliadores humanos para respostas com um número pequeno de palavras. Foi necessário verificar a pontuação obtida em função do número de palavras por resposta. A média de palavras das 130 respostas para a questão de Biologia foi de $27,84$, com desvio padrão de $17,11$, enquanto que para a questão de Geografia a média de palavras das 229 respostas foi de $67,65$, com desvio padrão de $30,25$.

A Figura 12 mostra as pontuações obtidas pelas respostas por avaliadores humanos em função do número de palavras para as duas questões:

Figura 12 – Pontuações por número de palavras



Para a questão de Biologia verificou-se que as maiores pontuações foram obtidas para aquelas respostas com o número de palavras entre a média mais ou menos o desvio padrão. Para a questão de Geografia verificou-se que respostas com um número baixo de palavras obtiveram uma pontuação baixa e respostas com número de palavras acima da média mais o desvio padrão tiveram pontuações mais elevadas.

Foi necessário a introdução de um método para ajuste do programa. Foi implementado um fator de penalização para corrigir discrepâncias entre as pontuações humana e automática da seguinte maneira: as pontuações atribuídas pelo programa para aquelas respostas com um número de palavras abaixo da média menos o desvio padrão eram multiplicadas pelo número de palavras da própria resposta dividindo o valor obtido pela média de palavras das respostas. Em seguida, o programa estima novamente a pontuação de cada resposta, calcula o erro cometido e a acurácia da estimativa. Os trabalhos de Olmos Olmos et al.

(2011) e Jorge-Botana et al. (2010) corrigiram este problema combinando a medida do cosseno com a distância euclidiana. Em trabalhos futuros pretendemos investigar qual a melhor abordagem para corrigir este problema com respostas curtas.

5.3 DISCUSSÃO DOS RESULTADOS

Esta pesquisa foi desenvolvida visando quatro principais objetivos:

i) Criar um modelo de co-ocorrência de unigramas e bigramas (base-line) para servir como performance para o modelo LSa proposto;

ii) Usar bigramas nas entradas da matriz inicial do modelo LSA;

iii) Calibrar o modelo LSA de acordo com o número de palavras por respostas;

iv) Comparar as distribuições das pontuações atribuídas pelo modelo LSA e por avaliadores humanos.

5.3.1 Modelo Base-line

A avaliação da eficiência do sistema LSA proposto foi através da comparação da acurácia do próprio sistema contra a acurácia de dois avaliadores humanos e também contra a acurácia de um sistema baseado apenas em n-gramas, o qual denominamos *Modelo base-line*.

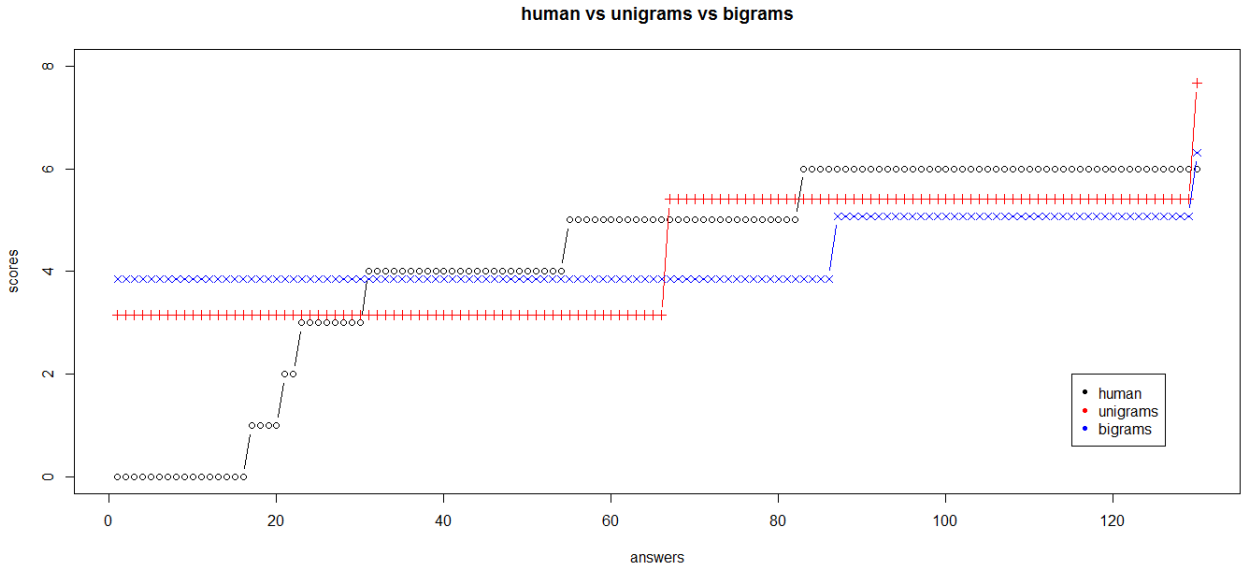
Na literatura temos o trabalho de Marinrez et al. (2005) que apresenta uma avaliação comparativa entre LSA e o algoritmo BLEU, o qual é baseado apenas em $n-$gramas. Esta foi a motivação para a implementação do modelo base-line, o qual foi implementado em três cenários: somente unigramas, somente bigramas, e unigramas combinado com bigramas através de um modelo de regressão linear múltipla.

O modelo base-line estima a similaridade entre uma resposta e a resposta referência considerando o número de unigramas ou bigramas que são comuns em ambas as respostas. Para todo os casos foram utilizados modelos de regressão linear para aproximar as pontuações do modelo base-line com a pontuação de avaliadores humanos, como feito no trabalho de Malandrakis, Iosif e Potamianos (2012), o qual fez uso da métrica informação mútua entre palavras em um modelo baseado em n-gramas. Este trabalho obteve uma correlação de 0.62 na comparação entre duas sentenças.

Questão de Biologia

Os resultados do modelo base-line para a questão de Biologia são apresentados graficamente na Figura 13:

Figura 13 – Resultados do modelo base-line para a questão de Biologia



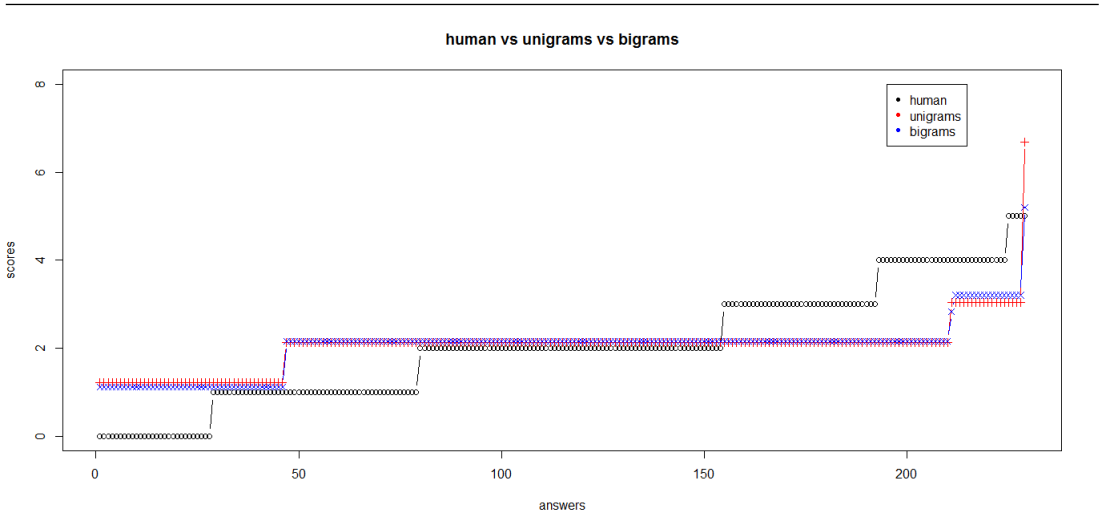
Pode ser observado graficamente na Figura 13 semelhança entre as pontuações de avaliadores humanos e modelo base-line no intervalo de 3 a 5 pontos. Existe uma diferença de mais de dois pontos para pontuações abaixo de 2 pontos e nota-se uma pequena diferença para pontuação de 6 pontos. O gráfico mostra que que o comportamento de unigramas e bigramas foi praticamente o mesmo.

Os índices de acurácia considerando somente unigramas e bigramas foram 78.5 % e 75.37 %, respectivamente. Para combinar unigramas e bigramas em uma única variável utilizamos um modelo de regressão linear múltipla e, para esta variável, o índice de acurácia foi de 78.93 %. O índice de acurácia entre as pontuações atribuídas por dois avaliadores humanos para esta questão foi de 93.94 %.

Questão de Geografia

Os resultados do modelo base-line para a questão de Geografia são apresentados graficamente na Figura 14:

Figura 14 – Resultados do modelo base-line para a questão de geografia



Pode ser obsevado graficamente na Figura 14 que o modelo base-line diverge por 1(um) ponto do avaliador humano para pontuações de 0 a 2 pontos. Para pontuações intermediárias os dois modelos tem praticamente o mesmo comportamento, e apresentam uma pequena diferença na pontuação de 5 pontos. O gráfico também mostra que o comportamento de unigramas e bigramas foi praticamente o mesmo.

Os índices de acurácia considerando somente unigramas e bigramas foram 83.89 % e 83.77 %, respectivamente. Para combinar unigramas e bigramas em uma única variável utilizamos um modelo de regressão linear múltipla e, para esta variável, o índice de acurácia foi de 83.85 %. O índice de acurácia entre as pontuações atribuídas por dois avaliadores humanos para esta questão foi de 83.93 %.

O modelo base-line alcançou um índice de acurácia próximo de avaliadores humanos para as respostas da questão de Geografia, o mesmo não acontecendo para as respostas da questão de Biologia.

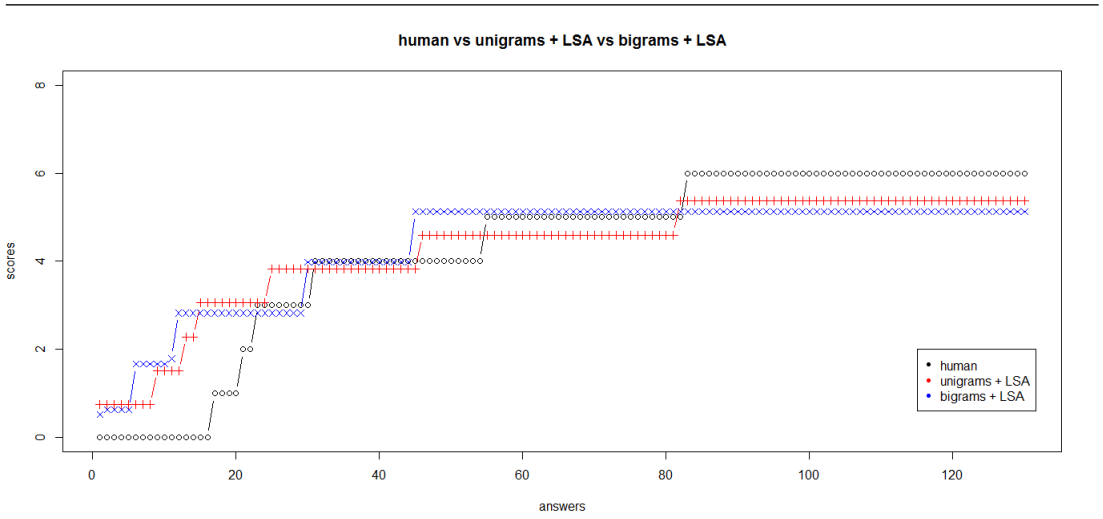
5.3.2 Modelo LSA

O modelo LSA estima a similaridade entre as respostas após a aplicação de um fator de penalização. O modelo também considera combinações no pré-processamento das respostas.

Questão de Biologia

Os resultados do modelo LSA para a questão de Biologia são apresentados graficamente na Figura 15:

Figura 15 – Resultados do modelo LSA para a questão de Biologia



Pode ser obsevado graficamente na Figura 15 uma pequena diferença entre avaliador humano e modelo LSA para a pontuação 0(zero) ponto. Entretanto pode ser visto uma importante correlação entre as curvas no restante do gráfico.

