Apoio a mediação pedagógica em um "Debate de Teses" utilizando técnicas de processamento de texto

Sabrina Siqueira Panceri[1], **Crediné Silva de Menezes**[1 2]

[1] Programa de Pós Graduação em Informática – Universidade Federal do Espírito Santo
Av. Fernando Ferrari, 514, Goiabeiras – Vitória – ES – Brasil – Caixa Postal 29075-910

[2]Faculdade de Educação – Universidade Federal do Rio Grande do Sul
Av. Paulo Gama, s/n, Farroupilha – Porto Alegre – RS - Brasil – Caixa Postal 90046-900

{sabrinapanceri, credine}@gmail.com

Abstract. *Aiming at the mediation, the teacher accompanying textual productions during the application of Pedagogical Architecture Debate Theses. Such monitoring requires a considerable amount of teacher time can be minimized when used computational tools to support this activity. This paper describes the construction of the SMA Alpes word processing core, through a combination of information retrieval techniques, text mining, latent semantic analysis and clustering. It is proposed a junction between the latent semantic analysis and the clustering method adopted. Applies the SMA Alpes to a thesis debate in 2013 and the results achieved demonstrate the feasibility of the proposal.*

Resumo. *Visando a mediação pedagógica, o professor acompanha as produções textuais durante a aplicação da Arquitetura Pedagógica Debate de Teses. Esse acompanhamento demanda uma quantidade de tempo considerável do professor, podendo ser minimizado ao serem utilizadas ferramentas computacionais como suporte a esta atividade. Este trabalho descreve a construção do núcleo de processamento de textos do SMA Alpes, através da combinação de técnicas de recuperação de informações, mineração de textos, análise semântica latente e clusterização. É proposta uma junção entre a análise semântica latente e o método de clusterização adotado. Aplica-se o SMA Alpes a um debate de teses realizado em 2013. Os resultados alcançados demonstram a viabilidade da proposta.*

1. Introdução

As interações nos espaços virtuais, na maior parte dos casos, são realizadas através da produção de pequenos textos. Comprova-se isso quando faz-se uso de ferramentas de comunicação instantânea *(chats)*, ou realizam-se postagens em uma rede social, ou até mesmo quando ocorrem interações em Ambientes Virtuais de Ensino e Aprendizagem (AVEA). Nos AVEA, as produções textuais são estimuladas por ferramentas como fóruns, *wikis*, ou pela aplicação de uma Arquitetura Pedagógica [Carvalho 2005]. A partir da concepção das Arquiteturas Pedagógicas, [Nevado 2011] propõem a Arquitetura Pedagógica Debate de Teses (APDT), que visa a construção de conhecimento coletivo através de interações orientadas, que são estabelecidas através de produções textuais.

Os textos produzidos com a aplicação da APDT são analisados pelo professor, para que sejam realizadas intervenções pedagógicas a fim de contribuir com a construção de conhecimento de seus alunos. De acordo com [Mason and Grove-Stephensen 2002], o docente gasta cerca de 30% de seu tempo na correção de atividades, tempo este que poderia ser

DOI: 10.5753/cbie.sbie.2015.977

utilizado em outras tarefas, como planejamento e/ou pesquisa. Com o objetivo de auxiliar o professor, foi concebido o Sistema Multiagente Alpes (SMA Alpes), apresentado nos trabalhos [Panceri and Menezes 2014a] e [Panceri and Menezes 2014b]. O presente artigo apresenta a continuidade desta pesquisa.

À vista disto, fez-se um estudo sobre Recuperação de Informações (RI) que viabilizou a concepção do SMA Alpes. O núcleo de processamento do SMA Alpes foi desenvolvido com base em técnicas das áreas de Recuperação de Informações (RI), Mineração de Textos (MT), Análise Semântica Latente (LSA) e Clusterização. O núcleo é responsável pela análise e processamento das produções textuais e alimentação das ferramentas que fornecem apoio aos professores. Duas abordagens foram desenvolvidas para criação dos grupos de similaridade com o aplicação do algoritmo de clusterização *K-Means*. Para validar o protótipo, este é aplicado sobre um debate de teses real, ocorrido em 2013. Valida-se os resultados alcançados através do cálculo da similaridade média entre os membros dos grupos criados. O apoio as mediações pedagógicas a partir dos agrupamentos são destacados. Além disso, a abordagem dada ao tratamento do debate de teses levou em conta o levantamento de trabalhos correlatos no contexto da mediação em espaços virtuais de debate.

