Ovídio José Francisco

Avaliação de técnicas de Recuperação de Informação para Organização e Extração de Conhecimento de Documentos de Reuniões

Sorocaba, SP

16 de julho de 2018

Ovídio José Francisco

Avaliação de técnicas de Recuperação de Informação para Organização e Extração de Conhecimento de Documentos de Reuniões

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCC-So) da Universidade Federal de São Carlos como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação. Linha de pesquisa: Aprendizado de Máquina.

Universidade Federal de São Carlos – UFSCar

Centro de Ciências em Gestão e Tecnologia – CCGT

Programa de Pós-Graduação em Ciência da Computação – PPGCC-So

Orientador: Prof.ª Dr. Katti Faceli

Coorientador: Prof. Dr. Rafael Geraldeli Rossi

Sorocaba, SP

16 de julho de 2018

Resumo

Em um contexto em que a grande parte das informações armazenadas pelas organizações está em formato textual, o desenvolvimento de ferramentas computacionais para extração e organização automática dessas informações é uma tarefa que retém atenção e relevância. Os modelos de extração de tópicos são recorrentemente empregados nessa tarefa. Esses modelos são capazes de estabelecer relações e encontrar padrões latentes em coleções de documentos textuais. No entanto, há um desafio adicional em documentos constituídos por múltiplos assuntos. Para textos onde há transição de assuntos faz-se necessário, em primeiro lugar, encontrar porções de texto que tratam de um único assunto. Para isso, as técnicas de Segmentação Textual são utilizada para dividir um texto em segmentos com um assunto relativamente independente. O uso combinado de algoritmos de Segmentação Textual e Extração de Tópicos, pode então ser aplicado para criar uma estrutura que ajuda a entender os dados textuais, os quais são inerentemente não estruturados. A criação de uma estrutura organizada em tópicos, que incorpora informações latentes sobre o *corpus* favorece técnicas de Recuperação de Informação. Essa abordagem permite a expansão do espaço de busca além do conjunto de termos original de cada segmento, e a identificação de trechos mais relevantes à consulta. Nesse trabalho, é apresentada uma metodologia para conectar as técnicas de Segmentação Textual aos modelos de Extração de tópicos a fim de gerar uma estrutura derivada de um *corpus* não estruturado, a qual concentra os textos originais acrescidos de informações latentes e organizadas por sua semelhança semântica. A pesquisa desse trabalho de mestrado investigou as técnicas de Segmentação Textual e Extração de Tópicos para o desenvolvimento de um sistema de Recuperação de Informação em uma coleção de atas de reunião. Desenvolveu-se uma sistema para entregar ao usuário os segmentos mais relevantes à consulta. Os trechos exibidos são agrupados por tópicos, o que também possibilita consultas exploratórias à base dados, por meio da navegação por grupos de segmentos relacionados semanticamente. Além disso, trechos pouco relevantes são omitidos, permitindo resultados concentrados no assunto pesquisado. Foram avaliados cinco técnicas de Segmentação Textual e três modelos de Extração de Tópicos da literatura. Com base no conjunto de atas estudado, criou-se um *corpus* anotado manualmente com informações sobre a composição temática das atas e as transições entre os assuntos. O *corpus* anotado serviu como referência para avaliação objetiva das técnicas de Segmentação Textual. Avaliou-se ainda, os resultados obtidos com os modelos de Extração Tópicos junto a profissionais ligados ao contexto de atas de reunião. Os resultados sugerem que as técnicas utilizadas e a metodologia apresentadas entregam respostas satisfatórios. Contudo, mais dados podem ser necessários para execução de novos experimentos a fim de aprimorar as técnicas utilizadas. A implementação do sistema e das ferramentas utilizadas para esse trabalho, bem como os resultados obtidos nas avaliações,

juntamente com o *corpus* anotado, são as principais contribuições desse trabalho no intuito de dar aportes ao desenvolvimento de novos trabalhos.

Palavras-chaves: Segmentação Textual. Extração de Tópicos. Recuperação de Informação.

Abstract

In a context where the most informations stored by the organizations is in the text format, the development of computer tools for automatic extraction and organization of this information is a task that holds attention and relevance. The extraction topic models are often employed in this task. This models are able to establish relations between documents and found latents patterns in sets of them. However, there is an additional challenge for documents composed by multiples subjects. For texts with subject shifts, it is necessary, in fist, found the chunks of texts that address a single subject. So, the Text Segmentation techniques are used to break a text in segments with a relatively independent issue. The combination of the Text Segmentation and Topic Extraction algorithms, can be used to make an structure that helps to understanding textual data, which are inherently non structured. The creation of an topic organized structure, that incorporates latent information concerning the *corpus*, favors Information Retrievals techniques. This approach allows query expansion space, besides the original set of terms of each segment, and the identification most relevant pieces of text. This work presents an methodology to connect the Text Segmentation techniques to the Topic Extraction Models, in order to generates an derivative structure from a non structured *corpus*, which concentrates the original texts plus the latent informations and organized by semantic likeness. The research for this thesis investigates the Text Segmentation techniques and the Topic Extraction Models to develop an Information Retrieval system for meeting minutes. We develops an system to give to the user the most relevant segments to the query. The segmentos presented are clustered by this topics, that also enables exploratory searches to the data base, through browsing by groups of semantically related segments. Furthermore, less relevant segments are omitted, allowing results focused on the researched subject. Five techniques of Text Segmentation and three Topic Extraction models from the literature was evaluated. Based on the set of meeting minutes analysed, we create an manually annotated *corpus* with informations on the thematic composition of the meeting minutes and about the issue shifts. The annotated *corpus* served as reference to objective evaluation of the Text Segmentation techniques. We also evaluate the results obtained with the Topic Extraction Models. This results were analysed by related to the meeting minutes context. The results points the employed techniques and the methodology presented gives satisfactory answers. However, more data can be necessary to leads new experiments in order to improve the used techniques. The system implementation and the tools used in the work, as well the results obtained in the evaluations, plus the annotated *corpus* are the principal contributions of this work, in order to support next works.

Key-words: Text Segmentation. Topic Extraction. Information Retrieval.

Lista de símbolos

D	Conjunto de documentos de uma coleção.
n	Número total de documentos em uma coleção.
m	Número total de termos em uma coleção de documentos.
T	Conjunto de termos de uma coleção.
t_i	i-ésimo termo do vocabulário da coleção de documentos.
w_i	Peso i-ésimo termo.
$\vec{d}$	representação vetorial do documento d.
$\vec{d}$	representação vetorial da consulta q.
q	consulta do usuário.
c_i	i-ésima sentença da coleção de documentos.
d_i	i-ésimo documento da coleção de documentos.
S	Lista de segmentos.
s_i	i-ésimo segmento do conjunto S
N	Total de sentenças do documento.
h	Total de segmentos do conjunto S
Z	Conjunto de tópicos.
z_i	i-ésimo tópico do conjunto de tópicos.
R_q	Conjunto de documentos relevantes à consulta q.

Sumário

1 Introdução

A popularização dos computadores possibilitou o armazenamento cada vez maior de conteúdos digitais, sendo bastante comum, o formato textual como livros, documentos, e-mails, redes sociais e páginas web. A produção de textos gera informações em volumes crescentes que superaram a capacidade humana de análise manual. Além disso, dados textuais quase sempre apresentam-se em formato não estruturado, caracterizado pela ausência de uma organização pré-definida que facilite a busca em sistemas computacionais. Contudo, dados textuais possuem uma organização linguística intrínseca o que possibilita a pesquisa com o uso de ferramentas automáticas para manipulação e consulta a bases dados textuais não estruturadas (CAO, 2017; MANNING; RAGHAVAN; SCHüTZE, 2008).

Assim, os processos de extração automática de conhecimento em coleções textuais são essenciais, e ao mesmo tempo, constituem um desafio, devido às suas características como o formato não estruturado e trechos com diferentes níveis de importância, desde informações essenciais até textos pouco informativos ou mesmo irrelevantes (AGGARWAL; ZHAI, 2012; JEONG; TITOV, 2010; TAGARELLI; KARYPIS, 2013).

Além dos tipos de informações mais comuns que são armazenados no formato textual, como e-mails, relatórios, artigos e postagens em redes sociais, têm-se também o armazenamento das atas de reuniões, as quais permitem às organizações a documentação oficial de reuniões em arquivos digitais, facilitando a sua confecção e compartilhamento, bem como consulta às decisões tomadas. Reuniões são tarefas presentes em ambientes de gestão e organizações de um modo geral, onde discute-se problemas, soluções, propostas, alterações de projetos e frequentemente são tomadas decisões importantes onde a comunicação entre os membros da reunião é feita de forma majoritariamente verbal.

Para que seu conteúdo possa ser registrado e externalizado, adota-se a prática de escrever seu conteúdo em atas (SCHWARTZ-ZIV; WEISBACH, 2013; LEE et al., 2011). Por exemplo, nas reuniões do conselho de um programa de pós-graduação de uma universidade, são decididos, quais são os critérios para credenciamento e permanência de docentes no programa. Ao longo do tempo, esse tema pode ser discutido e mencionado diversas vezes, podendo os critérios inclusive passar por significativas alterações, devido a diversos fatores. O coordenador do programa pode desejar recuperar qual foi a decisão mais recente, para poder aplicar os critérios a um potencial novo membro do programa, ou os membros do conselho podem desejar rever o histórico de tudo o que já foi discutido/decidido sobre o tema, para poder propor alterações nas regras, de forma mais adequada.

As atas de reunião possuem características particulares. Frequentemente apresentam um texto com poucas quebras de parágrafo e sem marcações de estrutura, como capítulos,

seções ou quaisquer indicações sobre o tema do texto. Devido a fatores como a não estruturação e volume dos textos, a localização de um assunto em uma coleção de atas é uma tarefa custosa, especialmente considerando o seu crescimento de seu número em uma instituição.

As organizações costumam manter seus documentos eletrônicos organizados em pastas e nomeá-los com informações básicas sobre a reunião a que se refere como a data e alguma referência cronológica, por exemplo *"37ª Reunião Ordinária do Conselho ..."*. Essa organização facilita a localização dos arquivos com ferramentas que fazem buscas pelo nome dos arquivos e pastas, sem levar em conta o teor dos documentos. Também é comum o uso de ferramentas que fazem buscas nos conteúdos dos documentos, buscando por ocorrências de palavras-chave nos textos. Essas ferramentas permitem buscas combinadas com operadores lógicos como *and*, *or* e *not* ou ainda suporte a expressões regulares. Esse recurso, conhecido como *grepping*[1], produz resultados satisfatórios em muitos casos. Por outro lado, traz algumas desvantagens como: 1) transfere certa complexidade da tarefa ao usuário 2) não há suporte a padrões mais flexíveis como a proximidade entre as palavras ou palavras que estejam na mesma sentença 3) informa apenas se um documento casa ou não com a consulta do usuário com base na presença ou ausência dos termos da consulta (AGGARWAL; ZHAI, 2012; MANNING; RAGHAVAN; SCHüTZE, 2008).