Os índices de acurácia considerando unigramas com LSA e bigramas com LSA foram 83.07 % e 83.46 %, respectivamente. O índice de acurácia entre as pontuações atribuídas por dois avaliadores humanos para esta questão foi de 93.85 %. Não foi aplicado nenhum modelo de regressão linear múltipla.

Nota-se que a diferença entre o índice de acurácia do programa LSA e o índice de acurácia entre dois avaliadores humanos foi de 10.39 %, que é uma diferença considerável. Entretanto, as diferenças entre as pontuações do programa LSA e do avaliador Humano não foram estatisticamente significativas.

A Tabela 12 apresenta uma comparação das distribuições das pontuações atribuídas pelo modelo LSA e avaliador humano.

Tabela 12 – Distribuição das pontuações LSA × Avaliador humano - Biologia

		Humano							Total	%
		0	1	2	3	4	5	6		
LSA	0	3	1	0	0	0	1	0	5	
	1	2	1	0	0	0	0	0	3	
	2	3	1	0	2	1	0	0	7	
	3	3	0	1	4	2	2	3	15	
	4	0	0	0	2	7	4	6	19	
	5	2	0	0	0	2	5	4	13	
	6	3	1	1	0	12	16	35	68	
Total		16	4	2	8	24	28	48	130	
Número de coincidências									55	42.31 %
Número de valores próximos									38	29.23 %

Observa-se na Tabela 12 que o modelo LSA atingiu cerca de 71.54 % entre pontuações coincidentes e com uma diferença de 1 (um) ponto. A distribuição das pontuações revela que o modelo LSA atribuiu 5 pontuações 0(zero) e o avaliador humano 16, justificando a diferença no gráfico para esta pontuação. Nas pontuações de 1 até 5, o modelo LSA atribuiu 125 pontuações e o avaliador humano 114, justificando a proximidade no gráfico para este intervalo de pontuações. O modelo LSA superestimou a pontuação máxima 6.

As listas de pontuações a que se refere a Tabela 12 são dadas abaixo:

$$
\begin{aligned}
pontuacoes_{humano} = [&6,4,6,1,6,6,0,3,6,6,0,4,0,5,0,0,5,0,5,6,6,5,6,6,0,6,3,4,5,3,\\
&5,6,6,6,5,4,3,4,6,4,6,1,6,6,0,6,6,4,6,6,6,4,6,4,6,4,6,5,6,6,0,\\
&5,6,6,5,6,4,4,5,5,5,2,6,6,4,5,4,3,0,6,6,6,5,5,5,6,4,1,2,1,4,0,\\
&4,6,6,5,0,0,5,4,3,5,6,0,3,4,0,6,6,4,5,4,5,5,5,6,4,6,4,0,5,4,6,\\
&6,6,5,6,5,3,5]
\end{aligned}
$$

$$
\begin{aligned}
pontuacoes_{LSA} = [&5,5,5,1,1,5,0,5,6,6,0,4,2,4,1,4,5,0,1,5,4,6,4,4,0,2,3,6,5,\\
&4,6,6,6,6,6,3,2,4,5,6,6,6,3,3,2,6,6,4,6,6,5,5,6,6,6,5,6,4,4,4,0,\\
&6,5,5,4,4,5,0,4,4,2,5,5,6,4,6,6,2,2,6,5,6,6,5,3,4,6,1,3,2,3,4,\\
&6,6,6,6,1,3,6,5,4,5,5,2,4,6,0,5,4,4,5,4,6,6,4,6,4,5,6,0,6,6,4,\\
&2,5,5,6,0,4,6]
\end{aligned}
$$

Foi realizado um teste de hipótese para a diferença entre as médias das pontuações *humano* e LSA. As médias *humano* e lsa são dadas por:

$$\overline{humano} = 4,2769 \text{ e } \overline{LSA} = 4,1154,$$

sendo os desvios padrão dados por:

$$\sigma_{humano} = 1,9962 \text{ e } \sigma_{LSA} = 1,8243$$

O teste será com intervalo de confiança no nível de 95%. A hipótese a ser testada é que não existe diferença estatisticamente significativa entre as médias das pontuações *humano* e LSA. Como as amostras são grandes usaremos o teste z para amostras conjugadas. Como o intervalo de confiança é no nível de 95%, pela tabela da distribuição normal temos que $z = 1,96$.

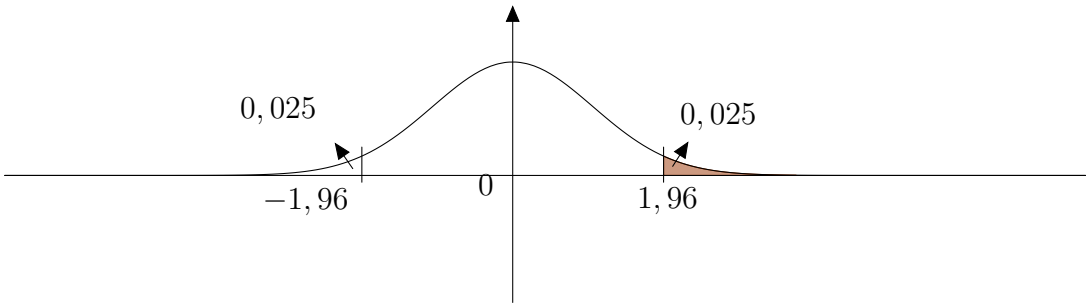


Figura 16 – Distribuição normal

A hipótese nula é dada por

$$H_0\text{: } \mu_{humano} = \mu_{LSA},$$

e a hipótese alternativa por

$$H_1\text{: } \mu_{humano} \neq \mu_{LSA},$$

Calculando o valor do z_{teste}, obtemos:

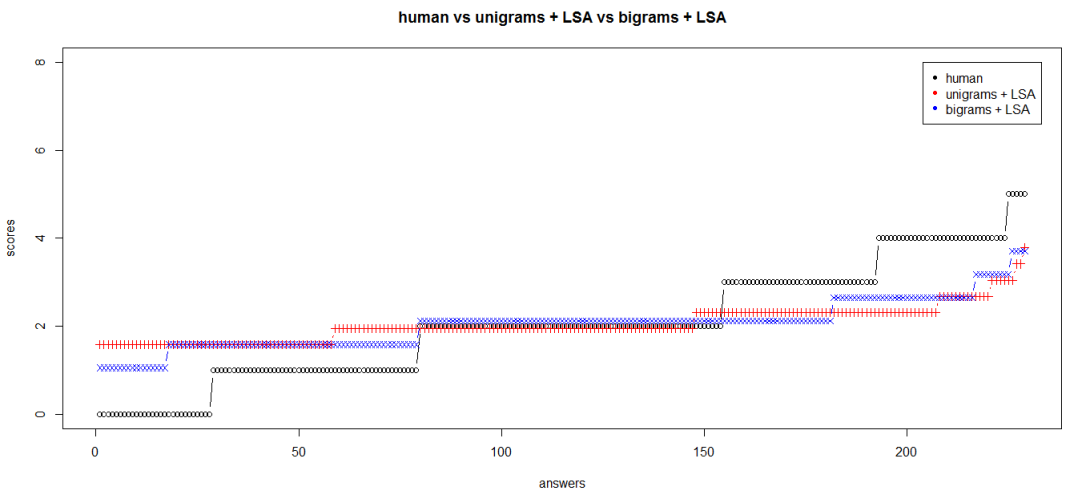
$$z_{teste} = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma^2_{humano}}{n_{humano}} + \frac{\sigma^2_{LSA}}{n_{LSA}}}} = \frac{4,2769 - 4,1154}{\sqrt{\frac{(1,9962)^2}{130} + \frac{(1,8243)^2}{130}}} = 0,6811.$$

O valor do z_{teste} caiu dentro da área de aceitação da hipótese nula; assim, a hipótes nula H_0 não foi rejeitada, e portanto, a diferença entra as médias das pontuações *humnano* e LSA não é estatísticamente significativa para as respostas de Biologia.

Questão de Geografia

Os resultados do modelo LSA para a questão de Geografia são apresentados graficamente na Figura 17:

Figura 17 – Resultados do modelo base-line para a questão de geografia



Pode ser observado graficamente na Figura 17 uma pequena diferença entre a pontuação 0 (zero) ponto entre avaliador humano e modelo LSA. Observa-se também pequena diferença para a pontuação 4 pontos. O gráfico mostra similaridades entre modelo LSA e avaliador humano para todas as demais pontuações.

Os índices de acurácia considerando somente unigramas e bigramas foram 84.94 % e 84.15 %, respectivamente. O índice de acurácia entre as pontuações atribuídas por dois avaliadores humanos para esta questão foi de 83.93 %. Não foi aplicado nenhum modelo de regressão linear múltipla.

Para melhor compreensão dos resultados obtidos para as respostas da questão de Geografia foi feita uma comparação das distribuições das pontuações entre as pontuações atribuídas pelo modelo LSA e por avaliadores humanos.

A Tabela 13 apresenta a distribuição das pontuações atribuídas pelo avaliador humano e pelo programa LSA

Tabela 13 – Distribuição das pontuações LSA × Avaliador humano - Geografia

		Humano							Total	%
		0	1	2	3	4	5	6		
LSA	0	10	2	2	0	0	0	0	14	
	1	11	24	12	6	3	1	0	57	
	2	4	13	33	6	9	0	0	65	
	3	3	7	16	14	5	1	0	46	
	4	0	4	12	11	8	1	0	36	
	5	0	1	0	1	5	1	0	8	
	6	0	0	0	0	2	1	0	3	
Total		28	51	75	38	32	5	0	229	
Número de coincidências									100	43.67 %
Número de valores próximos									83	36.24 %

Observa-se na Tabela 13 que o modelo LSA atingiu cerca de 79.91 % entre pontuações coincidentes e com uma diferença de 1 (um) ponto. A distribuição das pontuações revela que o modelo LSA atribuiu metade das pontuações do avaliador humano para pontuações abaixo de 1(um) ponto, justificando a diferença no gráfico para esta pontuação. O modelo LSA subestimou a pontuação mínima 0 (zero). Existe uma relativa equivalência na distribuição das pontuações para os demais intervalos. A diferença entre as pontuações não foi estatisticamente significativa.

5.3.3 Modelo base-line vs Modelo LSA

Questão de Biologia

Os índices de acurácia obtidos pelo modelo LSA foram melhores: 78.5 % vs 83.07 % na abordagem unigramas e 75.37 % vs 83.46 % na abordagem bigramas. Os números mostram que o modelo LSA é uma medida de similaridade mais fina com relação ao modelo base-line. Uma possível explicação reside no fato de que modelos n-gramas estima a similaridade entre uma resposta e a resposta de referência considerando apenas os unigramas ou bigramas comuns as duas respostas, enquanto que o modelo LSA considera a força de ligação entre as palavras. A diferença no índice de acurácia entre o modelo LSA e o avaliador humano foi um pouco mais de 10 %, o que ainda não é desejável. Uma possível razão para isto é o fato de que as respostas de Biologia foram comparadas com uma única resposta dada por um especialista humano.

Questão de Geografia

Os números mostram que o modelo LSA e o modelo base-line tiveram praticamente a mesma performance: 8494 % vs 83.89 % para a abordagem unigramas e 84.15 % vs 83.77 % para a abordagem bigramas. Uma possível razão para isto foi o fato de que a resposta de referência foi a concatenação de três respostas que obtiveram as maiores pontuações

dadas por avaliadores humanos. A distribuição das pontuações revela que o modelo LSA atribuiu pontuações variando de 1 a 4 para 37 respostas, enquanto que o avaliador humano 38 para o mesmo intervalo. O modelo LSA atribuiu pontuações variando de 5 a 6 para 85 respostas, enquanto que o avaliador humano 76 para o mesmo intervalo. Isto indica que o modelo LSA superestimou as pontuações maiores, porém foi bastante similar para as pontuações menores.

Pode ser concluido que o modelo base-line não teve a mesma performance nos dois conjuntos de respostas, enquanto que o modelo LSA teve a mesma performance, o que evidencia que o modelo LSA é mais robusto.