Este artigo está organizado da seguinte forma: a próxima seção os trabalhos correlatos a esta pesquisa. Na seção 3 tem-se uma breve fundamentação teórica sobre as áreas e técnicas utilizadas para o desenvolvimento deste trabalho. A seção 4 contém a descrição do contexto do problema. A seção 5 apresenta o processo de construção do núcleo de processamento textual do SMA Alpes. Na seção 6 mostra-se a aplicação do Alpes a um debate realizado em 2013, os resultados obtidos e a análise dos dados, além da sugestão das mediações pedagógicas. E, as considerações finais são apresentadas na seção 7.

2. Trabalhos Correlatos

[Azevedo et al. 2011] apresentam a aplicação do software MineraFórum, que realiza uma análise qualitativa das mensagens enviadas por alunos em fóruns de discussão. O MineraFórum utiliza técnicas de mineração de textos, e tem como foco principal analisar a relevância das mensagens enviadas pelo aluno em relação ao tema em discussão.

Em [Kim and Shaw 2014] apresenta-se um estudo de aplicação do PedaBot, que é uma ferramenta para mineração dos textos escritos em fóruns do estilo pergunta-resposta. O sistema foi desenvolvido com a combinação de técnicas de LSA, TF-IDF, técnicas de processamento de linguagem natural e recuperação de informações, e tem por objetivo fazer sugestões de textos e/ou outros tópicos sobre o assunto pesquisado pelo aluno. Os autores consideram que as indicações realizadas pelo sistema contribuem para construção de conhecimento de seus usuários.

Em [Avila and Soares 2013] o foco é a análise de respostas discursivas. Os autores apresentam uma comparação entre algoritmos utilizados para calcular a similaridade entre palavras, com as devidas alterações para que o cálculo seja realizado sobre frases. Além disso, a similaridade das respostas é calcula com base em respostas padrões previamente cadastradas no sistema.

[Kakkonen and Sutinen 2004] utiliza uma abordagem baseada em LSA para realizar a correção de respostas discursivas. O diferencial de sua proposta está na utilização de uma pequena base de textos para realizar o treinamento do sistema. Além disso, os textos utilizados para treinamento são os textos adotados pelo curso como materiais didáticos e/ou

referencial teórico. Dessa forma, a abordagem se diferencia por considerar esse tipo de material como base para realizar a avaliação das respostas.

Já [Klein et al. 2011] apresenta um sistema de correção de respostas discursivas que utiliza uma combinação de técnicas de LSA, TF-IDF, Similaridade de Cossenos e Distância Euclidiana, para realizar correção dos textos com uma taxa de acerto de 80%. O destaque deste trabalho é não utilizar o treinamento da base de análise.

A relação entre os trabalhos de [Azevedo et al. 2011], [Kim and Shaw 2014], [Avila and Soares 2013], [Kakkonen and Sutinen 2004] e [Klein et al. 2011] e a presente proposta está na semelhança de técnicas de pré-processamento textual, técnicas de mineração de textos e recuperação de informações, LSA e técnicas de agrupamento, buscando auxiliar o professor na tarefa de análise de textos produzidos pelos alunos.

O diferencial do presente trabalho encontra-se na combinação do algoritmo de clusterização *K-Means* com as técnicas de LSA. Sendo que, para treinamento da base de análise do módulo de LSA, utilizou-se os textos produzidos pelos alunos no próprio debate e textos utilizados como material didático de referência sobre o assunto abordado pela tese.

3. Fundamentação Teórica

A Recuperação de Informações (RI) tem como foco a busca de informações em documentos (ou textos), e a classificação destas buscas a fim de retornar os documentos mais importantes em uma coleção, dada uma quantidade limitadas de informações a serem resgatadas. De acordo com [Manning et al. 2008], "RI busca encontrar material (geralmente documentos) de natureza não estruturada (geralmente textos) que satisfaçam uma necessidade de informação, a partir de grandes coleções". Derivada da área de RI, a Mineração de Textos (MT) tem por objetivo a extração de informações úteis a partir de interações do usuário com uma coleção de documentos (ou textos) através de um conjunto de ferramentas de análise [Feldman and Sanger 2006].

Em ambas as áreas, RI e MT, técnicas de pré-processamento textual são aplicadas na coleção de documentos *(corpus)* para sua redução e criação da coleção de documentos processados que serão posteriormente analisados. As técnicas exploradas e utilizadas neste trabalho são: Remoção de termos frequentes (remoção de *stopwords*); Capitalização *(case-folding)*; Remoção de caracteres acentuados; e Lematização.