Ainda nesse contexto, usa-se outras técnicas de Recuperação de Informação como o Modelo de Espaço Vetorial para ranquear documentos atribuindo pontuações para a similaridade de cada par documento/consulta. Com isso, é possível apresentar os documentos ordenados conforme a sua relevância com a consulta (GUTIERREZ et al., 2016; CROFT; METZLER; STROHMAN, 2009; MANNING; RAGHAVAN; SCHüTZE, 2008). Contudo, essas técnicas baseiam-se na frequência de palavras, em que os documentos e consultas são vistos como conjuntos de termos sem levar em conta relações entre termos que compartilham um mesmo tópico dentro do domínio (WEI; CROFT, 2006). Por exemplo, as consultas "*alunos bolsa CAPES*" e "*suporte financeiro a pesquisa*" podem estar relacionados a um assunto em comum, nesse caso, a transferência de valores monetários como apoio a carreira acadêmica. Utilizando-se das técnicas até agora mencionadas, obteria-se resultados distintos para cada caso, uma vez que as consultas não compartilham termos e não há relação direta ente eles. Como efeito, os resultados de cada consulta limitariam-se a documentos que compartilham termos com a consulta. Essas técnicas produzem resultados melhores a medida que o usuário fornece termos mais acertados na consulta, o que por vezes é dependente de certo conhecimento e familiaridade com o domínio no qual a coleção de documentos está inserida. Além disso, o retorno ao usuário é uma lista documentos integrais, o que pode exigir uma segunda busca dentro de um documento para encontrar o trecho desejado.

[1] O nome *grepping* é uma referência ao comando `grep` do Unix

Uma vez que a ata registra a sucessão de assuntos discutidos na reunião, um sistema de recuperação de informação idealmente deve retornar ao usuário apenas os trechos que tratem do assunto pesquisado ao invés de documentos inteiros. Assim, cada trecho com um assunto predominate pode ser considerado um subdocumento. Portanto, em primeiro lugar, há a necessidade de descobrir onde há mudanças de assunto no texto.

Técnicas de segmentação automática de textos (segmentação textual) podem ser aplicadas com esse propósito. Elas podem dividir um documento em segmentos que contenham um assunto relativamente independente e gerar um conjunto de subdocumentos derivado da coleção de atas original (AGGARWAL, 2018; BOKAEI; SAMETI; LIU, 2015; SAKAHARA; OKADA; NITTA, 2014; MISRA et al., 2009; EISENSTEIN; BARZILAY, 2008).

Contudo, a segmentação textual apenas indica as transições de assuntos ao longo do texto, sem indicações sobre o teor dos segmentos. O assunto de cada trecho pode ser estimado por meio de modelos de extração de tópicos. Essa técnica possibilita a formação de grupos de segmentos que compartilham o mesmo assunto bem como indicar palavras que melhor descrevem o grupo (WEI, 2007). Com isso, obtém-se uma organização da coleção de documentos que favorece técnicas para navegação e consulta à coleção de documentos (MARCACINI; REZENDE, 2010). Tais modelos podem eleger um conjunto de termos importantes para um ou mais assuntos, bem como ranquear documentos por sua relevância para determinado tema (FALEIROS, 2016; YI; ALLAN, 2009).

Devido às características das atas, como a multiplicidade de assuntos, e ausência de meta-informação, as técnicas de segmentação podem ser empregadas em conjunto com modelos de extração de tópicos para criar uma estrutura de dados derivada da coleção de documentos original. Essa abordagem visa, em primeiro lugar, identificar os assuntos tratados em cada ata e gerar uma coleção de subdocumentos derivados da coleção de atas originais e, a partir disso, utilizar modelos de extração de tópicos para encontrar relações latentes entre os subdocumentos e termos da coleção. Como resultado, obtém-se uma organização da coleção em que os segmentos são agrupados por assuntos e acrescidos de um conjunto de termos que representam os principais tópicos ou assuntos identificados na coleção de documentos, dessa forma, incorporando conhecimento aos dados originais. Esse novos atributos e a organização das atas por temas podem ser usados para expandir o espaço de busca a fim de aprimorar técnicas de recuperação de informação em um sistema para extração de conhecimento em coleções de atas de reunião.

Essa abordagem traz vantagens para a recuperação de informação em coleções de documentos com múltiplos tópicos como atas de reunião. A segmentação das atas permite ao sistema criar uma base de documentos mais simples, em que os assuntos estão isolados em segmentos, formando assim, um *corpus* mais adequado aos métodos de aprendizado de máquina e recuperação de informação, empregados nesse trabalho. Além disso, permite ao

sistema final exibir apenas os trechos onde o assunto pesquisado está presente ao invés de entregar documentos integrais (TAGARELLI; KARYPIS, 2013; JEONG; TITOV, 2010; PRINCE; LABADIÉ, 2007; HUANG et al., 2003).

Modelos de extração de tópicos podem se integrar ao processo de recuperação de informação usando os agrupamentos e seus descritores como uma forma de descrição da coleção (ZHAI, 2017; YI; ALLAN, 2009), bem como relacionar termos distintos com um determinado tópico em comum (WEI; CROFT, 2006). A busca por palavras-chave em descritores transfere esforço computacional da varredura dos documentos para a etapa de extração de tópicos. Essa estratégia evita processamento e lentidão no momento da pesquisa de maneira semelhante à criação de índices para aumentar a eficiência em sistemas de recuperação de informação. Além disso, encontra relações entre termos e documentos sem necessidade de conhecimento externo sobre o domínio. Assim, nesse trabalho os modelos de extração de tópicos são utilizados para incorporar informação aos segmentos com a finalidade de aprimorar o ranqueamento dos resultados (MARCACINI; REZENDE, 2010; WEI; CROFT, 2006). Além disso, ao agrupar os segmentos por tópicos, tem-se uma organização dos documentos que permite a visualização de segmentos semelhantes permitindo a navegação e exploração aos grupos além das consultas por palavras-chave.

Diante desse cenário, o objetivo desse trabalho de mestrado é propor o desenvolvimento uma ferramenta para identificar, organizar e apresentar assuntos registrados em atas de reunião utilizando a estrutura latente de documentos segmentados em conjunto com técnicas de recuperação de informação. Como objetivos específicos, esse trabalho visa dar início a investigação de métodos de segmentação textual, extração de tópicos e recuperação de informação no contexto de atas de reunião. Para conhecer a eficiência das técnicas de segmentação textual, seus resultados foram analisados tendo como referência anotações coletadas de profissionais que desempenham atividades ligadas a atas e reuniões. Avaliar junto ao usuário a qualidade dos subdocumentos apresentados quanto ao agrupamento e relevância das informações contidas. Para isso, foi feito uma coleta de dados sobre a percepção de usuários sobre a qualidade dos resultados apresentados em uma consulta a coleção de atas por meio do sistema. Dessa forma, busca-se ajudar a suprir a necessidade de ferramentas para para esse cenário e contribuindo com uma metodologia de extração de informação em documentos com múltiplos assuntos. Além disso, disponibilizar o sistema implementado bem como os dados coletados durante a análise das técnicas de forma a contribuir com novos trabalhos relacionados a esse contexto.

No Capítulo 2 são apesentados os principais conceitos da Recuperação de Informação, de Mineração de Textos, Segmentação Textual e Extração de Tópicos. No Capítulo 3 é proposto um sistema para extração de conhecimento em atas de reunião e inicia-se o seu desenvolvimento com base nas técnicas mencionadas. Ainda nesse capítulo, é mostrado um estudo de caso da aplicação das técnicas de Segmentação Textual e Extração de Tópicos em

uma coleção de atas. No Capítulo 4 é apresentada uma avaliação objetiva em algoritmos de Segmentação Textual a fim de analisar seus resultados e escolher um segmentador a ser utilizado no sistema proposto. anotações No Capítulo 5 é apresentada uma avaliação subjetiva dos resultados dos modelos de Extração de Tópicos sob um *corpus* composto por segmentos de atas. Adicionalmente, avaliou-se o segmentador utilizado, uma vez que este está ligado aos resultados avaliados. Por fim as conclusões do trabalho, as limitações dessa pesquisa e os trabalhos futuros são apresentados no último capítulo dessa dissertação.

2 Conceituação Teórica

A popularidade dos computadores permite a criação e compartilhamento de textos. Com isso, a quantidade de informação disponível facilmente extrapola a capacidade de humana de leitura e análise das coleções de documentos, estejam elas disponíveis na Internet ou em computadores pessoais. A necessidade de simplificar e organizar grandes coleções de documentos criou uma demanda por técnicas que permitam ao usuário acessar informações de seu interesse. Para esse fim, foram desenvolvidas técnicas para descobrir, extrair e agrupar textos de grandes coleções, entre essas, a modelagem de tópicos (HOFMANN, 1999; DEERWESTER et al., 1990; LEE; SEUNG, 1999; BLEI, 2012) e técnicas para identificar e recuperar informações com base em buscas de usuários.

Os modelos de extração de tópicos podem ter sua performance aprimorada quando usada em conjunto com técnicas de segmentação textual em documentos que contém assuntos diversos. Dessa forma é possível identificar em um mesmo documento, múltiplos tópicos, bem com o trecho em que cada assunto é abordado (TAGARELLI; KARYPIS, 2013; NGUYEN, 2011). Essa abordagem possibilita a criação de uma base de dados estruturada a qual pode ser utilizada por métodos de Recuperação de Informação a fim de entregar ao usuário apenas os segmentos mais relevantes à consulta, bem como expandir o espaço de busca por meio de variáveis latentes identificas pelos modelos de extração de tópicos.

Nas próximas seções serão apresentados conceitos básicos sobre os modelos de Recuperação de Informação, o Preprocessamento de documentos, a Segmentação Textual e Extração de Tópicos. Mais adiante, no Capítulo 3 essas técnicas serão aplicadas no desenvolvimento de um sistema de Recuperação de Informação em uma coleção de atas de reunião.

2.1 Modelos de Recuperação de Informação

Devido à popularização dos computadores e a grande disponibilidade de documentos em formato digital, em especial na *web*, a área da Recuperação de Informação (RI) tem recebido atenção de pesquisadores nas últimas décadas. Recuperação de Informação é a área da computação que envolve a aplicação de métodos computacionais no tratamento e busca de informação em bases de dados não estruturados, usualmente grandes coleções de documentos textuais armazenados em dispositivos eletrônicos. De fato, não há dados completamente não estruturados ao se considerar a estrutura linguística latente em documentos textuais. O termo “não estruturado” se refere a dados que não oferecem uma estrutura clara para sistemas computadorizados (MANNING; RAGHAVAN; SCHüTZE,

2008; GUTIERREZ et al., 2016).