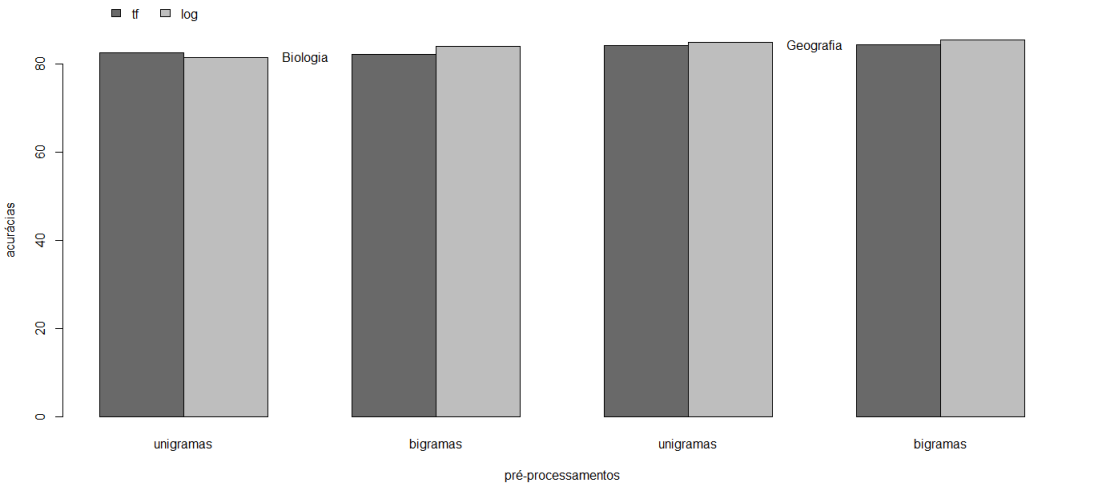
5.3.4 Desempenho dos parâmetros envolvidos

A calibração de parâmetros para que um modelo LSA possa atingir uma acurácia acima de 90 % ainda é um desafio. Nesta subseção é Apresentado um panorama do desempenho de cada tipo de parâmetro envolvido nesta pesquisa. Foram utilizaos dois tipos de pré-processamentos, **unigramas** e **bigramas**, combinados com três técnicas: considerar todas as stop words, remover todas as stop words e remover todas as stop words + stemming. Será feita uma avaliação do desempenho dos parâmetros envolvidos considerando a melhor configuração combinando as três técnicas para cada tipo de pré-processamento.

Ponderação local

Foram utilizadas duas ponderações locais: *termo-frequência* - ***tf*** e *logaritmo* - ***log***. O gráfico da Figura 18 mostra o desempenho do sistema variando apenas o parâmetro ponderação local na melhor configuração obtida.

Figura 18 – Desempenho do parâmetro ponderação local

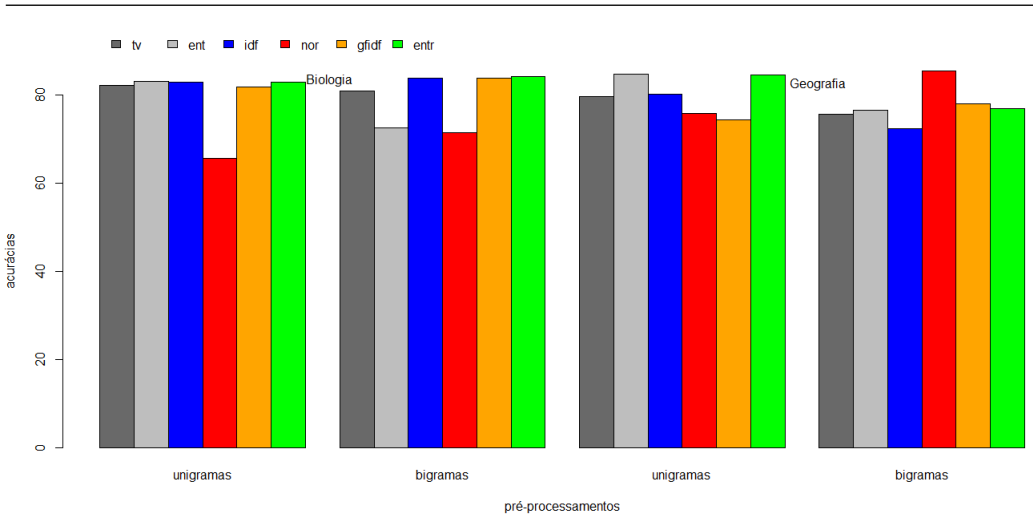


Da Figura 18 percebe-se que o parâmetro ponderação local teve praticamente o mesmo desempenho nos dois tipos de pré-processamentos para ambas questões de Biologia e Geografia.

Ponderação Global

Foram utilizadas seis ponderações globais: *trivial* - ***tv*** e *entropia* - ***ent***, *frequência documento inverso* - ***idf***, *normal* - ***nor***, *frequência global* - ***GfIdf*** e *entropia real* - ***entr***. O gráfico da Figura 19 mostra o desempenho do sistema variando apenas o parâmetro ponderação global na melhor configuração obtida.

Figura 19 – Desempenho do parâmetro ponderação local

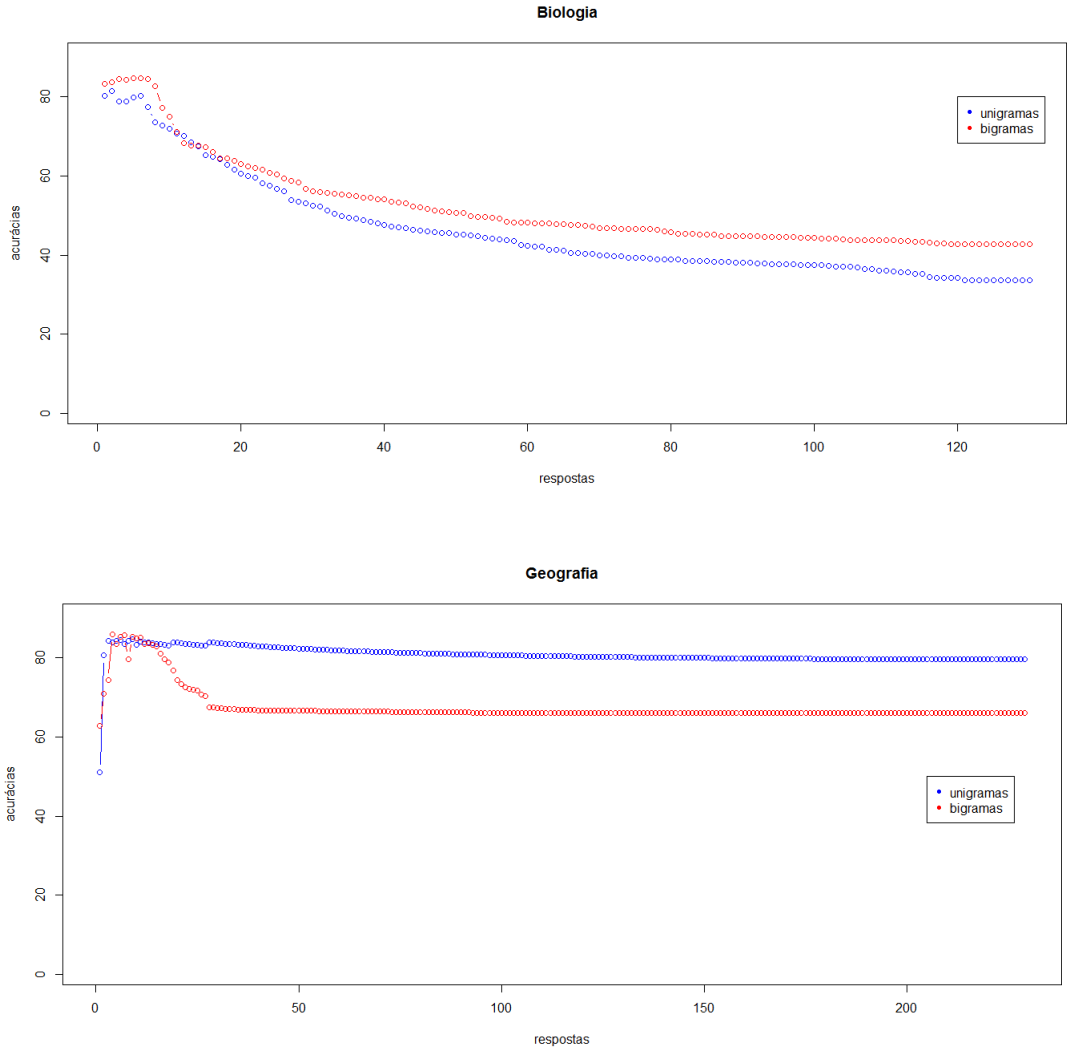


Da Figura 19 percebe-se que para a questão de Biologia o parâmetro ponderação global teve praticamente o mesmo desempenho nos dois tipos de pré-processamentos, exceto para a ponderação global normal, enquanto que para a questão de Geografia a ponderação global entropia teve o melhor desempenho para o pré-processamento e a ponderação global normal para bigramas.

Dimensionalidade

A variação do parâmetro k, que é a dimensão do espaço semântico (ou espaço latente), tem o impacto mais significativo sobre a eficiência de modelos LSA, pois é neste espaço que se compara dois textos. Nesta pesquisa foi considerada a variação de k desde 2 até o número de respostas para cada um dos dois conjuntos de respostas, como foi feito no trabalho de Wild et al. (2005). A Figura 20 apresenta o desempenho do parâmetro k em relação ao número de respostas.

Figura 20 – dimensão × acurácia



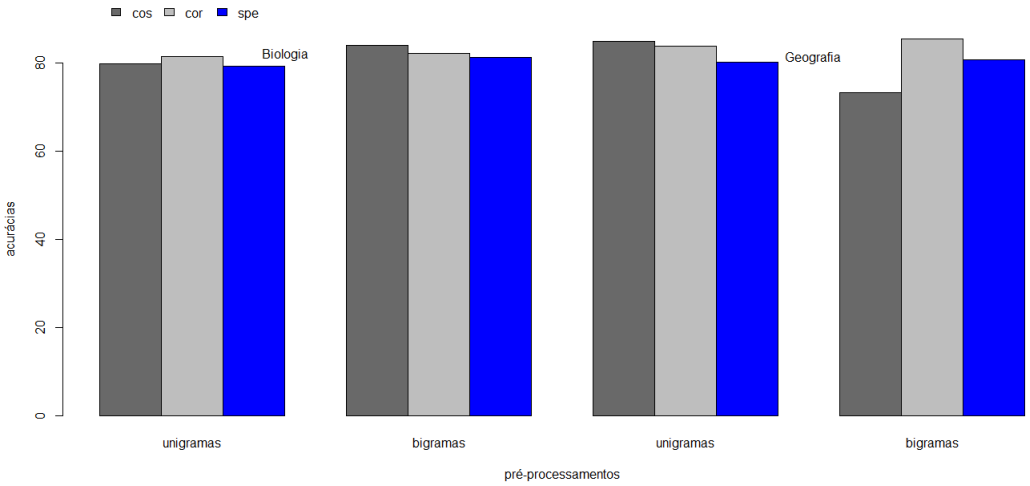
Pode ser observado graficamente na Figura 20 que os melhores resultados foram obtidos para valores baixos de k para os dois conjuntos de respostas. Duas razões explicam este fato. A primeira razão é que a SVD ordena os autovetores da maior para a menor magnitude do valor singular associado. A outra razão é o fato de que o conjunto D dos possíveis valores de k é discreto e limitado, portanto compacto, e a função que associa a cada valor de D uma acurácia ser uma função contínua, assumindo portanto, máximo em D.

Medidas de similaridade

Foram utilizadas três medidas de similaridade: *cosseno* - **cos**, *correlação de Pearson* - **cor** e *correlação de Spearman* - **spe**. O gráfico de Figura 21 mostra o desempenho so sistema variando apenas o parâmetro medida de similaridade na melhor configuração obtida.

Usando separadamente cada uma dessas medidas observa-se graficamente na Figura 21

Figura 21 – Desempenho do parâmetro medida de similaridade



que a correlação de Pearson obteve o melhor resultado, embora o cosseno seja a medida de similaridade mais utilizada em modelos LSA.