A representação dos documentos utiliza a identificação de características do mesmo através de um modelo espaço vetorial, onde cada documento representado na forma de um vetor apresenta a sequência de características encontradas e seus pesos. O modelo *"bag-of-words"*, um dos mais comuns e utilizados pelos sistemas de RI e MT, utiliza todos os termos do documento como uma característica, descartando a ordem em que eles são apresentados e armazenando apenas a quantidade de ocorrência desses termos nos documentos [Feldman and Sanger 2006].

A Análise Semântica Latente (LSA, do inglês *Latent Semantic Analysis*), apresentada por [Dumais 2004], é uma técnica estatística que visa a construção de matrizes de relações entre os documentos e seus termos, considerando que termos que estão num mesmo documento podem conter determinada relação semântica. Apesar de considerar a co-ocorrência dos termos num documento, LSA utiliza uma abordagem *"bag-of-words"*, logo não leva em consideração a sequência em que os termos aparecem no documento. Para que a aplicação da LSA seja bem sucedida, é necessário que seja feito o treinamento da base

que será utilizada para a análise de similaridade dos textos.

Para agrupar os documentos de uma coleção levando em consideração a semelhança de seus conteúdos é comum a utilização de algoritmos de Clusterização. Neste trabalho o algoritmo utilizado é o *K-Means* (KM). O KM utiliza uma coleção de documentos previamente tratados e representados num modelo espaço vetorial, particiona os vetores de documentos em um conjunto pré-definido de grupos *(clusters)*, cria os agrupamentos através do cálculo da média ao quadrado da Distância Euclidiana entre o vetor que representa o documento e o vetor que representa o centro (ou centróide) do grupo. O número de grupos e de centróides (K) deve ser informado previamente, sendo uma das características principais deste algoritmo. Contudo, a definição deste número deve ser baseada em conhecimentos prévios sobre a coleção que será analisada, tornando esta definição parte importante da eficiência de aplicação do KM. Outra característica importante do KM é o fato de serem realizadas n iterações para a definição dos grupos, onde cada iteração busca maximizar as semelhanças dos membros de um mesmo grupo e distanciá-los de outros grupos, até que não hajam mais modificações nos grupos que foram definidos [Manning et al. 2008, Feldman and Sanger 2006].

Neste trabalho, utilizou-se duas abordagens diferenciadas para criação do modelo espaço vetorial que será analisado pelo KM. A primeira abordagem, comum e retratada em diversos trabalhos, utiliza o algoritmo *TF-IDF (Term frequency x Inverse Document Frequency)*, que atribui a cada termo da coleção de documentos um peso baseado na quantidade de ocorrências do mesmo dentro da coleção. A fórmula para cálculo do peso de cada termo é dada pela função 1:

$$TFIDF(w, d) = TermFreq(w, d).log(N/DocFreq(w)) \quad (1)$$

Onde $TermFreq(w, d)$ é a frequência de cada palavra no documento, N é o número de todos os documentos da coleção, e $DocFreq(w)$ é o número de documentos que contém o termo w. A segunda abordagem, diferencial deste trabalho, utiliza LSA para gerar o modelo espaço vetorial que será analisado pelo KM. Neste caso, o modelo será composto por matrizes que possuem, além da frequência dos termos, a relação de co-ocorrência entre eles nos documentos da coleção.

Como o KM utiliza a Distância Euclidiana para calcular a similaridade entre os documentos e assim fazer os agrupamentos, e segundo [Feldman and Sanger 2006] a Similaridade de Cossenos (SC) *(Cosine Similarity)* é a medida de similaridade mais utilizada para realizar o agrupamento de textos, utiliza-se a SM calcular a similaridade entre os textos que foram agrupados.

4. Contexto do problema

A Arquitetura pedagógica Debate de Teses (APDT) é organizada em fases sequenciais, sendo cinco delas colocadas em prática com o apoio de um ambiente virtual específico[1]. Essas cinco fases são:

1. **Definição da tese:** O professor, no papel de mediador, define o assunto e cadastra a tese a ser discutida entre os alunos.
2. **Posicionamento e Argumentação Inicial:** O aluno, no papel de argumentador, sinaliza seu posicionamento sobre a tese em análise e o fundamenta ao escrever sua argumentação inicial.