A tarefa central da recuperação de informação é encontrar informações de interesse dos usuários e exibí-las. Nesses sistemas o usuário expressa sua necessidade por meio da formulação de uma consulta, usualmente composta por um conjunto de palavras-chave. Então, o sistema apresenta os resultados da busca, frequentemente documentos, em ordem de relevância com a consulta.

Um modelo de recuperação de informação deve criar representações de documentos e consultas a fim de predizer a necessidade expressa nos termos da consulta. Com base na entrada do usuário esses modelos buscam por documentos similares aos termos da consulta. Segue abaixo a descrição dos três modelos clássicos para recuperação de informação.

2.1.1 Modelo Booleano

O modelo booleano ou modelo lógico foi um dos primeiros modelos aplicados a recuperação informação sendo utilizado a partir de 1960. Nesse modelo uma consulta é considerada uma sequencia de termos conectador por operadores lógicos como AND, OR e NOT. Como resultado, classifica cada documento como relevante ou não relevante à consulta, sem gradação de relevância. Esses operadores lógicos podem ser manipulados por usuários com algum conhecimento em álgebra booleana para aumentar a quantidade de resultados ou restringí-la.

Uma desvantagem desse modelo é que não é possível medir a relevância de um documento em relação a consultado do usuário, devia a essa limitação não há informação que permita a ordenação dos resultados, que é uma característica relevante para muitos SRI. Já as vantagens desse modelo são a facilidade de implementação e a possibilidade de usuários experientes usarem os operadores lógicos como uma forma de controle sobre os resultados da busca. Por outro lado, para usuários inexperientes isso pode ser considerado uma desvantagem, uma vez que o uso de expressões lógicas não é intuitivo. Apesar dos problemas apresentados, visto sua simplicidade, esse modelo foi largamente utilizado em sistemas comerciais.

2.1.2 Modelo Espaço Vetorial

Uma das formas mais comuns para representação textual é conhecida como Modelo Espaço Vetorial (*Vectorial Space Model* - VSM), onde os documentos e consultas são representados como vetores em um espaço Euclidiano t-dimensional em que cada termo extraído da coleção é representado por uma dimensão (REZENDE, 2003). Considera-se que um documento pode ser representado pelo seu conjunto de termos, onde cada termo k_i de um documento d_j associa-se um peso $w_{ij} \geq 0$ que indica a importância desse termo no documento. De forma similar, para uma consulta q, associa-se um peso $w_{i,q}$ a cada termo

consulta. Assim o vetor associado ao documento d_j é dado por $\vec{d_j} = (w_{j,1}, w_{j,2}, ..., w_{j,p})$ e o vetor associado a consulta q é dado por $\vec{q} = (w_{q,1}, w_{q,2}, ..., w_{q,l})$. No modelo vetorial, a similidade entre um documento d_j e uma consulta q é calculada pela correlação entre os vetores $\vec{d_j}$ e $\vec{q}$, a qual pode ser medida pelo cosseno (Equação 2.7) do ângulo entre esses vetores, conforme mostrado adiante na Seção 2.2.

Avaliar a relevância de um documento sob uma consulta é fundamental para os modelos de RI. Para isso pode-se utilizar medidas estatísticas simples como a frequência do termo, conhecida como TF (do inglês *Term Frequency*) e a frequência de documentos, conhecida como DF (do inglês *Document Frequency*). A frequência do termo indica o número de vezes que um termo ocorre na coleção de documentos. A frequência de documentos, indica o número de documentos que contém ao menos uma ocorrência de um determinado termo. Considera-se que os termos que ocorrem frequentemente em muitos documentos, em geral, não trazem informações úteis para discriminar a relevância dos documentos, então, a fim de diminuir o peso de termos altamente frequentes, usa-se o fator IDF (*Inverted Document Frequency*), que é o inverso da número de documentos que contem um termo. O IDF é a medida de informação que um termo fornece com base em quão raro ou comum esse termo é para a coleção. Seja N o número de documentos de uma coleção e n_i o número de documentos onde o termo k_i ocorre, o cálculo de IDF é dado por:

$$IDF(k_i) = log\frac{N}{n_i} \ . \tag{2.1}$$

Entre as medidas mais populares para ranqueamento de buscas está a TF-IDF (*Term Frequency-Inverted Document Frequency*) que pondera a frequência de um termo em um documento com sua frequência na coleção total de documentos. Assim, a relevância de um termo para um documento é dada por:

$$w_{i,j} = freq_{i,j} \cdot IDF(k_i) \ , \tag{2.2}$$

onde $freq_{i,j}$ é a frequência do termo k_i no documento d_j. A medida TF-IDF atribui valores altos para termos que ocorrem frequentemente em um documentos, e valores menores para termos que ocorrem poucas vezes em um documento ou em muitos documentos da coleção. A ideia da medida TF-IDF e quantificar a importância de um termo em um documento com base em sua frequência no próprio documento e sua distribuição ao longo da coleção de documentos (CROFT; METZLER; STROHMAN, 2009; SALTON; BUCKLEY, 1988; SHAMSINEJADBABKI; SARAEE, 2012; SALTON; ALLAN, 1994).

Uma vez que o modelo, por meio da Equação 2.7, calcula a similaridade entre os documentos e a consulta do usuário, é possível ranquear os resultados por ordem de relevância. Além disso, sua relativa simplicidade e flexibilidade, favorecem a aplicação desse modelo em sistemas de recuperação de informação (TAN; STEINBACH; KUMAR,

2005; CROFT; METZLER; STROHMAN, 2009; MANNING; RAGHAVAN; SCHüTZE, 2008).

2.1.3 Modelo Probabilístico

O modelo probabilístico é baseado no princípio da ordenação probabilística (*Probability Ranking Principle*) onde dada um consulta q e um documento d_j relevante a q, o modelo tenta estimar a probabilidade do usuário encontrar o documento d_j. O modelo assumente que para uma consulta q há um conjunto de documentos R_q que contém exatamente os documentos relevantes e nenhum outro, sendo este um conjunto resposta ideal que maximiza a probabilidade do usuário encontrar um documento d_j relevante a q.

Seja $\overline{R_q}$ o complemento de R de forma que $\overline{R_q}$ contém todos os documentos não relevantes à consulta q. Seja $P(R_q|d_j)$ a probabilidade do documento d_j ser relevante à consulta q e $P(\overline{R_q}|d_j)$ a probabilidade de d_j não ser relevante à q. A similaridade entre um documento d_j e uma consulta q é definida por:

$$sim(d_j, q) = \frac{P(R_q|dj)}{P(\overline{R_q}|dj)} \tag{2.3}$$

A fim de obter-se uma estimativa numérica das probabilidades, o modelo assume o documento como uma combinação de palavras e seus pesos aos quais atribui-se valores binários que indicam a presença ou ausência de um termo, isto é, $w_{ij} \in \{0, 1\}$ e $w_{iq} \in \{0, 1\}$. Seja $p_i = P(t_i|R_q)$ a probabilidade do termo t_i ocorrer em um documento relevante à consulta q, e $s_i = P(t_i|\overline{R_q})$ a probabilidade do termo t_i estar presente em um documento não relevante. Seja ainda $\prod_{i:d_i=1}$ o produto dos termos com valor 1. Então, pode-se calcular:

$$sim(d_j, q) = \prod_{i:d_i=1} \frac{p_i}{s_i} \cdot \prod_{i:d_i=0} \frac{1 - p_i}{1 - s_i} \ , \tag{2.4}$$

onde $\prod_{i:d_i=1}$ significa o produto dos termos com valor 1.

O modelo também supõe que os termos ocorrem independentemente no documento, ou seja, a ocorrência de um termo não influencia a ocorrência de outro. Partindo dessas suposições, a Equação 2.4 passa por transformações que incluem aplicação da regra de Bayes e simplificações matemáticas, e chega-se a Equação 2.5 conhecida como equação de Robertson-Spark Jones a qual é considerada a expressão clássica para ranqueamento no modelo probabilístico. Detalhes da dedução dessa equação podem ser encontrados em (CROFT; METZLER; STROHMAN, 2009; MANNING; RAGHAVAN; SCHüTZE,

2008; RIJSBERGEN, 1979),

$$sim(d_j, q) = \sum_{i=1}^{t} w_{i,j} \cdot w_{i,q} \cdot \sigma_{i/R} \ , \tag{2.5}$$

onde t é o número total de termos da coleção e

$$\sigma_{i/R} = \log \frac{p_i}{1 - p_i} + \log \frac{1 - s_i}{s_i} \ . \tag{2.6}$$

Esse modelo tem com principal desvantagem a necessidade de estimar a separação inicial entre R_q e $\overline{R_q}$, pois não se conhece inicialmente o conjunto dos documentos relevantes a uma consulta, o qual deve ser aprimorado por meio de interações com o usuário. Além disso, o modelo não leva em consideração a frequência dos termos na indexação do documento. O modelo apresenta como vantagem a característica de atribuir probabilidades as similaridades entre documentos e consultas, o que permite ranquear dos resultados por ordem de relevância.

2.2 Pré-processamento

O pré-processamento é a etapa que difere o processo de Mineração de Textos do processo de Mineração de Dados. Uma vez que os texto estão em formato inerentemente não estruturado, o problema se resume em adequá-los para uma representação estruturada, concisa e manipulável por algoritmos de Aprendizado de Máquina, além de limpeza e redução de termos.

Os documentos da coleção frequentemente encontram-se em diferentes formatos dado a diversidade de *softwares* para edição e publicação de conteúdos digitais. Assim, o processo se inicia com a extração do texto em formato plano (puro), em seguida transformado em um formato mais adequado. Os dados textuais têm como características serem esparsos e apresentar alta dimensionalidade. Por exemplo, uma coleção de documentos frequentemente contém milhares de palavras, ao passo que um documento específico irá conter uma pequena parcela dessa diversidade, em torno de algumas centenas. Essas características, por consequência, exigem que os dados originais sejam reduzidos, porém preservando as caraterísticas mínimas para os algoritmos utilizados a seguir.

Remoção de *Stop Words*

Considerando a alta dimensionalidade dos textos, os termos menos significativos devem ser removidos. *Stop words* são palavras pouco relevantes que não contribuem para a distinção do texto em tópicos ou categorias podem ser removidas, como artigos, preposições,

pronomes, verbos de estado[1]. Trata-se também como *stop words* as palavras de uso muito frequente dentro de um determinado domínio não são capazes de discriminar documentos e também não devem fazer parte dos atributos. A eliminação das *stop words* é feita com base em um conjunto de palavras conhecido como *stoplist*.