CONCLUSÃO

A principal meta desta pesquisa foi a realização de estudos que viabilizem o desenvolvimento de uma tecnologia LSA para uso prático em avaliações automáticas de respostas escritas a questões discursivas. Algumas direções principais foram direcionadas:

No sentido de tornar o sistema mais robusto foi feito o pré-processamento com bigramas de palavras. Um sistema tradicional de LSA não considera a ordem das palavras ficando assim vulnerável a possíveis fraudes: um estudante bem informado sobre o seu funcionamento pode enganar o sistema, por exemplo, com a repetição de uma determinada palavra. Imaginando que o uso de bigramas pode tornar o sistema mais imune a este tipo de fraude, pois considera a ordem de palavras. Em experimentos foi observado que considerando frequências de bigramas como componentes dos vetores coluna da matriz inicial obteve-se melhores resultados para as respostas da questão de Biologia. Entretanto não teve qualquer impacto significativo na avaliação das respostas para a questão de Geografia. Em ambos os casos, o uso de bigramas não trouxe qualquer melhora significativa em comparação com unigramas.

Sabe-se que a eficiência de um modelo LSA depende fortemente da calibração de seus parâmetros e do domínio de aplicação. Isto dificulta reaplicar experimentos bem sucedidos num domínio em outro domínio. Nesta pesquisa foram utilizados dois domínios de aplicação: uma questão discursiva de natureza conceitual de Biologia e outra de natureza argumentativa de Geografia. Os melhores resultados foram obtidos com as respostas para a questão de Geografia, embora fosse usada a mesma arquitetura do modelo LSA para ambos conjuntos de respostas. A diferença foi o fato de ter sido considerada para a questão de Geografia a resposta de referência como sendo as três melhores respostas mais bem avaliadas por especialistas humanos, enquanto para a questão de Biologia foi um texto dado por um especialista humano. Diante disso pode ser concluido que:
i) considerar o texto de referência como sendo uma parte do próprio corpus forneceu os melhores resultados nesta pesquisa. A generalização deste fato para outros domínios é uma questão que deve ser investigada;
ii) Foi utilizado o mesmo programa LSA para os dois domínios estudados; diante disso pode ser imaginado que com uma calibração de parâmetros adequada, a portabilidade de um programa LSA pode ser bem sucedida em mais de um domínio de aplicação.

Foi visto na literatura que um modo de avaliar o desempenho de um modelo LSA é

comparar a acurácia do programa LSA com a acurácia entre dois especialistas humanos. Ocorreram duas situações distintas: i) para a questão de Biologia a diferença entre os índices de acurácia do avaliador humano e do programa LSA foi um pouco mais de 10 %. Uma provável justificativa para esta diferença foi o fato de ter considerado a resposta referência como sendo um único texto dado por um especialista humano. Provavelmente o resultado seria melhor se fossem considerados mais textos na resposta referência, ou então ter considerado a resposta referência como sendo as respostas mais bem avaliadas por especialistas humanos. Foi fixada a ideia de avaliar o desempenho do sistema através da comparação das respostas com um texto dado por um especialista humano, porisso não foram consideradas as outras possibilidades; ii) para a questão de Geografia os índices de acurácia entre dois especialistas humanos e do programa LSA foi praticamente o mesmo. Uma provável justificativa para este fato foi considerar como resposta referência a concatenação das respostas mais bem avaliadas por especialistas humanos. Uma conclusão é que o sistema tem melhor desempenho quando ocorre um aumento do vocabulário da resposta referência. É claro que não se pode concluir que isto deva ocorrer em outros domínios.

Optou-se também em verificar o desempenho do modelo LSA em relação ao desempenho de um sistema baseado apenas em unigramas e bigramas de palavras. Este modelo base-line considera tão somente a co-ocorrência de unigramas e bigramas. Para a questão de Biologia o modelo LSA teve melhor desempenho, enquanto que para a questão de Geografia o desempenho de ambos os sistemas foi praticamente o mesmo. Uma causa para isto foi provavelmente o aumento do vocabulário para a resposta referência da questão de Geografia. Em ambos os casos verificou-se que o modelo LSA foi superior tanto na abordagem unigrama quanto na abordagem bigramas.

Nos primeiros experimentos observou-se uma grade discrepância entre as pontuações do programa LSA em relação a pontuação do avaliador humano para algumas respostas com poucas palavras: se todas as palavras de uma destas respostas estivesse na resposta referência o programa atribuía a pontuação máxima, enquanto que o avaliador humano atribuía a pontuação mínima. Para corrigir esta discrepância foi aplicado um fator de penalização para respostas com um número reduzido de palavras. Ao rodar novamente os experimentos foi observado que a aplicação deste fator proporcionou um aumento de 8 % a 10 % nos índices de acurácia. Foram encontrados trabalhos na literatura que resolveram este problema usando a distância euclidiana como medida de similaridade. Com experimentos usando apenas a distância euclidiana os resultados não foram satisfatórios. Provavelmente isto tenha ocorrido visto que a distância euclidiana considera as escalas das medições: o programa LSA atribui uma pontuação no intervalo $[0, 1]$, enquanto o avaliador humano atribuía uma pontuação inteira entre 0 e 6.

Foi feita uma analise comparativa entre as distribuições das pontuações do avaliador *humano* e do programa *LSA*. Para as questões de Biologia e Geografia verificou-se que o modelo LSA atingiu cerca de 71.54 % e 79 %, respectivamente, de pontuações coincidentes ou com uma diferença de um ponto, o que não caracteriza discrepância. Em ambos os casos foi realizado um teste ao nível de significância de 1% com a hipótese de que não existe diferença estatísticamente significativa entre as médias das pontuações *humano* e *LSA*. Foi Ccncluido, em ambos os casos, que a hipótese nula não foi rejeitada, o que mostra consistência nos resultados obtidos.

Os resultados mostraram ainda que LSA pode ser utilizado para refinar resultados de métodos baseados unicamente em n-gramas, e com uma adequada calibração de seus parâmetros no domínio de aplicabilidade pode alcançar resultados próximos de avaliadores humanos. No caso da questão de Geografia, os experimentos forneceram um índice de acurácia de 84,94 % das pontuações LSA contra 84.93 % para pontuações de avaliadores humanos. Estes resultados mostram que a tecnologia LSA esta atingindo um grau de eficiência para uso prático em sistemas de avaliação automáticos em ambientes virtuais de aprendizagem.

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Recuperacao de Informacao: Conceitos e Tecnologias nas Maquinas de Busca*. [S.l.]: Bookman, 2013. Citado 2 vezes nas páginas 34 e 41.

BURSTEIN, J.; CHODOROW, M.; LEACOCK, C. Automated essay evaluation: The criterion online writing service. *Al Magazine*, v. 3, p. 27–36, 2010. Citado 2 vezes nas páginas 14 e 15.

CHALI, Y.; HASAN, S. A. (Ed.). *Automatically Assessing Free Texts*. Workshop on Speech and Language Processing Tools in Education: The COLING 2012 Organizing Committee, 2012. Citado na página 37.

DEERWESTER, S. et al. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, v. 41, p. 391– 407, 1990. Citado 4 vezes nas páginas 16, 21, 33 e 34.

DIANA, P. et al. Automatic assessment of students free text answers underpinned by the combination of a b leu inspired algorithm and latent semantic analysis. *Machine Translation*, 2005. Citado na página 37.

DUMAIS, S. T. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 1991. Citado 2 vezes nas páginas 36 e 72.

EMYGDIO, B. M. et al. Efficiency of similarity coefficients based on rapd markers in common bean genotypes. *Pesquisa Agropecuaria Brasileira*, v. 38, p. 243–250, 2003. Citado na página 63.

FERNANDES, J.; ARTIFICE, A.; FONSECA, M. Automatic estimation of the lsa dimension. In: *Kdir*. [S.l.: s.n.], 2012. Citado 5 vezes nas páginas 23, 37, 38, 42 e 58.

FOLTZ, P. W. Latent semantic analysis for text based research. *Behavior Research Methods Instruments and Computers*, v. 28, p. 197–202, 2009. Citado na página 72.

FOLTZ, P. W.; LAHAM, D.; LANDAUER, T. K. The intelligent essay assessor: Applications to educacional technology. *Interactive Multimedia Education Journal of Computer enhanced learning*, 1999. Citado na página 37.

HALEY, D. T. et al. Seeing the whole picture: evaluating automated assessment systems. *ITALICS*, v. 1, p. 203–224, 2007. Citado 2 vezes nas páginas 17 e 94.

HE, Y.; HUI, S. C.; QUAN, T. T. Automatic summary assessment for intelligent tutoring systems. *Computers & Education*, v. 53, p. 890–899, 2009. Citado 2 vezes nas páginas 14 e 37.

HEARST, M. A. The debate on automated essay grading. *IEE Intelligeng Systems archive*, v. 15, p. 22–37, 2000. Citado na página 15.

HU, X. et al. Lsa : First dimension and dimensional weighting. *Computer and Information Science*, 2003. Citado na página 37.

ISLAN, M. M.; HOQUE, M. L. Automated essay scoring using generalized latent semantic analysis. *Journal of Computers*, v. 7, p. 616–626, 2012. Citado 2 vezes nas páginas 14 e 37.

JORGE-BOTANA, G. et al. Latent semantic analysis parameters for essay evaluation using small scale corpora. *Journal of Quantitative Linguistics*, v. 1, p. 1–29, 2010. Citado 6 vezes nas páginas 17, 37, 44, 72, 73 e 75.

JORGE-BOTANA, G. et al. Automated lsa assessment of summaries in distance education: Some variables to be considered. *Journal of Educational Computing*, 2015. Citado na página 37.

KANEJIYA, D.; KUMAR, A.; PRASAD, S. Automatic evaluation of students answers using syntactically enhanced lsa. In: *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing - Volume 2*. [S.l.: s.n.], 2003. Citado na página 37.

KITCHENHAM, B. *Procedures for Performing Systematic Reviews*. 2004. Citado na página 33.

KLEIN, R.; KYRILOV, A.; TOKMAN, M. Automated assessment of short free-text responses in computer science using latent semantic analysis. *ACM Press*, 2011. Citado na página 37.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. *Discourse processes*, v. 25, p. 259–284, 1998. Citado na página 56.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. *Discourse processes*, 1999. Citado na página 72.

LANDAUER, T. K. et al. Information retrieval using a singular value decomposition model of latent semantic structure. *SIGIR '88 Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, 1988. Citado 3 vezes nas páginas 34, 36 e 38.

LAY, D. C. *Linear Algebra And Its Applications*. [S.l.]: English, 2012. Citado na página 24.

LAYFIELD, C. With lsa does matter. In: *EMS*. [S.l.: s.n.], 2012. Citado na página 37.

LEON, J. A. et al. Exploring the assessment of summaries: Using latent semantic analysis to grade summaries written by spanish students. *Procedia - Social and Behavioral Sciences*, 2013. Citado na página 37.

LIFCHITZ, A. Effect of tuned parameters on a lsa multiple choice questions answering model. *Behavior Research Methods*, v. 41, p. 1201–1209, 2009. Citado 2 vezes nas páginas 17 e 54.

LO, R. T.-W.; OUNIS, B. H. I. Automatically building a stopword list for an information retrieval system. *Journal of Digital Information Management*, 2005. Citado na página 50.

MAGLIANO, J. P.; GRAESSER, A. C. Computer-based assessment of student-constructed responses. *Behavior Research Methods*, v. 44, p. 608–621, 2012. Citado 2 vezes nas páginas 14 e 72.

MALANDRAKIS, N.; IOSIF, E.; POTAMIANOS, A. Deeppurple: estimating sentence semantic similarity using n-gram regression models and web snippets. *Association for Computational Linguistics*, 2012. Citado na página 75.

MARIN, D. P.; NIETO, I. P.; RODRIGUEZ, P. Natural language processing meets user modeling for automatic and adaptive free-text scoring. *Procesamiento del Lenguaje Natural*, v. 41, p. 225–232, 2008. Citado na página 14.

MARINREZ, D. P. et al. About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. *Revista Signos*, 2005. Citado 4 vezes nas páginas 15, 37, 72 e 75.

NAKOV, P.; POPOVA, A.; MATEEV, P. Weight functions impact on lsa performance. In: *EuroConference RANLP2001(Recent Advances in NLP)*. [S.l.: s.n.], 2001. Citado 4 vezes nas páginas 17, 37, 55 e 72.