[1]O ambiente para aplicação da arquitetura pedagógica Debate de Teses está disponível para acesso e utilização em http://lied.inf.ufes.br/debate2/

3. **Revisão por pares:** O aluno, agora no papel de revisor, tem acesso ao posicionamento e argumentação inicial de outros dois alunos. O objetivo do revisor é contrapor os argumentos de seu revisado, indicando alternativas a sua argumentação, a fim de apresentar um ponto de vista diferenciado sobre o assunto.
4. **Réplica:** O aluno, de volta ao papel de argumentador, responde as revisões realizadas sobre sua argumentação inicial.
5. **Posicionamento e Argumentação final:** O aluno, como argumentador, indica seu posicionamento final e o fundamenta escrevendo a argumentação final.

Os alunos produzem textos nas fases 2, 3, 4 e 5 que devem ser relacionados à tese proposta pelo professor. Para verificar se os alunos concluíram com êxito sua participação no debate e fazer as intervenções pedagógicas necessárias, o professor terá que analisar cada texto produzido. Como exemplo, ao utilizar a APDT numa turma com 10 alunos, propondo 3 teses para debate, o professor mediador terá um total de 180 textos para analisar, pois, para cada tese em debate são produzidos 6 textos.

Esta análise demanda uma quantidade de tempo considerável do professor, que podemos relacionar com os dados apresentados em [Mason and Grove-Stephensen 2002], onde os autores afirmam que, em média, o professor gasta cerca de 30% de seu tempo na correção de atividades, e neste caso, considera-se a aplicação da APDT como uma atividade, logo ela está incluída nessa porcentagem. Contudo, essa quantidade de tempo pode ser minimizada com a aplicação de ferramentas que realizem o processamento dos textos e produzam, de acordo com critérios pré-estabelecidos pelo professor, relatórios que o ajude a visualizar quais alunos precisam de sua ajuda e/ou apoio.

[Panceri and Menezes 2014b] identificam as mediações pedagógicas observadas durante a fase inicial desta pesquisa, e estas são apresentadas de forma resumida a seguir:

- Agrupar participantes por posicionamento e argumentação inicial;
- Fazer comentários sobre a argumentação inicial para orientar o participante;
- Verificar se as revisões feitas pelo participante são diferentes e relevantes;
- Verificar se não houveram comparações entre as argumentações revisadas;
- Conferir se as réplicas respondem aos questionamentos feitos nas revisões;
- Verificar se nas réplicas existem novos argumentos que reforcem o posicionamento inicial;
- Identificar "pontos importantes" para orientar a fase de reflexão[2];
- Elaborar síntese das argumentações finais para realizar *feedback* coletivo;
- Agrupar as argumentações finais semelhantes;
- Verificar qual o grau de influência das revisões e réplicas na argumentação final;
- Identificar quais teses podem ser trabalhadas a partir das argumentações finais;
- Analisar o padrão de resposta de cada participante;
- Verificar a evolução das argumentações;
- Avisar sobre o prazo de envio das etapas;
- Verificar se houve plágio nas revisões, réplicas e argumentações finais;
- Verificar se os textos foram escritos de forma respeitosa.

Em [Panceri and Menezes 2014a] foi proposta a concepção do SMA Alpes, composto pelas soluções (ferramentas): Relatório de Verificação de Réplicas; Relatório de Análise de Influências; Relatório de Evolução da Argumentação; Grupos de Argumentação;

[2]A fase de reflexão é realizada fora do ambiente virtual, através de uma conversa com a turma

Relatório de Padrões de Respostas; Relatório de Plágios; Relatório de Verificação das Revisões; Sugestões de teses.

Consequentemente, este trabalho apresenta a descrição da construção do núcleo de processamento de textos do SMA Alpes, através da combinação de técnicas de recuperação de informações, mineração de textos, análise semântica latente e clusterização. Neste trabalho, considera-se como coleção de documentos os textos produzidos pelos alunos em cada fase da APDT.

5. Núcleo de processamento textual

O núcleo de processamento foi desenvolvido utilizando a linguagem Python[3], na versão 2.7. A coleção de documentos analisados neste trabalho foi escrita no idioma Português, e para auxiliar no desenvolvimento de algumas funcionalidades utilizamos as bibliotecas NLTK[4] e Scikit-Learn[5].

A arquitetura do núcleo de processamento é representada pela Figura 1, na qual primeiro são resgatados os dados do banco de dados Debate de Teses, que são os "Textos". Em seguida, sobre os Textos, são aplicadas as técnicas de pré-processamento textual, realizando primeiro a substituição dos caracteres em maiúsculo para o seu correspondente em minúsculo, depois os caracteres acentuados são substituídos pelo seu correspondente sem acentuação, e por fim as pontuações são removidas. Eventuais *tags* HTML que foram salvas junto com os textos no banco de dados também são removidas. Com o auxílio da função *stopwords()* da NLTK, utiliza-se a lista de termos frequentes cadastrados para a língua portuguesa, comparando cada termo da coleção com as *stopwords* e, ao serem encontradas estas são removidas.