Corte por Frequências

Outra forma utilizada para seleção de termos é avaliar a importância de cada termo por meio de medidas estatísticas, como o TF (*term frequency*) e DF (document frequency). O método proposto em (LUHN, 1958) é uma técnica baseada na Lei de Zipf (ZIPF, 1932) também conhecida como Princípio do Menor Esforço, em que computando-se a frequência das palavras de um texto, e criando-se seu histograma em ordem decrescente, observa-se a chamada Curva de Zipf, na qual o k-ésimo termo mais comum ocorre com frequência inversamente proporcional a k. Os termos com alta frequência são considerados pouco relevantes por serem comuns à grande maioria dos documentos, enquanto termos mais raros não possuem caráter discriminatório suficiente. Assim, é possível estabelecer pontos de corte nos extremos da curva, a fim de manter termos com frequência intermediária, os quais são os mais representativos do documento (MARCACINI; REZENDE, 2010). Na Figura 1 é ilustrada a distribuição do termos mais relevantes em um documento e a curva de Zipf com dois cortes nas extremidades.

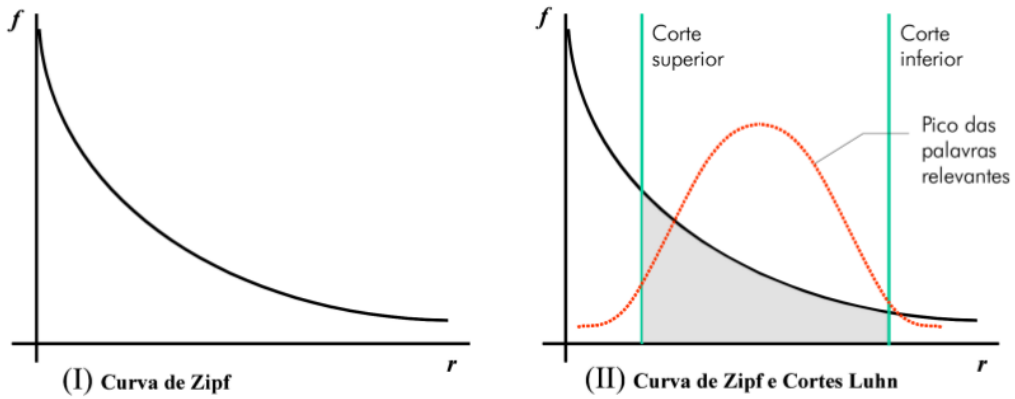


Figura 1 – A curva de Zipf e os cortes de Luhn (SOARES; PRATI; MONARD, 2008).

Stemming

A radicalização ou *stemming* é a redução das variações de uma palavra ao seu provável radical ou stem a fim de associar palavras semelhantes e diminuir a dimensionalidade da representação do texto. Nesse processo, as palavras são reduzidas ao seu provável radical ou *stem*, a fim de se associar palavras semelhantes e diminuir a dimensionalidade

[1] Apresentam uma situação inativa, onde o verbo não expressa uma alteração, mas apenas uma propriedade ou condição dos envolvidos.

da representação do texto. Por exemplo, os termos "*agenda*", "*agendamento*" e "*agendar*" dever ser todas reduzidas ao seu radical em comum "*agend*". Com isso, a dimensionalidade é diminuída ainda mais e tem-se um texto formado apenas por morfemas[2] com maior significância.

Em geral, algoritmos de *stemming* dependem do uso adequado da ortografia da língua em questão, inclusive com acentuação correta, sendo em alguns casos recomendado o uso de corretores automáticos na fase de pré-processamento. A língua portuguesa particularmente apresenta algumas dificuldades, na elaboração de algoritmos de *stemming*, das quais destacam-se o número elevado exceções e homófagos; palavras com mudanças no radical morfológico; nomes próprios que não podem ser radicalizados e frequência de termos estrangeiros. É possível identificar alguns erros apresentados pelos algoritmos de *stemming* que reduzem a qualidade os resultados da Mineração de Texto, como *oversteamming:* quando o algoritmo remove parte do radical e *understeamming:* quando o algoritmo não remove totalmente o sufixo.

O uso de *stemming*, de uma maneira geral, pode trazer algumas desvantagens das como a perda de contexto, pois palavras com sentidos diferentes podem resultar no mesmo radical, aumentando assim a quantidade de homônimos e a perca da precisão que diminui a variedade de palavras causando certa perda de informação. Contudo, eventuais perdas de informação por *stemming* não causam grandes impactos sobre a eficiência de algoritmos de text mining e seu uso se justifica pela redução da dimensionalidade da base de textos.

Representação de Textos

Uma das etapas mais importantes para as tarefas de Mineração Textos é a criação de uma representação adequada dos dados. Essa representação deve prover uma maneira estruturada para que possar ser utilizados por algoritmos de aprendizado de máquina. Os dados textuais se diferenciam de outros formatos estruturados como bancos de dados relacionais em que um dado é facilmente encontrado.

Uma das formas mais comuns para que a grande maioria dos algoritmos de aprendizado de máquina possa extrair padrões das coleções de textos é a representação no formato matricial conhecido como Modelo Espaço Vetorial (*Vectorial Space Model* - VSM) (REZENDE, 2003), onde os documentos são representados como vetores em um espaço Euclidiano m-dimensional em que cada termo extraído da coleção é representado por uma dimensão. Assim, cada componente de um vetor expressa a relação entre os documentos e as palavras. Essa estrutura é conhecida como matriz documento-termo (*document-term matrix*). Uma das formas mais populares dessa matriz é conhecida como *Bag Of Words* a qual é detalhada a seguir.

[2] Em Morfologia, um morfema é a menor unidade capaz de expressar significado.

Bag Of Words

Nessa representação, cada termo é transformado em um atributo (*feature*), em que a_{ij} é o peso do termo j no documento d_i e indica a sua relevância dentro da base de documentos (REZENDE, 2003). As medidas mais tradicionais para o cálculo desses pesos são a binária, onde o termo recebe o valor 1 se ocorre em determinado documento ou 0 caso contrário; *document frequency*, que é o número de documentos no qual um termo ocorre; *term frequency - tf*, atribui-se ao peso a frequência do termo dentro de um determinado documento; *term frequency-inverse document frequency, tf-idf*, pondera a frequência do termo pelo inverso do número de documentos da coleção em que o termo ocorre. Essa representação é mostrada pela Tabela 1.

	t_1	t_2	t_j	$\ldots$	t_n
d_1	a_{11}	a_{12}	a_{1j}	$\ldots$	a_{1n}
d_2	a_{21}	a_{22}	a_{2j}	$\ldots$	a_{2n}
d_i	a_{i1}	a_{i2}	a_{ij}	$\ldots$	a_{in}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_m	a_{m1}	a_{m2}	a_{mj}	$\ldots$	a_{mn}

Tabela 1 – Coleção de documentos na representação *bag-of-words*

Essa forma de representação sintetiza a base de documentos em um contêiner de palavras, ignorando a ordem em que ocorrem, bem como pontuações e outros detalhes, preservando apenas o peso de determinada palavra nos documentos. É uma simplificação de toda diversidade de informações contidas na base de documentos sem o propósito de ser uma representação fiel do documento, mas oferecer a relação entre as palavras e os documentos a qual é suficiente para a maioria dos métodos de aprendizado de máquina (REZENDE, 2003).

Medidas de Proximidade

No modelo espaço vetorial, a similidade entre um documentos x e y pode ser calculada utilizando-se a medida Cosseno. Essa medida é definida pela correlação entre os vetores $\vec{x}$ e $\vec{y}$, a qual pode ser calculada pelo cosseno do ângulo entre esses vetores. Dados dois documentos $x = (x_1, x_1, \ldots, x_t)$ e $y = (y_1, y_1, \ldots, y_t)$, calcula-se:

$$cosseno(x, y) = \frac{\vec{x} \bullet \vec{y}}{|\vec{x}| \times |\vec{y}|} = \frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t} x_i^2} \times \sqrt{\sum_{i=1}^{t} y_i^2}} \tag{2.7}$$

Valores de Cosseno próximos a 0 indicam um ângulo próximo a 90º entre $\vec{x}$ e $\vec{y}$, ou seja, o documento x compartilha poucos termos com a consulta y, enquanto valores próximos a 1 indicam um ângulo próximo a 0º, ou seja, x e y compartilham termos e são

similares. Outra medida utilizada para medir a similaridade entre documentos é conhecida como Jaccard, a qual é definida por:

$$jaccard(x_i, x_j) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} , \tag{2.8}$$

onde f_{11} é o número de termos presentes em ambos os documentos, f_{01} é o número de termos ausentes em x_i e presentes em x_j e f_{10} é o numero de termos ausentes em x_j e presentes em x_i. Semelhante a Cosseno, Jaccard é uma medida de similaridade que retorna um valor no intervalo [0,1], sendo que valores próximo a 1 indicam similaridade máxima (MARCACINI; REZENDE, 2010; TAN; STEINBACH; KUMAR, 2005; FELDMAN; SANGER, 2006).

2.3 Segmentação Textual

A tarefa de segmentação automática de textos, ou segmentação textual consiste em dividir um texto em partes que contenham um significado relativamente independente. Em outras palavras, é identificar as posições nas quais há uma mudança significativa de assunto. É útil em aplicações que trabalham com textos sem indicações de quebras de assunto, ou seja, não apresentam seções ou capítulos, como transcrições automáticas de áudio, vídeos e grandes documentos que contêm vários assuntos como atas de reunião e notícias (AGGARWAL, 2018; BOKAEI; SAMETI; LIU, 2015; SAKAHARA; OKADA; NITTA, 2014; MISRA et al., 2009; EISENSTEIN; BARZILAY, 2008).

Pode ser usada para melhorar o acesso a informação solicitada por meio de uma consulta, onde é possível oferecer porções menores de texto mais relevantes ao invés de exibir um documento grande que pode conter informações menos pertinentes. Além disso, encontrar pontos onde o texto muda de assunto, pode ser útil como etapa de pré-processamento em aplicações voltadas ao entendimento do texto, principalmente em documentos longos (CHOI; WIEMER-HASTINGS; MOORE, 2001).

As técnicas de segmentação textual consideram um texto como uma sequência linear de unidades de informação que podem ser, por exemplo, cada termo presente no texto, os parágrafos ou as sentenças. Cada unidade de informação é um elemento do texto que não será dividido no processo de segmentação e cada ponto entre duas unidades é considerado um candidato a limite entre segmentos. Nesse sentido, um segmento pode ser visto como uma sucessão de unidades de informação que compartilham o mesmo assunto.

Nessa seção serão apresentados os algoritmos frequentemente referenciados na literatura com diferentes características. A seguir são detalhados os algoritmos baseados em coesão léxica, *TextTiling* e *C99*, o *BayesSeg* e *TextSeg* que trazem abordagens probabilísticas e o *MinCutSeg* baseado em particionamento de grafos.