NAVEGA, S. (Ed.). *Manipulacao SSemantic de Textos. Os Projetos WordNet e LSA*. [S.l.]: Intelliwise, 2004. Citado na página 16.

NOORBEHBAHANI, F.; KARDAN, A. A. The automatic assessment of free text answers using a modified bleu algorithm. *Computer & Education*, v. 56, p. 337–345, 2011. Citado 2 vezes nas páginas 15 e 17.

OLMOS, R. et al. Using latent semantic analysis to grade brief summaries: some proposals. *International Journal of Continuing Engineering Education and Life-Long Learning*, 2011. Citado 3 vezes nas páginas 37, 73 e 75.

ORENGO, V. M.; HUYCK, C. A stemming algorith. *International Symposium on String Processing and Information Retriviel*, p. 186–193, 2001. Citado na página 51.

PAGE, E. B. The imminence of grading eessay by computer. *The Phi Delta Kappan*, v. 47, p. 238–243, 1966. Citado 2 vezes nas páginas 15 e 33.

RAMACHANDRAN, L.; GEHRINGER, E. F. Automated assessment of review quality using latent semantic analysis. *IEEE*, 2011. Citado na página 37.

REFAAT, M. M. et al. Automated assessment of students arabic free-text answers. *International journal of intelligent computing and information sciences*, v. 12, p. 213–222, 2012. Citado 2 vezes nas páginas 14 e 37.

REHDER, B. et al. Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse processes*, 1998. Citado na página 36.

RUSSELL, R. et al. *Issues and Challenges in Conducting Systematic Reviews to Support Development of Nutrient Reference Values: Workshop Summary*. [S.l.]: Technical Reviews, No. 17.2, 2009. Citado na página 33.

SALTON, G.; WONG, A.; YANG, C. A vector space model for automatic indexing. *Communications of the ACM*, v. 18, p. 613 – 620, 1975. Citado na página 62.

STEINHART, D. J. *Summary Street: An Intelligent Tutoring System for Improving Student Writing through the Use of Latent Semantic Analysis*. Tese (Doutorado) — University of California, 2001. Citado na página 37.

WHITTINGTON, D.; HUNT, H. (Ed.). *Approaches to the Computerized Assessment of Free Text Responses*. Loughborough University: 3rd International Computer Assisted Assessment Conference, 1999. Citado na página 33.

WIEMER-HASTINGS, P. Adding syntactic information to lsa. In: *PROCEEDINGS OF THE 22ND ANNUAL CONFERENCE OF THE COGNITIVE SCIENCE SOCIETY*. [S.l.: s.n.], 2000. Citado na página 37.

WILD, F. et al. Factors influencing effectiveness in automated essay scoring with lsa. In: *Proceeding of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*. [S.l.: s.n.], 2005. Citado 5 vezes nas páginas 23, 37, 58, 72 e 84.

WOLFE, M. B. W. et al. Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, 1998. Citado na página 72.

ZEN, K.; ISKANDAR, D. N. F. A.; LINANG, O. Using latent semantic analysis for automated grading programming assignments. In: *International Conference on Semantic Technology and Information Retrieval*. [S.l.: s.n.], 2011. Citado 3 vezes nas páginas 14, 37 e 72.

APÊNDICE A – RESUMOS DOS TRABALHOS PESQUISADOS

A.1 AVALIAÇÃO NA QUALIDADE DO ESTUDO

Foi feita no Capítulo 3 uma avaliação sistemática qualitativa dos trabalhos relacionados com sistemas de avaliação automática baseados em LSA aplicados em ambientes educacionais. A revisão foi feita basicamente descrevendo-se a metodologia de cada sistema seguindo o que foi sugerido no trabalho de Haley et al. (2007). Para cada trabalho será feito um pequeno quadro sobre o local, autores e resumo do trabalho. No resumo será descrito a proposta e inovação, os detalhes técnicos, bem como a calibração dos parâmetros, e os resultados obtidos. Esta avaliação será feita por ordem cronológica anual desde 1988 até 2015.

1988

1. Information Retrieval using a Singular Value Decomposition Latent Semantic Structure

Local University of Chicago, Bellcore

Autores Deerwester, Dumais, Furnas, Landauer, Harshman, Streeter, Lochbaum

Resumo O trabalho descreve um novo método para indexação e recuperação automática de informações. O modelo tende a melhorar as estimativas de associação entre palavras e textos diminuindo o tempo de consultas de documentos. O método usou o processo de decomposição a valores singulares para decompor a matriz palavra-texto e aproxima-la por uma outra matriz. Palavras e textos são representados em um espaço de dimensão menor em relação ao espaço das colunas da matriz palavra-texto. O modelo LSA foi testado em documentos de duas coleções, a MED [1] e a CISI [2].

[1] National Library of Medicine
[2] Institute for Scientific Information

1990

2. Indexing by Latent Semantic Analysis

Local University of Chicago, Bell Communications Research (NJ), University of Western Ontario

Autores Deerwester, Dumais, Furnas, Landauer, Harshman

Resumo O trabalho descreve um novo método para indexação e recuperação automática de informações de textos. O método pretende superar métodos que usam apenas frequência de palavras para recuperação de informações. A abordagem considera uma estrutura semântica que esta implícita na associação entre as palavras dos textos. O trabalho afirma que este método melhora a filtragem de textos relacionados com as palavras que compõe uma consulta. O trabalho inova com a explicação do método SVD e dos passos da redução da dimensionalidade. O modelo LSI foi aplicado em 1033 resumos médicos num total de 5823 palavras, utilizou a técnica de remoção de stop words no pré-processamento, não utilizou nenhum esquema de ponderação, dimensionalidade igual a 100, e usou o cosseno do ângulo como medida de similaridade. Em dois níveis de recuperação, a precisão do modelo LSI ficou acima de modelos que consideram apenas frequências de palavras.

1991

3. Improving the retrieval of information from external sources

Local Bellcore

Autor Susan T Dumais

Resumo O trabalho relata que a principal barreira para uma recuperação de informação bem sucedida é a grande variabilidade no significado de palavras. A mesma ideia pode ser descrita por palavras diferentes, e vice-versa, uma mesma palavra pode ser usada para significar coisas diferentes. O trabalho propõe um método estatístico chamado indexação semântica latente, que modela uma estrutura mais ampla implícita na ordem e associação das palavras, o que melhora o desempenho de recuperação de informação em até 30%. O trabalho inova com o uso de funções ponderação, pois o seu uso provoca uma melhora adicional de 40% a 67%. O sistema é aplicado em resumos médicos, informações científicas e em textos de ciência da computação; o número de resumos analisados variou de 82 até 1460, com o número de palavras variando de 374 a 5831. O sistema usou log-entropia como ponderação e a dimensionalidade variou de 60 a 100. O método LSI melhora em torno de 20% métodos baseados apenas em frequências de palavras.

1992

4. Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval

Local Bellcore

Autor Susan T Dumais

Resumo O trabalho descreve melhorias para o método LSI. Afirma que uma função ponderação adequada melhora o desempenho do sistema em torno de 40%. O sistema é aplicado em 150 testes de ciência da informação e utilizou vários esquemas de ponderação, sendo a dimensionalidade igual a 100. O método obteve uma precisão de quase 80% com o esquema de ponderação log-entropia.

1994

5. Using linear algebra for intelligent information retrieval

Local Department of Computer Science, University of Tennessee

Autores M.W. Berry, S.T. Dumais and G.W. O'Brien

Resumo O trabalho afirma que a maioria das abordagens para recuperar informações textuais depende de uma harmonia lexical entre as palavras que compõe o próprio texto. Como existe uma grande diversidade no uso das palavras, os métodos lexicais são incompletos e imprecisos. O trabalho afirma ainda que o método de decomposição a valores singulares (SVD) aproveita a estrutura de ordem superior implícita na associação entre as palavras. O método é chamado indexação semântico latente pois o espaço onde ocorre as relações associativas entre as palavras não é evidente. O método traz a inovação de mostrar como o SVD é atualizado. O sistema é aplicado em coleções de textos em mais de uma língua para recuperação de informações. O sistema usou log-entropia como ponderação e a dimensionalidade variou de 70 a 100. Para tarefas de recuperação o método LSI mostrou uma melhora de 16% em relação a métodos que usam palavras-chave.

1996

6. Latent semantic analysis for text-based research

Local New Mexico State University, Slippery Rock U, U of Pittsburgh

Autores Foltz, Britt, Perfetti

Resumo O trabalho resume três experiências que ilustram como LSA pode ser usado em análises de textos. Duas experiências descrevem métodos para analisar o assunto de um determinado ensaio e determinar de que texto o assunto foi retirado. O sistema classifica a qualidade da informação mencionada no ensaio. A terceira experiência descreve o uso de LSA para medir a coerência e a compreensão dos textos. O sistema foi aplicado em 607 ensaios sobre o canal do Panamá num total de 4829 palavras. O sistema não utilizou nenhum esquema de ponderação, dimensionalidade igual a 100 e o cosseno como medida de similaridade utilizada. O sistema obteve uma correlação de 0.68 comparado com avaliadores humanos.

1997

7. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge

Local University of Colorado, Bellcore

Autores Susan T. Dutnais and Thomas K Landauer

Resumo O trabalho propõe uma solução para o problema de Plato usando LSA. Mesmo com o mínimo de informações as pessoas adquirem conhecimentos. A aprendizagem de vocabulários de textos é um caminho para esta solução. A teoria LSA foi utilizada para aprendizagem de vocabulários e outros fenômenos psicolinguísticos. LSA usa a co-ocorrência de palavras para adquirir conhecimentos sobre o vocabulário. LSA usa SVD para redução para o espaço semântico e a escolha correta da dimensão deste espaço é de fundamental importância para eficiência do modelo. Simulações foram feitas para um grande número de dimensões. O trabalho conclui que com a variação da dimensão os resultados são irregulares.

8. How Well Can Passage Meaning be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans

Local University of Colorado, Boulder

Autores Thomas K. Landauer, Darrell Laham, Bob Rehder, and M. E. Schreiner

Resumo O trabalho investiga o quanto do significado de um texto podemos extrair não considerando a ordem das palavras que compõe o próprio texto. A investigação foi feita com a aplicação de um modelo estatístico para avaliação de respostas curtas de estudantes sobre temas científicos comparado com avaliação humanos. Surpreendentemente houve pouca diferença entre avaliadores humanos e e estimativa do modelo estatístico. O modelo foi aplicado em 94 ensaios escritos por alunos de pós-graduação sobre anatomia, função e finalidade do coração humano. Não foi utilizado nenhum esquema de ponderação e o cosseno foi a medida de similaridade utilizada. A correlação entre a pontuação de LSA e a média da pontuação entre os dois avaliadores humanos foi de 77%.

9. E-Assessment using Latent Semantic Analysis in the Computer Science Domain : A Pilot Study

Local Open University

Autores Petre Thomas and Debra Haley and Marian Petre

Resumo O trabalho diz que LSA tem seu foco na avaliação formativa em domínios gerais. A adequação de LSA para avaliação formativa no domínio de ciência da computação não é bem conhecida. O trabalho encoraja pesquisadores para o uso de LSA no domínio técnico de ciência da computação. O trabalho explica a teoria por trás de LSA, descreve algumas aplicações e apresenta os resultados usando LSA para avaliação automática de pequenos ensaios de uma classe de graduação de arquitetura de sistema computacional.

1998

10. An Introduction to Latent Semantic Analysis

Local University of Colorado at Boulder

Autores Thomas K Landauer and Peter W. Foltz and Darrell Laham

Resumo O trabalho faz uma introdução de LSA e afirma que LSA extrai através de cálculos estatísticos a representação do significado no uso de palavra. A ideia por traz de LSA é que o agrupamento de palavras produz um conjunto de ligações mútuas e pode determinar uma certa semelhança no significado das mesmas palavras. LSA tem sido adequada para refletir o conhecimento humano, como por exemplo, na comparação entre a pontuação LSA e a pontuação humana. LSA avalia a qualidade e a quantidade de conhecimento contido em um determinado ensaio. O trabalho relata que a técnica LSA foi aplicada em vários estudos; em um deles, LSA fez uma

avaliação de similaridades entre textos. O modelo foi aplicado em coleções de textos padronizadas. Não foi utilizado nenhum esquema de ponderação e a dimensionalidade variou entre 50 e 400. O cosseno foi a medida de similaridade utilizada. Em relação a métodos prévios, a melhora foi em torno de 30%.