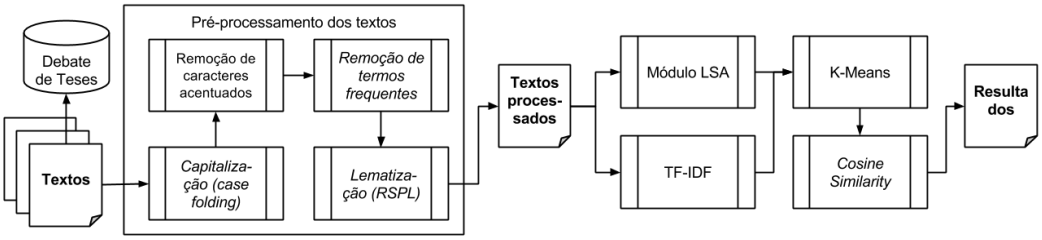

Figure 1. Arquitetura do núcleo de processamento

Como última etapa da fase de pré-processamento dos textos, aplica-se o RSLP - Removedor de Sufixos para Língua Portuguesa [Orengo and Huyck 2001]. Baseado em regras, o RSLP é utilizado para remover os sufixos dos termos e retornar apenas o *lemma* principal de cada termo. Ao final do pré-processamento dos textos, tem-se a coleção de "Textos Processados". Duas abordagens foram implementadas, uma utiliza o algoritmo TF-IDF para transformar a coleção de textos processados em matrizes que contém o peso de cada termo em relação a sua frequência nos documentos, a outra utiliza o módulo de LSA para criação das matrizes de frequência e co-ocorrência dos termos com os documentos. Para implementação do módulo de LSA utilizou-se a biblioteca *gensim*[6]. Dentro do módulo de LSA foram desenvolvidas duas formas de treinamento da base: Forma 1: criação dos dicionários de termos com base em textos sobre o assunto proposto pela tese e utilizados

[3] www.python.org
[4] www.nltk.org
[5] scikit-learn.org/dev/index.html
[6] Disponível em http://radimrehurek.com/gensim/index.html

como material de referência e/ou didático; Forma 2: criação dos dicionários de termos com base nos textos produzidos por todos os alunos em suas argumentações iniciais.

O agrupamento dos documentos semelhantes é feito com a aplicação do algoritmo de clusterização *K-Means*. O KM foi implementado através da utilização da classe *KMeans()* do pacote *cluster* da biblioteca *Scikit-Learn*. Após a definição dos grupos pelo algoritmo *K-Means*, utiliza-se a Similaridade de Cossenos como medida de similaridade para analisar e comparar os textos dos membros de um mesmo grupo, e verificar se eles realmente possuem similaridade entre si. Finalizada esta análise, os "Resultados" são apresentados para o usuário.

6. Aplicação do Alpes a um Debate de Teses

Para validar o núcleo desenvolvido, analisou-se um debate de teses realizado durante a "XIV Maratona de Empreendedorismo da UFRGS", no ano de 2013 [Michels 2014]. Neste debate, 21 dos 74 alunos cadastrados concluíram sua participação, desenvolvendo todas as etapas do debate para as três teses propostas. A análise foi aplicada sobre a argumentação inicial, com o objetivo de agrupar os alunos que apresentavam conhecimentos prévios semelhantes sobre o tema em debate e assim alimentar a ferramenta "Grupos de Argumentação".

Isto posto, foram realizados experimentos de acordo as duas abordagens propostas. Sendo que, a segunda abordagem utiliza as duas formas de treinamento diferenciadas descritas anteriormente. Portanto, temos: **Abordagem A** - TF-IDF + KM; **Abordagem B** - LSA + Forma 1 + KM; e **Abordagem C** - LSA + Forma 2 + KM. Vale ressaltar que com a utilização de LSA, o sistema está apto a realizar a análise de textos produzidos sobre qualquer área do conhecimento, pois, os dicionários de semelhança entre os termos são gerados a partir da alimentação do sistema, ficando a critério do professor escolher qual forma de treinamento irá utilizar. O resumo dos resultados obtidos através dos experimentos realizados com a aplicação das abordagens A, B e C para análise das argumentações iniciais elaboradas sobre as três teses debatidas é apresentado pela Tabela 1.