2.3.1 Algoritmos de Segmentação Textual

Os primeiros trabalhos dessa área se apoiam na ideia de que a mudança de assunto em um texto é acompanhada de uma proporcional mudança de vocabulário. Essa ideia, chamada de coesão léxica, sugere que a distribuição das palavras é um forte indicador da estrutura do texto, e demonstrou-se que há uma estreita correlação entre quedas na coesão léxica em janelas de texto e a transição de assuntos (KOZIMA, 1993). Em seu trabalho, Kozima calculou a coesão léxica de uma janela de palavras usando *spreading activation* em uma rede semântica especialmente elaborada para o idioma Inglês. Contudo, a implementação de um algoritmo para outros domínios depende da construção de uma rede adequada.

O conceito de coesão léxica permite a aplicação da técnica de janelas deslizantes para encontrar os segmentos de um texto, em que se verifica a frequência dos termos em um fragmento do documento. Inicialmente, estabelece-se a partir do início do texto, um intervalo de t termos, chamado janela que em seguida é deslocada em passos de k termos adiante até o final do texto. A cada passo, analisa-se os termos contidos na janela (REYNAR, 1998).

TextTiling

A abordagem baseada em coesão léxica motivou a elaboração dos primeiros algoritmos para segmentação textual, entre eles o *TextTiling*. O *TextTiling* baseia-se na ideia de que um segmento pode ser identificado pela análise dos termos que o compõe. Inicialmente, o *TextTiling* recebe uma lista de candidatos a limite entre segmentos, usualmente finais de parágrafo ou finais de sentença. Utilizando a técnica de janelas deslizantes, para cada posição candidata são construídos 2 blocos, um contendo as sentenças que a precedem e outro com as que a sucedem. O tamanho desses blocos é um parâmetro a ser fornecido ao algoritmo e determina o tamanho mínimo de um segmento. Esse processo é ilustrado na Figura 2.

Em seguida, os blocos de texto são representados por vetores que contém as frequências de suas palavras. Diferente da proposta de Kozima, o *TextTiling* utiliza cosseno (Equação 2.7) como medida para a similaridade entre os blocos adjacentes. Um limite ou transição entre segmentos é identificado sempre que a similaridade entre as unidades que antecedem e precedem o ponto candidato cai abaixo de um limiar, indicando uma diminuição da similaridade entre os blocos adjacentes. Ou seja, identifica-se uma transição entre segmentos pelos vales na curva de dissimilaridades. Para cada final de sentença representada por c_i atribui-se uma profundidade dada por $(c_{i-1} - c_i) + (c_{i+1} - c_i)$ e será um limite entre segmentos caso a profundidade exceda $\overline{s} - \sigma$, onde $\overline{s}$ é a média da profundidade de todos os vales do documento e σ, o desvio padrão. Na Figura 3 é ilustrado a curva de dissimilaridade entre os blocos adjacentes.

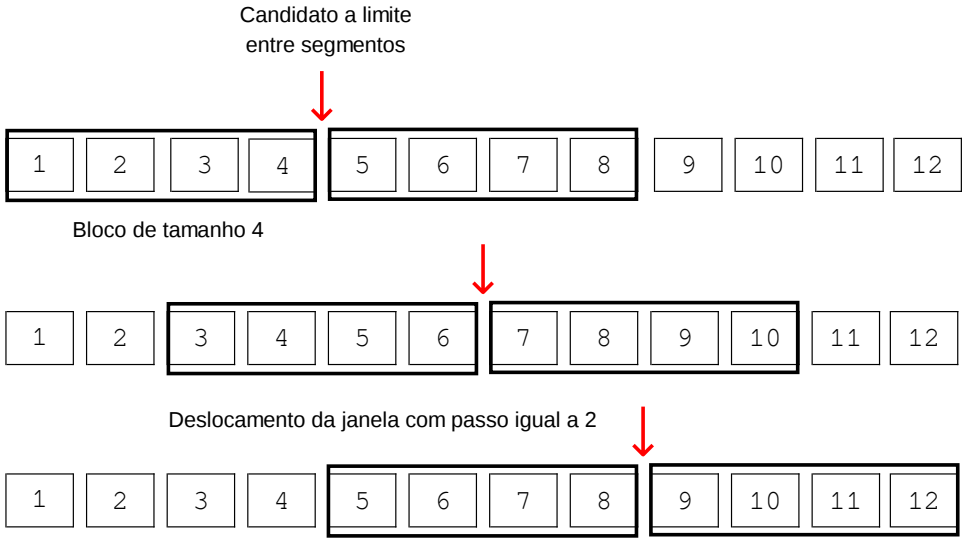


Figura 2 – Processo de deslocamento da janela deslizante. Os quadrados numerados representam as sentenças e os retângulos representam os blocos de texto a serem comparados. O deslocamento movimenta o candidato a limite e por consequência os blocos que o antecede e sucede.

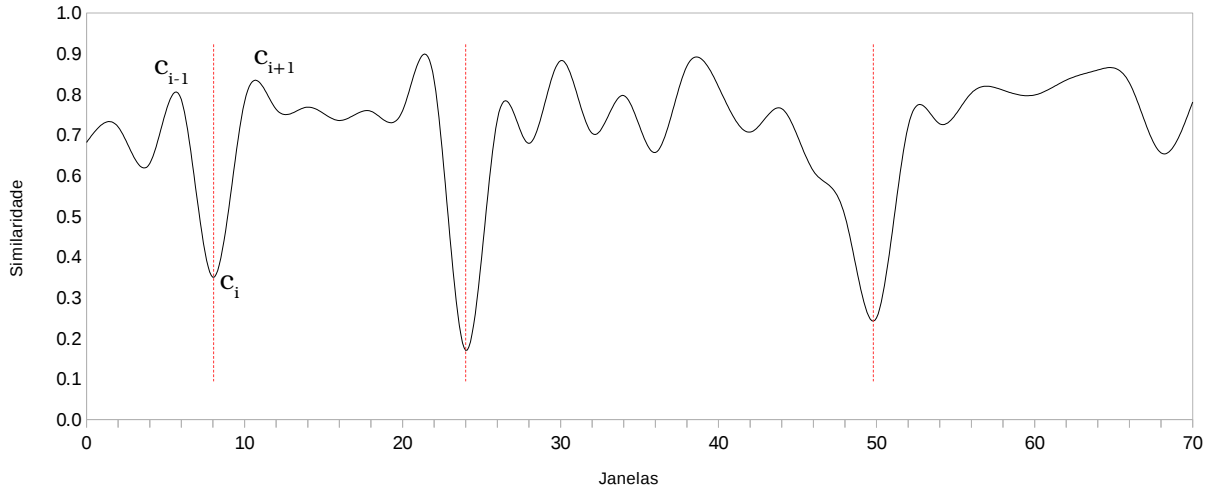


Figura 3 – Curva de dissimilaridades entre blocos de texto adjacentes. As linhas pontilhadas representam diminuições de similaridade que indicam limites entre segmentos.

O TextTiling apresenta como vantagens a facilidade de implementação, o que favorece o desenvolvimento de trabalhos similares (NAILI; CHAIBI; GHEZALA, 2016; BOKAEI; SAMETI; LIU, 2015; CHAIBI; NAILI; SAMMOUD, 2014; KERN; GRANITZER, 2009; GALLEY et al., 2003), e sua utilização como *base line* em outros trabalhos (CARDOSO; PARDO; TABOADA, 2017; DIAS; ALVES; LOPES, 2007). Por outro lado, algoritmos mais complexos, como os baseados em matrizes de similaridade, apresentam acurácia relativamente superior como apresentado em (CHOI, 2000; KERN; GRANITZER, 2009; MISRA et al., 2009).

C99

Outro algoritmo frequentemente referenciado na literatura é o C99 (CHOI, 2000) o qual é baseado em uma matriz de *ranking* das similaridades. A utilização de da coesão léxica pode não ser confiável para segmentos pequenos nessa abordagem, pois a ocorrência adicional de uma palavra pode causar certo impacto e alterar o cálculo da similaridade. Além disso, o estilo da escrita normalmente não é constante em todo o texto. Por exemplo, textos iniciais dedicados a introdução costumam apresentar menor coesão do que trechos dedicados a um tópico específico. Portanto, comparar a similaridade entre trechos de diferentes regiões não é apropriado. Devido a isso, as similaridades não podem ser comparadas em valores absolutos. Contorna-se esse problema fazendo uso de matrizes de similaridade para encontrar os segmentos de texto. Para isso, o *C99* constrói uma matriz que contém as similaridades de todas as unidades de informação (normalmente sentenças ou parágrafos).

Na Figura 4 é mostrado um exemplo de uma matriz de similaridade onde a intensidade do ponto(i, j) representa a similaridade entre as sentenças i e j. Observa-se que a matriz é simétrica, assim cada ponto na linha diagonal representa a similaridade quando $i = j$ (ou seja, com a mesma sentença) e revela quadrados com maior concentração de pontos ao longo da diagonal. A concentração de pontos ao longo da diagonal indica porções de texto com maior coesão léxica.

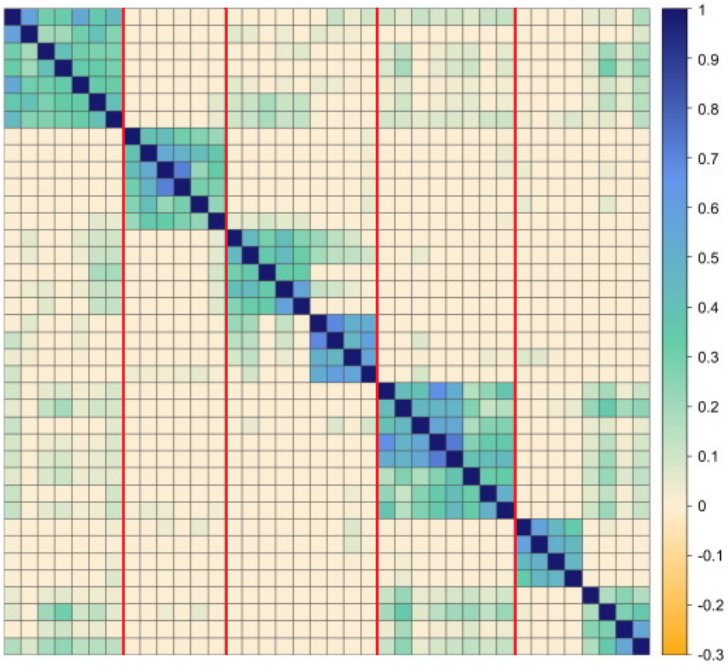


Figura 4 – *DotPlot* da similaridade entre sentenças onde as linha verticais representam segmentos reais (EISENSTEIN; BARZILAY, 2008).