11. Learning from text: Matching readers and texts by Latent Semantic Analysis

Local University of Colorado, Boulder, New Mexico State University

Autores Michael B. W. Wolfe and M. E. Schreiner and Bob Rehder and Darrell Laham and Peter W. Foltz and Walter Kintsch and Thomas K Landauer

Resumo O trabalho analisa a hipótese de que a capacidade do leitor de aprender um texto é função do conhecimento do leitor e das dificuldades das informações no próprio texto. Um modelo LSA foi aplicado avaliar a aprendizagem de estudantes em quatro textos sobre coração humano e sistema circulatório em diferentes níveis de dificuldades. Os resultados mostraram que houve maior aprendizagem dos estudantes para textos com nível de dificuldade mediana e que uma pré-leitura aumenta a qualidade da aprendizagem. Não foi utilizado nenhum esquema de ponderação e o cosseno e a correlação quadrática foram as medidas de similaridade utilizadas.

12. The Measurement of Textual Coherence with Latent Semantic Analysis

Local New Mexico State University, University of Colorado

Autores Peter W. Foltz and Walter Kintsch and Thomas K. Landauer

Resumo O trabalho usa LSA para medir a coerência de textos. Os textos são comparados no espaço semântico e o modelo fornece a similaridade entre eles. Foram re-analisados dois conjuntos de textos de dois estudos sobre coerência e compreensão de textos. O primeiro foi um livro de treinamento da força aérea do Vietnã e o segundo um livro sobre doenças cardíacas. O esquema de ponderação utilizado foi log-entropia e a dimensionalidade igual a 300. O cosseno foi a medida de similaridade utilizada. Os resultados indicam que o modelo é capaz de prever o efeito da coerência na compreensão dos textos.

13. Using Latent Semantic Analysis to assess knowledge: Some technical considerations

Local University of Colorado, Boulder

Autores Bob Rehder, M. E. Schreiner, Michael B. W. Wolfe, Darrell Laham, Thomas K Landauer, and Walter Kintsch

Resumo O trabalho afirma que LSA pode comparar um ensaio escrito por um aluno com um ou mais textos de referência através do cosseno do ângulo. Entretanto, afirma que considerar o papel desempenhado por termos técnicos pode melhorar o desempenho de um modelo LSA. A similaridade semântica entre um ensaio escrito por um aluno e um textos de referência é uma medida confiável de conhecimento do aluno. Dentre as características do ensaio escrito deve-se considerar o papel desempenhado pelos termos técnicos:

1. Foi construída uma lista de termos técnicos sobre o coração humano e sistema circulatório
2. As palavras dos ensaios foram separadas em dois conjuntos: termos técnicos e não-técnicos
3. O modelo LSA analisou separadamente cada um dos dois conjuntos de termos

Considerando o osseno como medida de similaridade, o modelo LSA analiso 106 ensaios escritos por estudantes. A correlação com o conjutnto de termos técnicos foi de 69%, com o conjunto de termos não-técnicos foi de 59% e com um texto de referência foi de 71%.

1999

14. The Intelligent Essay Assessor: Applications to Educational Technology

Local New Mexico State University, Knowledge Analysis Technologies, University of Colorado

Autores Peter W. Foltz and Darrell Laham and Thomas K. Landauer

Resumo O artigo descreve o aplicativo Intelligent Essay Assessor (IEA) para pontuação na qualidade do conteúdo de ensaios. O aplicativo é treinado inicialmente em textos de referência, em seguida caracteriza os ensaios de estudantes e compara os mesmos com os textos de referência. Em vários aspectos, as pontuações do IEA são as mesmas pontuações de especialistas humanos. Alcança uma correlação de 80% entre LSA e avaliadores humanos. O sistema não utilizou nenhum esquema de ponderação e o cosseno foi a medida de similaridade utilizada.

15. Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis

Local University of Memphis

Autores Peter Wiemer-Hastings and Katja Wiemer-Hastings and Arthur C. Graesser

Resumo O trabalho apresenta o Auto Tutor que é um tutor inteligente que interage com o aluno através de um diálogo de linguagem natural. A pesquisa aborda o problema da compreensão do que é falado pelo aluno. LSA foi a técnica escolhida para resolver este problema. O trabalho faz uma descrição geral e mostra como LSA pode ser usada em um sistema de tutoria. Os resultados são discutidos para um sistema de tutoria geral. O corpus foi constituído por trabalhos de estudantes em testes TOEFL e também por artigos e livros textos de ciência da computação. O esquema de ponderação utilizado foi log-entropia e o cosseno foi a medida de similaridade utilizada.

2000

16. Developing Summarization Skills through the Use of LSA-Based Feedback

Local University of Colorado, Platt Middle School, Boulder Colorado

Autores Eileen Kintsch and Dave Steinhart and Gerry Stahl and Cindy Matthews and Ronald Lamb

Resumo O trabalho descreve o desenvolvimento do Summary Street, que é um software educacional que usa LSA para apoiar a elaboração e revisão de atividades em sala de aula. O Summary Street fornece vários tipos de feedback. O artigo discute o processo colaborativo educacional inerente ao Summary Street. Um dos experimentos foi realizado com 830 textos num total de 17.688 palavras sobre o funcionamento do coração humano. O cosseno foi a medida de similaridade utilizada.

17. Getting Better Results With Latent Semantic Indexing

Local Informatics–Sofia University

Autor Preslav Nakov

Resumo O trabalho apresenta uma visão geral de alguns fatores que influenciam a qualidade dos resultados obtidos quando é usada análise semântica indexada. Os fatores são separados em cinco grupos e analisados separadamente e em conjunto. Alguns operadores lógicos, como E e OU são considerados. O corpus é constituído por textos religiosos e sacros. O esquema de ponderação utilizado foi termo frequência vs frequência documento inverso, a dimensionalidade foi igual a 23. A medida de similaridade foi definida como uma combinação linear entre o cosseno do ângulo, distância euclidiana, distância de manhattan, distância de minkowski e a correlação de Pearson.

18. Using Latent Semantic Analysis to Evaluate the Contributions of Students in AutoTutor

Local University of Memphis

Autores Arthur C. Graesser and Peter Wiemer-Hastings and Katja Wiemer-Hastings and Derek Harter and Natalie Person and Tutoring Research Group

Resumo O trabalho aplica LSA como componente do mecanismo que avalia a qualidade das contribuições de estudantes no diálogo tutorial do sistema AutoTutor. O modelo LSA descreve classes diferentes de alunos em relação a suas contribuições; bom, vaga, errada. O modelo usa dois livros textos e 30 artigos de computação como referência. A dimensionalidade foi igual a 200. O corpus foi 192 respostas de estudantes de computação. A correlação entre LSA e a média de quatro especialistas humanos foi de 49%, enquanto a correlação entre dois especialistas humanos foi de 78 %.

19. Adding syntactic information to LSA

Local University of Edinburgh, Edinburgh EH8 9LW Scotland

Autores Peter Wiemer-Hastings

Resumo Muitos esforços tem sido empreendidos para extrair automaticamente a estrutura sintática de um texto. Não existe um consenso sobre a importância da informação sintática na representação do significado. Um modelo LSA extrai o significado da informação sem levar em conta a ordem das palavras, o que torna secundário o papel da sintaxe. Porém LSA a performance equivalente a seres humanos para certos significados. A adição da informação sintática pode ajudar LSA? Este trabalho tenta responder esta pergunta. Foram realizados dois experimentos: no primeiro, foi calculada a similaridade entre os verbos de uma sentença escrita por um aluno e uma sentença de referência. No segundo, foi dada uma pontuação alternativa para correspondências adequadas entre os sujeitos de uma proposição. O cosseno foi a medida de similaridade utilizada.

2001

20. Summary Street : an intelligent tutoring system writing through the use of for improving student latent semantic analysis

Local University of California

Autor David J. Steinhart

Resumo O texto é uma Tese de Doutorado que descreve o projeto, evolução e testes para implementação do Summary Street(SS). O SS é um sistema de tutoria inteligente que usa LSA para apoiar atividades de escrita e revisão de estudantes. O trabalho é feito em três domínios: civilizações antigas, sistema de circulação humano e fontes de energia. Cinquenta e dois estudantes escreveram resumos sobre os três domínios. Professores atribuirão uma pontuação de 0 a 10. Em ambos os casos, os alunos que utilizaram o SS obtiveram resultados significativos. O cosseno foi a medida de similaridades utilizada.

21. Development of Physics Text Corpora for Latent Semantic Analysis

Local University of Memphis

Autores Donald R. Franceschetti et al .

Resumo O trabalho analisa respostas de alunos para questões de Física usando LSA. O corpus foi constituído por diferentes domínios da Física; os melhores resultados foram obtidos para questões conceituais de Física Fundamental. As respostas foram classificadas em categorias diferentes tais como exposição, problemas de exemplo, material histórico-cultural, raciocínio, e assim por diante. Cada parágrafo foi considerado como um documento. A dimensão do espaço semântico variou de 100 até 500. O cosseno foi utilizado como medida de similaridade.

22. Weight functions impact on LSA performance

Local Sofia University “St. Kliment Ohridski”

Autores Preslav Nakov and Antonia Popova and Plamen Mateev

Resumo O trabalho apresenta resultados experimentais do uso de LSA para análise de textos da literatura inglesa utilizando vários tipos de funções peso. As funções de peso local utilizadas foram termo-frequência e logaritmo; enquanto que as funções de peso global utilizadas foram a trivial, a normal, a relação entre a frequência global de um termo e o número de textos em que o termo aparece e a frequência documento inverso. Os experimentos foram realizados em duas coleções de textos da literatura inglesa coletados diretamente da web: 272 para The Adventures of Sherlock Holmes e 269 para Huckleberry Finn. Foi escolhida dimensionalidade igual a 15. A aplicação das diversas funções peso foi crucial para escolha da dimensionalidade. O cosseno foi a medida de similaridade utilizada.

2002

23. A Computational Theory of Complex Problem Solving Using Latent Semantic Analysis

Local University of Colorado, Boulder, University of Granada, Spain

Autores José Quesada, Walter Kintsch Emilio Gomez

Resumo Problemas complexos (CSP) são considerados problemas híbridos. Este trabalho introduz uma nova e abstrata conceituação de pesquisa baseada em inovações: 1) problemas que tratam protocolos como objetos em um espaço característico e, 2) uma métrica similar que é definida neste espaço problema. LSA é usado para analisar performance em CPS. Exemplos básicos de aplicações são fornecidos, e vantagens e limitações são discutidos. O corpus foi constituído por 3441 ensaios. O cosseno foi a medida de similaridade utilizada.

24. Using Production to Assess Learning: An ILE That Fosters Self-Regulated Learning

Local Universit Pierre-Mendes, France

Autores Philippe Dessus and Benoit Lemaire

Resumo O trabalho apresenta um detalhamento do sistema APEX, um sistema de dois laços que fornece textos para serem lidos por estudantes. No primeiro laço, chamado leitura, o estudante formula uma pergunta e diz quais dos textos fornecidos estão relacionados com a pergunta. Então o sistema faz a sumarização dos textos. No segundo laço, chamado escrevendo, o estudante escreve resumos dos textos fornecidos e recebe uma avaliação do sistema. O sistema não usa nenhum esquema de ponderação, sugere um intervalo para a dimensionalidade de 100 a 300 e usa o cosseno como medida de similaridade.

25. On the computational basis of learning and cognition: Arguments from LSA

Local Universit Colorado

Autor Thomas K Landauer

Resumo O trabalho é um capítulo do livro In The Psychology of Learning and Motivation, Volume 41. Neste capítulo o autor argumenta que LSA tem performance surpreendente na compreensão verbal humana.