Table 1. Agrupamentos realizados sobre o Posicionamento Inicial

TESE 1

Abordagem A							*Abordagem B*							*Abordagem C*						
Grupos	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	*Grupos*	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	*Grupos*	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**
Qtde	2	3	3	4	6	3	*Qtde*	2	3	4	3	5	4	*Qtde*	3	2	3	4	4	3
SIM	67,8%	75,4%	55,0%	80,4%	57,1%	62,0%	*SIM*	72,5%	38,0%	47,1%	60,4%	65,5%	74,9%	*SIM*	68,3%	84,1%	69,4%	48,46	47,1%	60,4%
Qtde	3	3	3	8	4		*Qtde*	3	4	5	6	3		*Qtde*	6	4	3	5	3	
SIM	55,0%	62,0%	75,4%	52,3%	80,4%		*SIM*	60,4%	47,1%	65,5%	72,5%	38,0%		*SIM*	57,2%	48,5%	68,3%	69,4%	60,4%	
Qtde	9	5	3	4			*Qtde*	8	3	7	3			*Qtde*	7	3	5	6		
SIM	49,0%	78,4%	62,0%	68,1%			*SIM*	69,5%	60,4%	63,0%	38,0%			*SIM*	49,4%	60,4%	69,4%	57,2%		
Qtde	6	9	6				*Qtde*	8	5	8				*Qtde*	9	8	4			
SIM	65,5%	51,1%	75,3%				*SIM*	55,6%	62,3%	69,5%				*SIM*	69,6%	61,4%	48,5%			

TESE 2

Abordagem A							*Abordagem B*							*Abordagem C*						
Grupos	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	*Grupos*	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	*Grupos*	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**
Qtde	3	4	2	4	4	4	*Qtde*	2	5	2	4	6	2	*Qtde*	5	2	7	3	3	1
SIM	92,0%	84,3%	86,0%	58,0%	33,5%	82,5%	*SIM*	93,9%	87,7%	94,6%	82,7%	83,8%	0,0%	*SIM*	80,5%	86,4%	80,2%	48,0%	79,4%	-
Qtde	3	7	3	3	5		*Qtde*	3	8	2	4	4		*Qtde*	6	3	7	4	1	
SIM	92,0%	46,9%	85,9%	67,0%	82,1%		*SIM*	88,0%	82,4%	94,6%	22,7%	82,7%		*SIM*	64,2%	79,4%	80,2%	79,5%	-	
Qtde	9	4	3	5			*Qtde*	4	10	3	4			*Qtde*	5	7	5	4		
SIM	81,3%	92,6%	28,7%	59,5%			*SIM*	22,7%	81,7%	81,9%	82,7%			*SIM*	32,7%	80,2%	80,5%	79,5%		
Qtde	7	10	4				*Qtde*	6	4	11				*Qtde*	12	5	4			
SIM	46,9%	82,1%	92,6%				*SIM*	40,5%	82,7%	84,0%				*SIM*	80,8%	32,7%	79,5%			

TESE 3

Abordagem A							*Abordagem B*							*Abordagem C*						
Grupos	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	*Grupos*	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**	*Grupos*	**G1**	**G2**	**G3**	**G4**	**G5**	**G6**
Qtde	3	12	1	2	1	2	*Qtde*	5	5	5	2	1	2	*Qtde*	3	3	9	2	3	1
SIM	65,4%	66,8%	-	76,9%	-	67,1%	*SIM*	63,0%	83,4%	65,7%	60,7%	-	42,9%	*SIM*	76,3%	60,7%	63,2%	42,9%	75,36	-
Qtde	8	4	4	4	1		*Qtde*	7	5	6	2	1		*Qtde*	5	9	4	1	2	
SIM	64,6%	65,8%	58,1%	81,1%	-		*SIM*	66,1%	83,4%	62,1%	42,9%	-		*SIM*	75,8%	63,2%	62,1%	-	42,9%	
Qtde	12	1	4	4			*Qtde*	5	7	3	6			*Qtde*	3	5	9	4		
SIM	66,0%	-	81,1%	58,1%			*SIM*	83,4%	66,1%	49,9%	62,1%			*SIM*	49,9%	75,8%	63,2%	62,1%		
Qtde	5	4	12				*Qtde*	9	5	7				*Qtde*	6	9	6			
SIM	47,5%	81,1%	66,0%				*SIM*	54,8%	83,4%	66,1%				*SIM*	51,5%	63,2%	67,1%			