Em seguida, cada valor na matriz de similaridade é substituído por seu *ranking* local. Para cada elemento da matriz, seu *ranking* será o número de elementos vizinhos com valor de similaridade menor que o seu. Assim, para cada elemento determina-se uma

região quadrada de tamanho l em que o elemento em questão será comparado com $l \times l - 1$ elementos vizinhos. Na Figura 5 é destacado um quadro 3 x 3 de uma matriz. Tomando como exemplo o elemento com valor $0,5$, a mesma posição na matriz de *rankings* terá o valor 4, pois esse é o número de vizinhos com valores inferiores a $0,5$ dentro do quadro analisado na matriz de similaridades. Da mesma forma, para o valor $0,2$ a matriz de *rankings* conterá o valor 1 na mesma posição. Após a construção da matriz de ranking obtêm-se um maior contraste entre os pontos, o que facilita a detecção de limites quando a queda de similaridade entre sentenças é mais sutil.

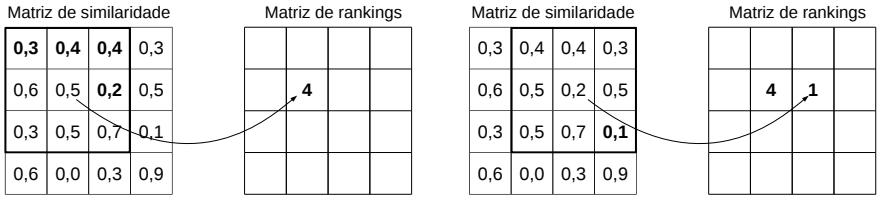


Figura 5 – Exemplo de construção de uma matriz de rankings.

Finalmente, com base na matriz de *ranking*, o C99 utiliza um método de agrupamento baseado no algoritmo *DotPloting* (REYNAR, 1998) que usa regiões com maior densidade em uma matriz de similaridades para determinar como os segmentos estão distribuídos. Um segmento é definido por duas sentenças c_i e c_j (respectivamente a primeira e última sentença do segmento) que representam uma região quadrada ao longo da diagonal da matriz. Seja $S = \{s1, ..., s_h\}$ a lista de h segmentos, som_b a somatória dos valores dos *rankings* de um segmento $s \in S$ e a_s a sua área. Então, a densidade é computada por: Calcula-se a densidade dessa região como mostrado na Equação 2.9.

$$Den = \frac{\sum_{s=1}^{h} som_s}{\sum_{s=1}^{m} a_b} \tag{2.9}$$

O processo inicia com um único segmento formado por todas as sentenças do documento e o divide recursivamente em m segmentos. Cada passo divide B no ponto (i, j) que maximiza Den (Equação 2.9). O processo se repte até atingir o número de segmentos desejados ou um limiar de similaridade.

TextSeg

Desenvolveu-se também abordagens probabilísticas para segmentação textual, por exemplo, o método proposto por (UTIYAMA; ISAHARA, 2001), aqui chamado de *TextSeg* encontra a segmentação por meio de um modelo estatístico. Dado um texto representado por um conjunto de palavras $W = \{w_1, w_2, \ldots, w_n\}$ e um conjunto de

segmentos $S = \{s_1, s_2, \ldots, s_h\}$ que segmenta W, a probabilidade da segmentação S é dada por:

$$P(S|W) = \frac{P(W|S)P(S)}{P(W)} \tag{2.10}$$

Com isso, é possível encontrar a sequência de segmentos mais provável $\hat{S} = argmax_S\ P(W|S)P(S)$. Nesse trabalho assume-se que os segmentos são estaticamente independentes entre si e as palavras nos segmentos são independentes dado o segmento que as contém. Essa simplificação permite decompor o termo $P(W|S)$ em um produtório de ocorrência de das palavras dado um segmento,

$$P(W|S) = \prod_{i=1}^{m} \prod_{j=1}^{n} P(w_j^i|S_i)\ , \tag{2.11}$$

onde $P(w_j^i|S_i)$ é a probabilidade da j-ésima palavra ocorrer no segmento S_i a qual é definida na Equação 2.12. Seja $f_i(w_j)$ a frequência da j-ésima palavra no i-ésimo segmento, n_i é o número de palavras em S_i e k é o número de palavas diferentes em W. Calcula-se:

$$P(w_j^i|S_i) = \frac{f_i(w_j) + 1}{n_i + k} \tag{2.12}$$

A suposição de independência entre segmentos e as palavras neles contidas, são é verificada no mundo real. Para segmentos muito pequenos a estimativa das probabilidades das palavras pode ser afetada. Além disso, o modelo não leva em conta a importância relativa das palavras (MALIOUTOV; BARZILAY, 2006).

BayesSeg

Os métodos baseados em coesão léxica que utilizam métricas como cosseno quantificam a similaridade entre sentenças baseando-se apenas na frequência das palavras, Essa abordagem, ignora certas características do texto que podem dar pistas sobre a estrutura do texto. Por exemplo, frases como "*Prosseguindo*", "*Dando continuidade*", "*Ao final da reunião*" podem ajudar a detectar o inicio ou final de segmento. A fim de aproveitar esses indicadores, pode-se usar um framework bayesiano que permite incorporar fontes externas ao modelo. O método *BayesSeg* (EISENSTEIN; BARZILAY, 2008) aborda a coesão léxica em um contexto bayesiano onde as palavas de um segmento surgem de um modelo de linguagem multinomial o qual é associado a um assunto. Essa abordagem é similar à métodos probabilísticos de extração de tópicos como o Latent Dirichlet Allocation (LDA) (BLEI; NG; JORDAN, 2003), com a diferença que ao invés de atribuir tópicos ocultos a cada palavra, esses são usados para segmentar o documento. Nesse sentido, detecta-se um limite entre sentenças quando a distribuição de tópicos entre elas for diferente. O *BayesSeg*

baseia-se na ideia que alguns termos são usados em tópicos específicos enquanto outros são neutros em relação aos tópicos do documento e são usados para expressar uma estrutura do documento, ou seja, as frases-pista vem de um único modelo generativo. A fim de refletir essa ideia, o modelo é adaptado para influenciar a probabilidade da sentença de ser uma final ou início de segmento conforme a presença de frases pista.

MinCut

O *MinCutSeg* (MALIOUTOV; BARZILAY, 2006) aborda a segmentação textual como um problema de particionamento de grafo, em que cada nó representa um sentença e os pesos das arestas representam a similaridade entre duas sentenças (Figura 6). Nessa abordagem, a segmentação textual corresponde ao particionamento do grafo que representa o texto.

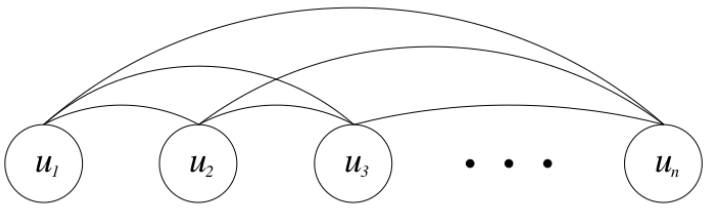


Figura 6 – Representação de texto baseada em grafo (MALIOUTOV; BARZILAY, 2006)

Essa abordagem é inspirada no trabalho de (SHI; MALIK, 2000) que propõe um critério para particionamento de grafos chamado *normalized-cut criterion* inicialmente desenvolvido para segmentação de imagens estáticas a qual foi aproveita a restrição de linearidade dos textos para segmentação textual.

Seja $G = V, E$ um grafo ponderado, unidimensional em que V é o conjunto de vértices que correspondem às sentenças e E é o conjunto de arestas que correspondem as similaridades entre as sentenças. Seja $sim(u, v)$ o valor de similaridade entre o par de vértices u e v. O *MinCutSeg* visa particionar G em dois grafos disjuntos A e B de modo a minimizar o corte definido pela somatória das arestas que ligam $u \in A$ à $v \in B$(Equação 2.13):

$$corte(A, B) = \sum_{u \in A, v \in B} sim(u, v) \tag{2.13}$$

Além de maximizar a diferença entre as partições A e B, é necessário que essas seja homogenias em relação a similaridade de suas sentenças, conforme requerimento definido por (SHI; MALIK, 2000) em que o valor do corte deve ser normalizado pelo volume das

partições dado por:

$$vol(A) = \sum_{u \in A, v \in V} sim(u, v) \tag{2.14}$$

Em seguida, define-se o critério de corte normalizado (NCorte) como o resultado da normalização do corte pelo volume, conforme mostrado na Equação 2.15.

$$NCorte(A, B) = \frac{corte(A, B)}{vol(A} + \frac{corte(A, B)}{vol(B)} \tag{2.15}$$

Uma vez que um texto normalmente é dividido em mais que dois segmentos, é necessário extender o modelo para atender a essa necessidade. Seja $A_1 \dots A_k$ uma partição e $V - A_k$ a diferença entre o grafo V e a partição k. O critério para múltiplos cortes normalizados é então estendido para:

$$NCorte_k(V) = \frac{corte(A_1, V - A_1)}{vol(A_1)} + \dots + \frac{corte(A_k, V - A_k)}{vol(A_k)} \tag{2.16}$$

A decomposição do modelo em uma somatória de termos individuais permite empregar técnicas de programação dinâmica para o problema de cortes multidirecionais em grafos. Mais detalhes da formulação dessa solução estão disponíveis em (MALIOUTOV; BARZILAY, 2006).

Embora o problema minimizar cortes normalizados em grafos seja um problema do tipo NP-Completo[3], no contexto de segmentação textual esse problema é restrito a manter a linearidade dos vértices. A segmentação linear em um grafo implica que todos os vértices entre as extremidades esquerda e direitas de uma partição pertencem à essa partição, consequentemente o espaço de soluções possíveis é reduzido o que permite a execução do algoritmo em tempo polinomial.

2.3.2 Medidas de Avaliação em Segmentação Textual

As medidas de avaliação tradicionais como precisão e revocação permitem medir o desempenho de modelos de Recuperação de Informação e Aprendizado de Máquina por meio da comparação dos valores produzidos pelo modelo com os valores observados em uma referência. Usa-se uma tabela, chamada matriz de confusão, para visualizar o desempenho de um algoritmo. Na Tabela 2 é apresentada uma matriz de confusão para duas classes (Positivo e Negativo) (AGGARWAL, 2018; MANNING; RAGHAVAN; SCHüTZE, 2008).

[3] NP-Completo configura um tipo de problema para o qual não se conhece uma solução determinística que possa ser computada em tempo polinomial. Papadimitriou provou que o problema de corte mínimo em grafos está incluso nessa categoria.

	Predição Positiva	Predição Negativa
Positivo real	VP (Verdadeiro Positivo)	FN (Falso Negativo)
Negativo real	FP (Falso Positivo)	VN (Verdadeiro Negativo)

Tabela 2 – Matriz de confusão.