2003

26. LSA: First dimension and dimensional weighting

Local Universit Memphis

Autor X Xu and Z Cai and D Francesche and Penumatsa and P. Graesser and M M Louwerse and D S McNamara

Resumo O trabalho reporta duas descobertas acerca de LSA: i) observa as propriedades especiais da dimensão k do espaço semântico; ii) observa que a ponderação desempenha um importante papel em LSA. Foram testados diversos valores para a dimensão k e testados diferentes esquemas de ponderação. Faz a recomendação de um novo algoritmo para LSA. Os experimentos foram feitos com o Auto Tutor. O sistema verificou a similaridade entre os verbos de uma proposição escrita por um estudante e uma resposta esperada e testa uma nova estratégia para pontuação entre duas proposições.

27. Automatic Evaluation of Students' Answers using Syntactically Enhanced LSA

Local Indian Institute of Technology, New Delhi

Autores Dharmendra Kanejiya and Arun Kumar and Surendra Prasad

Resumo O trabalho relata que LSA tem sido utilizado para avaliar aprendizagem de alunos. Apresenta o SELSA (LSA sintaticamente melhorada) que é uma generalização de LSA. O SELSA considera a palavra e sua vizinhança sintática dada pela etiqueta da palavra anterior. Realiza experimentos de avaliação automática de questões de ciência da computação (serviços básicos). O trabalho conclui que o sistema SELSA avalia melhor algumas respostas comparado com LSA, entretanto tem menor correlação em relação a avaliadores humanos em relação a LSA. O corpus foi constituído por dois livros textos de computação e de 10 artigos sobre hardware, sistema operacional e internet. Foram 192 respostas para oito questões de cada um desses tópicos; as respostas foram comparadas com 20 respostas de referência. A dimensionalidade variou de 200 até 400, o esquema de ponderação foi logaritmo e o cosseno foi a medida de similaridade.

28. Towards Deeper Understanding of the Latent Semantic Analysis Performance

Local University of California at Berkeley, Bulgarian Academy of Sciences

Autores Preslav Nakov and Elena Valchanova and Galia Angelova

Resumo O trabalho estuda os fatores que influenciam a performance de LSA. Diferente de pesquisas realizadas anteriormente que estudam parâmetros tais como ponderação, dimensionalidade, medidas de similaridade, etc, este trabalho verifica o impacto de um outro fator fundamental: a definição da palavra. Experimentos foram realizados para comparar dois corpus através da categorização dos textos. Os resultados mostraram que fatores como ponderação, dimensionalidade e medidas de similaridade são mais importantes, entretanto o processamento linguístico também influencia no desempenho do sistema. O corpus foi constituído por 702 artigos de notícias on-line búlgaros divididos em 15 categorias. Foram utilizados dois esquemas de ponderações locais (termo-frequência e logaritmo) e cinco globais (trivial, normal, o quociente da frequência do termo pelo número de documentos que contém este termo, frequência documento inverso e entropia). A dimensionalidade foi igual a 300 e o cosseno foi a medida de similaridade utilizada.

2004

29. Comparison of Human and Latent Semantic Analysis (LSA) Judgements of Pairwise Document Similarities for a News Corpus

Local Intelligence, Surveillance and Reconnaissance Division Information Sciences Laboratory, Australian

Autores Brandon Pincombe

Resumo O trabalho faz análise de similaridades entre humanos e LSA para um conjunto de 50 textos de notícias. O sistema usa dois esquemas de ponderações locais, termo-frequência e logaritmo, e três globais, normal, frequência documento inverso e entropia. O cosseno foi a medida de similaridade utilizada.

30. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation

Local West Bohemia University

Autores Josef Steinberger and Karel Jezek

Resumo O trabalho trata do uso de LSA na sumarização de textos. LSA identifica sentenças semanticamente importantes e faz a comparação entre o texto original e o resumo. Compara sete resumos clássicos com os respectivos resumos por LSA. Usa o cosseno como medida de similaridade.

2005

31. About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment

Local Universidad Autónoma de Madrid (Espanha), Instituto para la Investigación Científica y Tecnológica (Itália)

Autores Diana Pérez and Enrique Alfonseca and Pilar Rodríguez and Alfio Gliozzo and Carlo Strapparava and Bernardo Magnin

Resumo O trabalho apresenta uma combinação de LSA com as técnicas de PLN: stemming, remoção de stop words e desambiguação do sentido da palavra. O principal objetivo é melhorar a avaliação automática de respostas de estudantes em textos livres. Esta combinação é testada no sistema de avaliação assistida (SAA) Atenea. O sistema Atenea faz perguntas aleatoriamente ou de acordo com o perfil do estudante e atribui uma pontuação. Os resultados mostram uma relativa melhora na correlação do sistema Atenea com avaliadores humanos. Não usa nenhum esquema de ponderação e usa o cosseno e a correlação de Person como medidas de similaridades.

32. Automatic Assessment of Students free-text Answers underpinned by the Combination of a BLEU-inspired algorithm and Latent Semantic Analysis

Local Universidad Autónoma de Madrid (Espanha), Instituto para la Investigación Científica y Tecnológica (Itália)

Autores Diana Pérez and Enrique Alfonseca and Pilar Rodríguez and Alfio Gliozzo and Carlo Strapparava and Bernardo Magnin

Resumo O trabalho apresenta uma avaliação comparativa entre o algoritmo BLEU e LSA e propõe que as respostas dos alunos sejam com base nesta combinação. O corpus foi constituído de 1929 respostas de estudantes coletadas em um curso de sistemas operacionais e de 142580 textos da Ziff- Davis que é parte da North America Collection. O sistema usa a correlação de Pearson como medida de similaridade.

33. Factors Influencing Effectiveness in Automated Essay Scoring with LSA

Local Vienna University of Economics and Business Administration

Autores Fridolin Wild and Christina Stahl and Gerald Stermsek and Yoseba Penya and Gustaf Neumann

Resumo O trabalho aborda a discussão sobre fatores que influenciam a pontuação automática de ensaios por LSA. Apresenta evidências para os efeitos dos parâmetros

pré-processamento, ponderação, dimensionalidade, e tipo de semelhança sobre os resultados. Esta eficiência é referida comparando as pontuações atribuídas pela máquinas e por humanos para um caso do mundo real. O trabalho mostra como cada fator influencia a qualidade da pontuação automática do ensaio, entretanto, não são independentes entre si. O corpus foi constituído por respostas de estudantes sobre marketing consistindo de 43 arquivos pré-avaliados por especialistas humanos e média de 56.4 palavras. A referência foi um glossário de 302 arquivos com média de 56.1 palavras. Foi utilizada uma lista de stop words com 373 termos em alemão. Foram combinadas três ponderações locais (termo-frequência, logaritmo e binária) e três ponderações globais (normal, frequência documento inverso, entropia). A dimensionalidade foi igual a 10 e foram utilizadas três medidas de similaridade: correlação de Pearson, correlação de Spearman e o cosseno.

2006

34. Applying part-of-speech enhanced lsa to automatic essay grading

Local University of Joensuu, Finland

Autores Tuomo Kakkonen and Niko Myller and Erkki Sutinen

Resumo O trabalho afirma que a sintaxe tem papel importante na representação do significado de sentenças. Assim é totalmente praticável a extensão de LSA com sintaxe para capturar informações contextuais. Esta abordagem é utilizada empiricamente para classificar ensaios. Os resultados mostraram que a adição da sintaxe pode aumentar a acurácia do sistema em torno de 10%. O sistema foi aplicado em 73 ensaios de educação (pós-graduação), 45 ensaios vocacionais e 27 de engenharia de software (graduação). Não foi utilizado nenhum esquema de ponderação e a correlação de Spearman foi a medida de similaridade utilizada.

35. Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore

Local Universite Catholique de Louvain

Autor Yves Bestgen

Resumo O trabalho propõe a aplicação de LSA na extração do conhecimento semântico de um corpus. O principal objetivo é melhorar a acurácia de um algoritmo de análise morfológica. O algoritmo foi executado várias vezes para verificar se o conhecimento semântico era levado em conta. O corpus foi constituído por textos genéricos. Os resultados mostraram que o conhecimento semântico influencia na acurácia do sistema.

2007

36. Essential Dimensions of Latent Semantic Indexing (LSI)

Local Ursinus College, Collegeville

Autor April Kontostathis

Resumo O trabalho propõe o desenvolvimento de um modelo LSA que tem como principal parâmetro a dimensão do espaço semântico. O corpus foi constituído das seguintes coleções e respectivos números de documentos: MED (1033), CISI (1450), CACM (3204), CRAN (1400), LISA (6004), NPL (11429) e OHSUMED (348566). Para estas coleções a dimensionalidade ficou em torno de 300. O esquema de ponderação utilizado foi log-entropia e a correlação de Spearman foi a medida de similaridade utilizada.

2008

37. Using Latent Semantic Analysis for Extractive Summarization

Local University of Colorado Boulder

Autor Kirill Kireyev

Resumo O trabalho trata do uso de técnicas LSA para fornecer uma maneira simples e robusta de gerar e atualizar a extração de resumos. O corpus foi constituído de 439.947 de artigos do New York Times. A dimensionalidade foi de 327. O cosseno foi a medida de similaridade utilizada.

2009

38. Automatic summary assessment for intelligent tutoring systems

Local The Open University, UK, Nanyang Technological University, Nanyang Avenue, Singapore, Hochiminh City University of Technology Hochiminh City, Vietnam

Autores Yulan He and Siu Cheung Hui and Tho Thanh Quan

Resumo O trabalho diz que a avaliação de resumos escritos é uma tarefa que consome muito tempo do professor e que a avaliação assistida por computador pode ajuda-lo a realizar uma classificação de forma mais eficaz. O trabalho afirma que LSA e n-gramas não têm atingido resultados satisfatórios separadamente e propõe uma abordagem que une LSA com n-gramas para avaliação automática de resumos escritos. Os resultados mostram que abordagem melhora os resultados. Um software com esta

abordagem foi desenvolvido. Seis testes foram utilizados para avaliar a performance da abordagem proposta. Os sumários referência foram obtidos da Cambridge O-Level English Language Examination. O cosseno foi a medida de similaridade utilizada.

39. Effect of Tuned Parameters on a LSA Multiple Choice Questions Answering Model

Local Université Pierre et Marie, Cognition Humaine et Artificielle, EPHE-CNRS, Paris, France

Autores Alain Lifchtz and Sandra Jehan Larose and Guy Denhiere

Resumo O trabalho propõe o estudo de fatores que influenciam significativamente o desempenho de LSA. A dimensão do espaço semântico é o parâmetro principal. Um software foi projetado para calibrar a dimensão do espaço semântico para questões de múltipla escolha. O corpus foi constituído por 1049 testes de múltipla escolha de Biologia dos ensinos público e particular. O esquema de ponderação foi log-entropia, a dimensionalidade variou de 5 a 14 e o cosseno foi a medida de similaridade utilizada.

40. New algorithms assessing short summaries in expository texts using latent semantic analysis

Local Universidad Autónoma de Madrid, Madrid, Spain

Autores Ricardo Olmos and José A León and Guillermo Jorge-Botana and Inmaculada Escudero

Resumo Neste trabalho foram comparados resumos de um texto expositivo com quatro resumos de referência. O trabalho propõe três novas abordagens: 1) algoritmo LSA tradicional, 2) melhor redução para a dimensão do espaço semântico, 3) distância euclidiana como medida de similaridade, pois incorpora ao mesmo tempo o comprimento do vetor e o cosseno do ângulo. Ao todo participaram 192 estudantes do ensino médio e seis especialistas humanos. Todos leram e resumiram um texto expositivo. Os Resultados mostraram que LSA é uma técnica de avaliação confiável.

2010

41. Applying Latent Semantic Indexing on the TREC 2010 Legal Dataset

Local Ursinus College, Collegeville

Autores Andy Garron and April Kontostathis

Resumo O trabalho aplica LSI e EDLSI numa tarefa de aprendizagem para a TREC 2010. O foco é a aprendizagem de máquina para o processo LSI. Mesmo sendo a

dimensionalidade limitada em 70, a performance do sistema foi maior ou igual a métrica F1.