Na Tabela 1, "GK" representa os grupos formados de acordo com o valor de K, ou seja, para $K = 6$, são formados seis grupos (G1 a G6), e assim por diante. A linha "Qtde" mostra o número de membros do grupo, e na linha "SIM" estão os valores de similaridade interna referente àquele grupo. A similaridade interna é o valor resultante da aplicação da SC entre os membros do grupo. Com base nos resultados apresentados pela Tabela 1, destacam-se as seguintes análises:

- Para a coleção de documentos com base na argumentação inicial sobre a **Tese 1**:
  - Para $K = 6$, ou seja, para formação de seis grupos, as três abordagens apresentam uma distribuição de membros por grupo semelhante. Contudo, os agrupamentos realizados para Abordagem A apresentam uma melhor distribuição destes membros, uma vez os valores de similaridade interna dos grupos apresentam uma variação menor em relação as outras duas abordagens.
  - Para $K = 5$, as abordagens B e C apresentam o mesmo número de membros por grupo e uma baixa variação da similaridade interna dos grupos.
  - Para $K = 4$, os grupos com maior valor de similaridade interna foram gerados com a aplicação da abordagem A.
  - Para $K = 3$, a abordagem B gera grupos com a menor variação da similaridade interna entre eles.
- Para a coleção de documentos com base na argumentação inicial sobre a **Tese 2**:
  - Para $K = 6$, com a aplicação da abordagem A, a variação entre as similaridades internas dos grupos se destaca, uma vez que quatro grupos possuem similaridade interna superior a 80% e os outros dois não ultrapassam 60%. Com a abordagem B, temos um agrupamento que apresenta similaridade interna igual a 0%, e com a abordagem C, um grupo é formado apenas por um participante. O que demonstra a incompatibilidade da argumentação de um dos participantes, em relação aos outros.
  - Para $K = 4$, a abordagem C realiza a distribuição com menor variação entre o número de membros do grupo.
  - Para todos os valores de K, pelo menos um dos grupos gerados com a abordagem A apresenta similaridade interna superior a 90%.
- Para a coleção de documentos com base na argumentação inicial sobre a **Tese 3**:
  - Para $K = 6$, na abordagem A, dois participantes não são agrupados com os demais, assim, são formados dois grupos com apenas um membro.
  - Para $K = 6$ e $K = 5$, nas abordagens B e C, o texto produzido por um dos participantes não possui semelhança com os textos produzidos pelos demais participantes, dessa forma este participante fica isolado em um grupo.
  - Para $K = 4$, com a abordagem A, os grupos formados apresentam similaridade interna superior a 55%, contudo um dos participantes fica isolado em um grupo.
  - Para $K = 3$, as abordagens B e C apresentam agrupamentos com menor variação entre o número de membros. E, a similaridade interna dos grupos apresenta resultados maiores que 50%.

6.1. Apoio à mediação pedagógica

Ao analisar a argumentação inicial feita pelo aluno, o professor mediador visualiza o grau de entendimento do aluno e da turma sobre o assunto que está em debate. Com base nos agrupamentos feitos com a aplicação do núcleo do Alpes, a atividade de análise dos textos

e posterior agrupamento dos alunos é realizado de forma automatizada. Assim, o professor terá auxílio computacional para aplicar as seguintes mediações pedagógicas:

- Auxiliar o aluno que não conseguiu desenvolver sua argumentação sobre o assunto tratado pela tese, indicando referencial teórico sobre o assunto;
- Indicar revisores que tenham opiniões distintas sobre o assunto, ao comparar as argumentações iniciais. Desta forma, os revisores irão contrapor a argumentação inicial com argumentos diferenciados a fim de gerar desequilíbrios no aluno que está sendo revisado;
- Realizar comentários sobre a argumentação inicial, incentivando o aluno a desenvolvê-la com outros elementos;
- Comparar os agrupamentos gerados com a análise da argumentação inicial com os grupos gerados com a análise da argumentação final, com o objetivo de verificar quais alunos continuaram juntos num mesmo grupo;
- Ao comparar a argumentação inicial e a final de um mesmo aluno, e possível verificar quais foram as principais alterações entre elas, e assim verificar a evolução das argumentações.