No contexto de segmentação textual, um falso positivo é um limite identificado pelo algoritmo que não corresponde a nenhum limite na segmentação de referência, ou seja, o algoritmo indicou que em determinado ponto há uma quebra de segmento, mas na segmentação de referência, não há quebra no mesmo ponto. De maneira semelhante, um falso negativo é quando o algoritmo não identifica um limite existente na segmentação de referência, ou seja, em determinado ponto há, na segmentação de referência, um limite entre segmentos, contudo, o algoritmo não o identificou. Um verdadeiro positivo é um ponto no texto indicado pelo algoritmo e pela segmentação de referência como uma quebra de segmentos, ou seja, o algoritmo e a referência concordam que em determinado ponto há uma transição de assunto. Na avaliação de segmentadores, não há o conceito de verdadeiro negativo. Este seria um ponto no texto indicado pelo algoritmo e pela segmentação de referência onde não há uma quebra de segmentos, uma vez que os algoritmos apenas indicam onde há um limite, essa medida não é necessária (KAZANTSEVA; SZPAKOWICZ, 2012; BEEFERMAN; BERGER; LAFFERTY, 1999).

Com base nos valores da matriz de confusão, pode calcular a Precisão, a qual indica a proporção de limites corretamente identificados pelo algoritmo, ou seja, correspondem a um limite real na segmentação de referência. Porém, não diz nada sobre quantos limites reais existem. É calculada dividindo-se o número de limites identificados automaticamente pelo número de candidatos a limite (Equação 2.17).

$$Precisão_{seg} = \frac{VP}{VP + FP} \tag{2.17}$$

A revocação, é a proporção de limites verdadeiros que foram identificados pelo algoritmo. Porém não diz nada sobre quantos limites foram identificados incorretamente. É calculada dividindo-se o número de limites identificados automaticamente pelo número limites verdadeiros (Equação 2.18).

$$Revocação_{seg} = \frac{VP}{VP + FN} \tag{2.18}$$

Existe uma relação inversa entre precisão e revocação. Conforme o algoritmo aponta mais segmentos no texto, este tende a melhorar a revocação e ao mesmo tempo, reduzir a precisão. Esse problema de avaliação pode ser contornado utilizado a medida F^1 que é a

média harmônica entre precisão e revocação onde ambas tem o mesmo peso (Equação 2.19).

$$F^1{}_{seg} = \frac{2 \times Precisão \times Revocação}{Precisão + Revocação} \tag{2.19}$$

As medidas tradicionais baseiam-se na matriz de confusão a qual considera um verdadeiro positivo quando duas segmentações colocam o final de um segmento no mesmo ponto, sem considerar pequenas diferenças. Assim essas medidas podem ser mais adequadas quando necessita-se medir a eficiência do algoritmo com maior exatidão. Por consequência, essas medidas sempre penalizam o algoritmo quando um segmento não não coincide perfeitamente com a referência, pois computam apenas os erros do algoritmo quando se detecta falsos positivos ou falsos negativos, o que nesse contexto de segmentação textual pode não ser suficiente, dado a subjetividade da tarefa. Além dessas medidas, que consideram apenas se um segmento foi perfeitamente definido conforme uma referência, pode-se também considerar a distância entre o segmento extraído automaticamente e o segmento de referência (KERN; GRANITZER, 2009). Chama-se *near misses* o caso em que um limite identificado automaticamente não coincide exatamente com a referência, mas é necessário considerar a proximidade entre eles.

Na Figura 7 é apresentado um exemplo com duas segmentações extraídas automaticamente e uma referência. Em ambos os casos não há nenhum verdadeiro positivo, o que implica em zero para os valores de precisão, acurácia, e revocação, embora o resultado do algoritmo A possa ser considerado superior ao primeiro se levado em conta a proximidade dos limites.

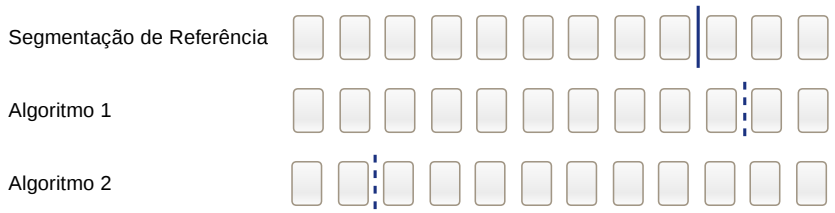


Figura 7 – Exemplos de *near missing* e falso positivo. Os blocos indicam uma unidade de informação e as linha verticais representam uma transição de assunto.

Considerando o conceito de *near misses*, algumas medidas de avaliação foram propostas. Proposta por (BEEFERMAN; BERGER; LAFFERTY, 1999), P_k atribui valores parciais a *near misses*, ou seja, limites sempre receberão um peso proporcional à sua proximidade, desde que dentro de um janela de tamanho k. Para isso, esse método move uma janela de tamanho k ao longo do texto. A cada passo verifica, na referência e no algoritmo, se as extremidades (a primeira e última sentença) da janela estão ou não dentro do mesmo segmento, então, penaliza o algoritmo caso este não concorde com a referência. Portanto, dado dois termos de distância k, P_k verifica se o algoritmo coloca os

termos no mesmo segmento ou em segmentos distintos e o penaliza caso não concorde com a referência. Dadas uma segmentação de referência ref e uma segmentação automática hyp, ambas com N sentenças, P_k é computada como:

$$P_k(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} (\delta_{ref}(i, i+k) \bar{\oplus} \delta_{hyp}(i, i+k)) \ , \tag{2.20}$$

na qual $\delta_S(i, j)$ é a função indicadora que retorna 1 se as sentenças c_i e c_j estão no mesmo segmento e 0 caso contrário, $\bar{\oplus}$ é o operador `XNOR` (ambos ou nenhum) que retorna 1 se ambos os argumentos forem iguais. O valor de k é calculado como a metade da média dos comprimentos dos segmentos reais. Como resultado, é retornada a dissimilaridade entre a segmentação calculada pela contagem de discrepâncias divida pela quantidade de segmentos analisados. Essa medida pode ser interpretada como a probabilidade de duas sentenças extraídas aleatoriamente pertencerem ao mesmo segmento.

WindowDiff (PEVZNER; HEARST, 2002) é uma medida alternativa à P_k. De maneira semelhante, move uma janela pelo texto e penaliza o algoritmo sempre que o número de limites proposto pelo algoritmo não coincidir com o número de limites esperados para aquela janela. Assim, o algoritmo é penalizado quando não concordar com a segmentação de referência quanto ao número de segmentos na janela. Mais formalmente, para cada intervalo k, compara o número de segmentos obtidos pela referência r_i com o obtido pelo algoritmo a_i e penaliza o algoritmo se $r_i \neq a_i$. Na Equação 2.21 é mostrada a definição de WindowDiff onde $b(i, i+k)$ representa o número de limites entre as sentenças i e $i+k$ e N, o total de sentenças no texto.

$$WindowDiff(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(ref_i - ref_{i+k}) - b(hyp_i - hyp_{i+k})| > 0) \tag{2.21}$$

Assim, *WindowDiff* consegue manter a sensibilidade a *near misses* e além disso, considerar o tamanho das janelas. A fim de melhor equilibrar o peso dos falsos positivos em relação a *near misses*, dobra-se a penalidade para falsos positivos, evitando-se a supervalorização dessa medida.

As medidas *WindowDiff* e P_k, consideram a quantidade e proximidade entre os limites, sendo mais tolerantes a pequenas imprecisões. Essa é uma característica desejável, visto que as segmentações de referência possuem diferenças consideráveis. *WindowDiff* equilibra melhor os falsos positivos em relação a *near misses*, ao passo que P_k os penaliza com peso maior. Isso significa que segmentadores melhores avaliados em P_k ajudam a selecionar as configurações que erram menos ao separar trechos de texto com o mesmo assunto, enquanto *WindowDiff* é mais tolerante nesse aspecto. De maneira geral, observa-se melhores resultados de *WindowDiff* quando os algoritmos aproximam a quantidade de

segmentos automáticos da quantidade de segmentos da referência. Por outro lado, P_k avalia melhor as configurações que retornam menos segmentos. Contudo, não é possível definir um valor adequado, uma vez que os segmentadores humanos frequentemente apontam segmentações diferentes.

2.3.3 Anotação de Segmentos

A avaliação de segmentadores frequentemente requer uma segmentação de referência. Essa referência deve refletir um segmentação real sendo confiável para apoiar a avaliação da qualidade de técnicas de segmentação.

A construção de um *corpus* anotado demanda tempo e disponibilidade de anotadores humanos, o que a torna uma tarefa relativamente custosa. Assim, é necessário seguir procedimentos que assegurem que a tarefa seja concluída com o exito esperado que o resultado produzido seja válido, confiável e consistente para fins de pesquisas científicas. Para isso, (HOVY; LAVID, 2010) propuseram uma metodologia para anotação em *corpus* que pode ser resumida em sete passos: (1) escolha do evento a ser anotado, (2) seleção do *corpus*, (3) selecionar e treinar os anotadores, (4) especificar o processo de anotação, (5) modelar uma interface para anotação, (6) escolher e aplicar medidas de avaliação e (7) disponibilizar e manter o produto.

Um dos primeiros trabalhos a produzir um *corpus* com anotações de segmentos foi (HEARST, 1997) no qual um *corpus* constituído por doze artigos de revistas foram anotados por sete técnicos pesquisadores. Cada artigo continha entre 1.800 e 2500 palavras. O autor considerou um limite entre segmentos real onde pelo menos três anotadores marcavam uma transição de tópico. No trabalho de (KAZANTSEVA; SZPAKOWICZ, 2012) utilizou-se um livro ficcional contendo vinte capítulos que foi segmentado por seis alunos de graduação que além de marcar os pontos de transição entre segmentos, forneceram um descrição breve sobre cada segmento identificado.

Outros trabalhos abordaram *corpus* compostos pela transcrição de audios. Por exemplo, (PASSONNEAU; LITMAN, 1997) transcreveu vinte narrativas sobre um filme que foi segmentada e anotada por sete voluntários. Cada narrativa, continha cerca de 13.500 palavras. Os anotadores não receberam nenhum treinamento formal para a tarefa, mas apenas foram solicitados a usar suas noções de comunicabilidade para identificar as mudanças de tópicos. No trabalho de (GALLEY et al., 2003) investigou-se a transcrição de um conjunto de vinte e cinco reuniões obtidas do *ICSI Meeting corpus* (JANIN et al., 2003) em que pelo menos três anotadores analisaram os pontos onde ocorreram trocas da pessoa que fala e apontaram como sendo ou não uma mudança de assunto.