42. Latent Semantic Analysis Parameters for Essay Evaluation using Small-Scale Corpora

Local Universidad Autonoma de Madrid, Spain

Autores Guillermo Jorge-Botana and José A Leon and Ricardo Olmos and Inmaculada Escudero

Resumo O trabalho afirma que LSA tem alcançado resultados positivos numa comparação com avaliadores humanos. No entanto, existem diferenças consideráveis na metodologia de modelos LSA. Estas diferenças impedem a identificação de melhores parâmetros: remoção de stop words, funções peso e dimensionalidade. O trabalho foca na busca da melhor combinação de parâmetros para um modelo LSA para avaliação automática de ensaios numa comparação com especialistas humanos. O corpus foi constituído por 80 respostas de estudantes para a seguinte questão aberta: O que é uma fobia social ?. O modelo utilizou termo-frequência, logaritmo, frequência documento inverso e entropia como funções de ponderação, foram testados vários valores para a dimensionalidade, e o cosseno e a distância euclidiana foram as medidas de similaridades utilizadas.

43. Diffusion of latent semantic analysis as a research tool: A social network analysis approach

Local Hacettepe Universit, Ankara, Turkey

Autores Yasar Tonta and Hamid R Darvish

Resumo O trabalho afirma que LSA é uma técnica relativamente nova e tem aplicações desde a análise do discurso até ciências cognitivas. Este trabalho traça o desenvolvimento e a difusão de LSA como ferramenta de pesquisa identificando trabalhos que usam LSA em seus títulos, 65 trabalhos, e artigos que tem LSA como um de seus temas, 250 artigos. O período considerado foi de 1990 até 2008.

44. Latent Semantic Analysis: Five Methodological Recommendations

Local University of North Texas, Northern Kentucky University

Autores Nicholas Evangelopoulos and Xiaoni Zhang and Victor R. Prybutok

Resumo O trabalho afirma que análise de dados textuais ainda é um desafio para pesquisadores e que LSA é uma abordagem que pode ser utilizada para enfrentar este

desafio. O trabalho relata cinco metodologias que devem orientar um pesquisador de LSA. Discussões sobre como estas decisões metodológicas devem ser tomadas são realizadas. O trabalho ilustra estas questões com quatro estudos envolvendo a análise de resumos de artigos publicados no European Journal of Information Systems.

2011

45. Sparse Latent Semantic Analysis

Local Carnegie Mellon University

Autores Xi Chen and Yanjun Qi and Bing Bai and Qihang Lin and Jaime G Carbonell

Resumo O trabalho propõe um novo modelo chamado Sparse LSA que produz uma matriz projeção esparsa via regularização. Diferente do modelo LSA tradicional, o modelo Sparse LSA seleciona apenas um pequeno número de palavras relevantes em cada tópico, e assim, fornece uma representação compacta das relações entre as palavras. O modelo Sparse LSA requer menos memória para o armazenamento da matriz. Foram realizados vários experimentos em diferentes conjuntos de dados para comparação entre Sparse LSA e outros métodos. Os resultados mostraram que Sparse LSA tem melhor desempenho que LSA e explica melhor as relações entre tópicos e palavras.

46. Automatic Estimation of the Dimension

Local University of Lisbon

Autores Jorge Fernandes and Andreia Artıfice and Manuel J Fonseca

Resumo O trabalho afirma que métodos baseados em LSA exigem que o usuário defina a dimensão do espaço semântico utilizado e propõe uma fórmula para estimar esta dimensão com base no número de textos do corpus. Resultados mostram que a fórmula alcança resultados semelhantes aos resultados obtidos testando-se a melhor dimensão para o espaço semântico. O corpus foi composto por dez categorias de textos: ciência e tecnologia, cinema e tv, esportes, economia e administração, informática e internet, games, música, política, saúde e veículos automotivos. Foram utilizados 128 textos por categoria. A dimensionalidade foi de 107. Foi utilizada a medida F1 como medida de similaridade.

47. Automated Assessment of Short Free-Text Responses in Computer Science using Latent Semantic Analysis

Local University of the Witwatersrand Johannesburg, South Africa, University of California Merced, USA

Autores Richard Klein and Angelo Kyrilov and Mayya Tokman

Resumo O trabalho relata que questões de múltipla escolha fornecem uma visão limitada do conhecimento do estudante e diz que professores preferem o usar questões abertas para avaliar o conhecimento do estudante. Estas questões são difíceis de pontuar, pois existe uma grande variação no significado semântico das palavras. O trabalho afirma que modelos LSA podem inferir o significado semântico de um texto. O trabalho descreve a concepção, implementação e avaliação de um sistema de avaliação automática, com base em LSA, para pontuação de questões abertas. O corpus utilizado foi composto por respostas a questões de um curso de informática já avaliadas por um especialista humanos. O esquema de ponderação utilizado foi termo-frequência e logaritmo, a dimensionalidade variou no intervalo de 1 até 7, e o cosseno e a distância euclidiana foram as medidas de similaridade utilizadas. O sistema atinge uma correlação acima de 0.8 em comparação com a pontuação de especialista humanos.

48. Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents

Local Katholieke Universiteit Leuven

Autores Tom Magerman and Bart Van Looy and Bart Baesens and Koenraad Debackere

Resumo O trabalho realiza uma avaliação completa do método LSA e seus parâmetros. Verifica semelhanças entre 88.248 documentos patenteados e 948.432 publicações científicas com base em 40 medidas de similaridade, quatro esquemas de ponderação e dimensionalidade igual a 10. Os melhores resultados foram obtidos com termo-frequência e o cosseno.

49. Using latent semantic analysis to grade brief summaries: some proposals

Local Ricardo Olmos and José A León and Inmaculada Escudero and Guillermo Jorge-Botana

Autores Universidad Autonoma de Madrid

Resumo O trabalho apresenta propostas na melhora de modelos LSA para avaliação de pequenos resumos narrativos e expositivos (menos de 50 palavras). Foram analisados ensaios já existentes e propostos novos algoritmos que simulam o comportamento humano para melhorar a confiabilidade LSA. Um método de avaliação que combina LSA com o modelo linguístico Rouge-N foi apresentado com o objetivo de melhorar a qualidade das avaliações em diferentes níveis acadêmicos. O corpus foi constituído por 372 documentos retirados da internet e enciclopédias on line e livros textos. No total foram 5.995 palavras lematizadas. A dimensionalidade foi de 75. A distância euclidiana foi a medida de similaridade utilizada.

50. Automated assessment of review quality using latent semantic analysis

Local North Carolina State University Raleigh,

Autores Lakshmi Ramachandran and Edward F Gehringer

Resumo O trabalho foca na identificação dos fatores que identificam a qualidade de uma revisão: a qualidade e o tom dos comentários, e o número de anotações. Foi utilizado um sistema informatizado baseado em LSA para classificar comentários pela sua qualidade e tom. Os dados usados para revisão e re-revisão passaram por diferentes etapas de pré-processamento. Os dados para os experimentos foram retirados da Expertiza, um ambiente de aprendizagem colaborativo baseado na web. O cosseno foi a medida de similaridade utilizada.

51. Using Latent Semantic Analysis for Automated Grading Programming Assignments

Local Universiti Malaysia Sarawak

Autores Kartinah Zen and D N F Awang Iskandar and Ongkir Linang

Resumo O trabalho relata que programas de computadores são classificados manualmente. Como a classificação manual é tediosa e demorada o trabalho propõe uma ferramenta para classificação automática de computadores. Foi feita uma adaptação de LSA para classificação de programas em computadores como ao processo de classificação manual. As pontuações são geradas pelo cosseno de similaridade e os resultados mostram que LSA não é capaz de detectar as ordens dos programas e nem símbolos; no entanto, detecta as atribuições mais rapidamente e consistentemente, o que resulta na diminuição do tempo do processo de classificação manual.

2012

52. Automatically Assessing Free Texts

Local University of Lethbridge, Lethbridge, AB, Canada

Autores Yllias Chali and Sadid A Hasan

Resumo O trabalho revela que avaliação de conteúdos de textos livres é uma tarefa difícil para humanos. O trabalho tem foco na avaliação automática de ensaios escritos por estudantes. Usa um método baseado em LSA e nos experimentos um conjunto de dados obtidos a partir de um curso de terapia ocupacional. São 91 respostas escritas por estudantes pré-avaliadas pelo professor. O sistema usou os seguintes esquemas de ponderação: termo-frequência, logaritmo, entropia, normal e frequência documento inverso. O cosseno foi a medida de similaridade utilizada.

53. With LSA Size DOES Matter

Local University of Malta

Autor Colin Layfield

Resumo O trabalho diz que LSA é parte do campo de processamento de linguagem natural que permite comparações semânticas entre textos por operações vetoriais. Modelos LSA tem sido usados para avaliação automática de ensaios. Uma variável é a quantidade de textos. Existe um consenso que quanto mais textos quanto melhor o desempenho de um modelo LSA; entretanto existem poucos exemplos concretos que demonstram este fato. Este trabalho mostra a partir de dados do mundo real que quanto mais texto melhores serão as comparações de similaridades semânticas entre esses textos. O estudo utilizou dois conjuntos de respostas: um com 88 respostas e outro com 137 respostas para uma mesma pergunta sobre avaliação da capacidade funcional de uma disciplina do último ano de um curso de Terapia Ocupacional. Utilizou o esquema de ponderação termo-frequência vs frequência Documento inverso. O cosseno foi a medida de similaridade utilizada.

54. Automated Essay Scoring Using Generalized Latent Semantic Analysis

Local Bangladesh University of Engineering & Technology

Autores Monjurul Islam and A S M Latiful Hoque

Resumo O trabalho mostra o desenvolvimento de um sistema usando LSA generalizada (GLSA). O sistema GLSA considera n-gramas na matriz inicial. Resultados mostram que o sistema GLSA proposto atinge um alto nível de acurácia comparado com

avaliadores humanos. O corpus foi constituído por 960 ensaios escritos por estudantes de pós graduação sobre divisão digital, plantação e governança. Utilizou o esquema de ponderação termo-frequência vs frequência Documento inverso. O cosseno foi a medida de similaridade utilizada.

55. Automated Assessment of Students' Arabic Free-text Answers

Local Mansoura University, Egypt

Autores M M Refaat and A A Ewees and M M Eisa and Ab. A. Sallam

Resumo O trabalho diz que questões discursivas são importantes na avaliação e que necessitam de um longo tempo para serem avaliadas. Muitas pesquisas foram feitas para avaliação automática de respostas a questões discursivas. O artigo propõe um método baseado em LSA para avaliar automaticamente respostas de textos livres na língua árabe. A correlação do método com avaliadores humanos foi satisfatória. Foram coletadas 29 respostas de estudantes sobre sistemas de designer escritas em língua árabe. A média foi de 75 palavras por resposta. Utilizou o esquema de ponderação termo-frequência vs frequência Documento inverso. O cosseno foi a medida de similaridade utilizada.

2013

56. Exploring the Assessment of Summaries: Using Latent Semantic Analysis to Grade Summaries Written by Spanish Students

Local Universidad Automona de Madrid

Autores J A León and R Olmos and I Escudero and G Jorge-Botana and D Perry

Resumo O trabalho propõe um método integrado para avaliar resumos automaticamente usando LSA. O método é baseado em uma equação de regressão que incorpora dois parâmetros LSA: semelhança semântica e comprimento do vetor. O estudo foi realizado com 396 estudantes de quatro etapas da educação. Os resumos de pequenos textos narrativos foram avaliados em uma escala de 0 a 10 por quatro especialistas humanos. Estas pontuações foram comparadas com as pontuações atribuídas pelo modelo LSA. Os resultados mostraram que a incorporação dos dois parâmetros foi mais eficiente do que a medida do cosseno tradicional.

2015

57. Automated LSA Assessment os summaries in Distance Education: Some Variables to Be Considered

Local Universidad Automona de Madrid

Autores J A León and R Olmos and I Escudero and G Jorge-Botana and D Perry

Resumo O trabalho faz a discussão de algumas variáveis no processo de avaliação automática de resumos no ensino à distância. O sistema avalia automaticamente o conteúdo, a coerência textual e o peso de palavras para estimar uma pontuação de resumos. Os primeiros resultados mostraram uma interdependência entre a coerência de parágrafos e variáveis superficiais de texto. Um modelo de regressão foi utilizado para aproximar as pontuações do modelo LSA com pontuações atribuídas por especialistas humanos.