7. Considerações finais

Ao aplicar a arquitetura pedagógica Debate de Teses para auxiliar no processo de aprendizagem sobre determinado tema, a análise dos textos produzidos nas etapas a fim de viabilizar as mediações pedagógicas destacadas na subseção 6.1, demandam tempo e esforço do professor. Desta forma, ao utilizarmos ferramentas computacionais que minimizem esse o tempo e esforço, proporciona-se condições que facilitam a aplicação das mediações pedagógicas.

Os agrupamentos gerados com a aplicação do Alpes, apresentam-se como solução favorável ao processo de mediação pedagógica, ao incluir em um mesmo grupo alunos que tenham desenvolvido textos similares na fase de argumentação inicial, categorizando os alunos com conhecimentos prévios semelhantes sobre o assunto que está em debate.

Concluiu-se, ao comparar a aplicação de TF-IDF + KM com a aplicação de LSA + KM, o uso de LSA na preparação dos dados que serão agrupados pelo K-Means se mostra favorável, pois as relações identificadas entre os termos com a aplicação de LSA, podem, potencialmente, aumentar os pesos atribuídos aos termos mais importantes da coleção, possibilitando que a análise de similaridade aplicada sobre os textos alcance melhores resultados. Além disso, ao utilizar técnicas de LSA para conferir se os textos produzidos pelos alunos estão de acordo com o tema proposto pela tese, utiliza-se uma abordagem genérica que viabiliza a utilização do sistema para análise de textos produzidos sobre qualquer área de conhecimento.

As seguintes melhorias serão implementadas em versões futuras do Alpes: a) Utilizar a WordNet.Br para verificar a relação semântica entre os termos da coleção de documentos; b) Utilizar o *corpus* Bosque do projeto Floresta Sintática[7] para implementar a normalização de termos (troca de termos sinônimos por um único termo); c) Utilizar uma combinação de LSA com *n-gramas*.

References

Avila, R. L. F. d. and Soares, J. M. (2013). Uso de técnicas de pré-processamento textual e algoritmos de comparação como suporte à correção de questões dissertativas: experi-

[7] http://www.linguateca.pt/Floresta/

mentos, análises e contribuições. In *Anais do 24º Simpósio Brasileiro de Informática na Educação*, volume 1, pages 727–736.

Azevedo, B. F. T., Behar, P. A., and Reategui, E. B. (2011). Análise das mensagens de fóruns de discussão através de um software para mineração de textos. In *Anais do 22º Simpósio Brasileiro de Informática na Educação*, volume 1, pages 20–29.

Carvalho, Marie Jane S. e Nevado, R. N. e. M. C. S. d. (2005). Arquiteturas pedagógicas para educação à distância: Concepções e suporte telemático. In *Anais do 16º Simpósio Brasileiro de Informática na Educação*, volume 1, pages 351–360.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230.

Feldman, R. and Sanger, J. (2006). *Text Mining Hebook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, NY, USA.

Kakkonen, T. and Sutinen, E. (2004). Automatic assessment of the content of essays based on course materials. In *Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on*, pages 126–130. IEEE.

Kim, J. and Shaw, E. (2014). Scaffolding student online discussions using past discussions: Pedabot studies. *Artificial Intelligence Review*, 41(1):97–112.

Klein, R., Kyrilov, A., and Tokman, M. (2011). Automated assessment of short free-text responses in computer science using latent semantic analysis. In *Proceedings of the 16th annual joint conference on innovation and technology in computer science education*, pages 158–162. ACM.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Mason, O. and Grove-Stephensen, I. (2002). Automated free text marking with paperless school. © Loughborough University.

Michels, A. B. (2014). Do fazer ao compreender no contexto da educação a distância: Uso de arquiteturas pedagógicas no processo de empreender. Master's thesis, Programa de Pós Graduação em Educação - UFRGS.

Nevado, Rosane Neves e Menezes, C. S. d. e. V. J. R. R. M. (2011). Debate de teses – uma arquitetura pedagógica. In *Anais do 22º Simpósio Brasileiro de Informática na Educação*, volume 1, pages 820–829.

Orengo, V. and Huyck, C. (2001). A stemming algorithm for the portuguese language. In *String Processing and Information Retrieval, 2001. SPIRE 2001. Proceedings. Eighth International Symposium on*, pages 186–193.

Panceri, S. S. and Menezes, C. S. d. (2014a). Alpes: Um sistema multiagente para análise de produções textuais no contexto de um debate de teses. In *Anais do 25º Simpósio Brasileiro de Informática na Educação*, volume 1.

Panceri, S. S. and Menezes, C. S. d. (2014b). O suporte computacional como auxílio a mediação pedagógica em um debate de teses. In *TISE'14*, volume 1.