Nesses trabalhos utilizou-se os anotadores como juízes para produzir uma referência em que decidiu-se sobre cada candidato a limite entre segmentos por meio da opinião da

maioria. Além desses trabalhos, outros se valeram de segmentações produzidas artificialmente. Por exemplo, (CHOI, 2000) produziu um *corpus* formado por 700 documentos. As referências foram geradas pela concatenação de sentenças extraídas de documentos diferentes. De maneira semelhante, (CHAIBI; NAILI; SAMMOUD, 2014) utilizou a concatenação de artigos de noticias para produzir os documentos. Os autores consideram um limite real o ponto que divide dois artigos originais.

Os trabalhos citados anteriormente utilizaram procedimentos diferentes para produzir segmentações de referência para seus trabalhos. Como já citado, (HOVY; LAVID, 2010) propôs que o processo de anotação em *corpus* pode der sintetizado e dividido em sete passos.

Escolha do *corpus*

A criação de *corpus* raramente é restrita a um único propósito. O material original deve ser preferencialmente constituído de documentos disponíveis livremente à comunidade, a fim de facilitar a comparação, extensão e avaliação de trabalhos futuros. Devido a diversidade linguística de diferentes domínios e gêneros de textos, a escolha dos documentos de amostra deve procurar ser representativa ao domínio a ser abordado. O *corpus* é considerado representativo quando o assunto a abordado na amostra tem correspondência com a interpretação do público geral desse domínio.

Escolha da teoria a ser explicada

A anotação deve ajudar a explicar uma teoria, ou seja, fornecer informações úteis à sua compreensão. Essa teoria irá guiar a especificação do processo de anotação, quais informações deseja-se extrair e como interpretá-las. Quanto mais complexa for a teoria ser explicada, mais complexa será a tarefa de anotação bem como as instruções que os anotadores deverão seguir. Além disso, deve-se estabelecer de início o nível de detalhamento necessário. A complexidade da teoria e nível de detalhamento impactam na condução da anotação e da estabilidade da anotação.

Selecionar e treinar os anotadores

O treinamento e o nível de conhecimento dos anotadores ainda é uma questão em aberto. Alguns pesquisadores afirmam que estes devem ser especialistas no domínio do *corpus*. Outros afirmam que pessoas adequadamente treinadas podem produzir resultados satisfatórios. Considerando a necessidade de treinamento, tem-se a subjetividade das tarefas que dificulta a elaboração de instruções precisas. Tarefas que permitem a especificação de procedimentos que levam em conta a possibilidade de diferentes casos e variáveis, põem em dúvida a necessidade da criação de um *corpus* anotado. Por outro lado, a ausência

de treinamento implica que as anotações terão como base o conhecimento prévio dos anotadores e sua preconcepção a cerca do domínio o que diminui o nível de concordância entre os anotadores e dificulta a replicação de outros trabalhos.

Especificar o procedimento de anotação

Alguns processos de anotações podem levar longos períodos, criando a necessidade de dividir a tarefa em fases. Nesses casos, frequentemente os anotadores fazem reuniões periódicas a fim de relatar eventuais problemas. Em caso de baixa concordância, pode-se abrir espaço para discussão a fim de que encontrar um ponto de convergência, a qual é chamada de fase de "reconciliação" que embora recomendada, em alguns casos pode ocasionar um enviesamento dos resultados, outra estratégia para diminuir uma eventual baixa concordância é solicitar que os anotadores marquem o nível de certeza sobre as anotações.

Modelar uma interface para anotação

Um *software* com interface amigável, além de facilitar o trabalho, evita erros durante o processo. O ganho em tempo e a melhoria na qualidade dos resultados justifica a criação de uma interface. Exemplos *softwares* para anotação na área de Processamento de Linguagem Natural e Bioinformática podem ser encontrados em (GRUENSTEIN; NIEKRASZ; PURVER, 2007).

Escolher e aplicar medidas de avaliação

Quando observa-se baixa concordância entre os anotadores, entende-se que há uma falha no processo de anotação ou na teoria a ser explicada, o que implica que o dados produzidos não servem para a fins de pesquisa ou aplicações práticas. A medida dessa concordância deve determinar a confiabilidade dos resultados. A medida mais utilizada em Processamento de Linguagem Natural é o coeficiente *kappa* (CARLETTA, 1996) que retorna um valor no intervalo de 0 até 1, onde 1 significa uma concordância perfeita e 0 que não houve concordância. Seja $P(A)$ a proporção de vezes que os anotadores concordam e $P(E)$ a proporção de concordância esperada ao acaso. O cálculo de *kappa* é dado por:

$$kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.22}$$

Essa medida, apresenta como limitação a entrada de apenas dois casos. Como alternativa, a medida conhecida como *Fleiss's k* (SHROUT; FLEISS, 1979) pode ser utilizada quando há mais que dois anotadores, porém restringe-se a anotações com apenas duas categorias. Na avaliação de segmentadores, as medidas P_k (Equação 2.20) e *WindowDiff*

(Equação 2.21) podem ser utilizadas, uma vez que são medidas de similaridade, como visto em (KAZANTSEVA; SZPAKOWICZ, 2012; CARDOSO; PARDO; TABOADA, 2017).

Disponibilizar e manter o produto

Uma vez criado, o *corpus* anotado deve ser disponibilizado para uso em outros trabalhos. Recomenda-se fornecer o *corpus* original além dos resultados obtidos, observando-se desde o início e ao longo do tempo a propriedade e eventuais licenças sob o *corpus* original.

2.4 Modelos de Extração de Tópicos

Os modelos de extração de tópicos fornecem uma estratégia que visa encontrar nas relações entre documentos, padrões latentes que sejam significativos para o entendimento dessas relações (WEI, 2007). Tais modelos podem ranquear um conjunto de termos importantes para um ou mais assuntos, bem como ranquear documentos por sua relevância para determinado tema (FALEIROS, 2016; YI; ALLAN, 2009). Atualmente, destacam-se os modelos probabilísticos de extração de tópicos como LDA (BLEI; NG; JORDAN, 2003) e PLSA (HOFMANN, 1999). São abordagens amplamente utilizadas (ZHU et al., 2012) e frequentemente referenciadas em trabalhos que buscam extrair conhecimento e organizar bases textuais (AGGARWAL, 2018; O'CALLAGHAN et al., 2015; STEYVERS; GRIFFITHS, 2007). Nesse trabalho, a expressão *tópico* é usada para designar um assunto considerando que o mesmo foi extraído por meio de técnicas automáticas, ficando a expressão *assunto* utilizada como seu teor popular.

O processo de extração de tópicos atribui um peso a cada documento-tópico e uma relação termo-tópico que pode representar a probabilidade de ocorrência de um termo em um documento dado que o tópico está presente. A partir dessas representações, é possível agrupar documentos que compartilham o mesmo tópico bem como os termos que melhor descrevem o tópico (AGGARWAL, 2018). Com isso, obtém-se uma organização da coleção de documentos que favorece técnicas para navegação e consulta à coleção de documentos (MARCACINI; REZENDE, 2010). Além disso, essas abordagens de extração de tópicos fornecem a construção de novos atributos que representam os principais tópicos ou assuntos identificados na coleção de documentos, sendo uma oportunidade de incorporar conhecimento de domínio aos dados (GUYON; ELISSEEFF, 2003).

Para extrair esses tópicos, algumas técnicas foram propostas. Em termos de metodologia, a maioria dos trabalho enquadram-se em duas principais categorias, os modelos não-probabilísticos e os modelos probabilísticos.

2.4.1 Modelos Não Probabilísticos

Nos modelos não-probabilísticos a matriz documento-termo é projetada em um espaço com menor dimensionalidade chamado *Latent Semantic Space*. Seja $d \in D = \{d_1, \cdots, d_n\}$ o vetor que representa a coleção de documentos, $t \in T = \{t_1, \cdots, t_m\}$ seus termos distintos e $z \in Z = \{z_1, \cdots, z_k\}$ seus tópicos. Esses métodos aprendem decompondo a matriz documento-termo W, em duas matrizes Z e A, tal que a resultante de ZA seja uma aproximação da matriz W original. Mais formalmente tem-se:

$$Z \cdot A = \hat{W} \approx W \tag{2.23}$$

Sendo n o número de termos, m o número de documentos da coleção, k a quantidade de tópicos a serem extraídos, a matriz A corresponde a matriz documento-tópico e possui dimensão $k \times n$. Z corresponde a matriz termo-tópico e possui dimensão $m \times k$. Uma vez que $k \ll n, m$, então A e Z são menores que a matriz de entrada, o que resulta em uma versão comprimida da matriz original, pois $k \cdot n + m \cdot k \ll n \cdot m$. Ao final, obtém-se uma representação documento-tópico que atribui um peso para cada tópico em cada documento da coleção e uma representação termo-tópico que representa a probabilidade de ocorrência de um termo em um documento dado que o tópico está presente no documento.

Nesse sentido, o *Latente Semantic Indexing* (LSA) (DEERWESTER et al., 1990) usa a técnica chamada *Singular Value Decomposition* (SVD) para encontrar padrões no relacionamento entre assuntos e termos em uma coleção de texto não estruturada. Entretanto, esse método não fornece uma interpretação para elementos com valores negativos (DEERWESTER et al., 1990) (CHENG et al., 2013).

Outro modelo popular é o *Non-Negative Matrix Factorization* (NMF) (LEE; SEUNG, 1999). as matrizes resultantes não possuem elementos negativos, permitindo uma interpretação mais intuitiva de seus valores. O processo de fatoração proporciona o agrupamento das colunas da matriz W o que possibilita, a propriedade *clustering* a esse modelo.

2.4.2 Modelos Probabilísticos

Os modelos probabilísticos consideram os documentos como uma mistura de tópicos e um tópico como uma distribuição probabilística sobre os termos. O processo de elaboração do documento a partir desses tópicos é chamado de processo generativo ou modelo generativo, o qual é desconhecido, porém, pode ser estimado com base nos termos presentes no documento, também chamados de variáveis observáveis. Assim, o processo de extração de tópicos consiste em estimar o modelo generativo que deu origem aos documentos de uma coleção.

O PLSA (HOFMANN, 1999) foi um dos primeiros a estender o modelo LSA e formalizar a extração de tópicos probabilísticos. De maneira similar ao LSA, esse modelo decompõe uma matriz esparsa a fim de reduzir a dimensionalidade. O PLSA cria um modelo estatístico chamado *aspect model* que associa os tópicos às variáveis observáveis atribuindo probabilidades às ligações entre os tópicos e os documentos e entre as palavras e os tópicos. Assim, cada documento pode ser representado como a probabilidade de um tópico estar presente, $P(z|d)$. E a probabilidade de um termo ocorrer dado que um tópico esta presente, $P(t|z)$. Em comparação ao LSA, é considerado um método mais robusto por proporcionar uma interpretação probabilística. Por outro lado, esse modelo apresenta desvantagens como o número de parâmetros do modelo que cresce linearmente com o número de documentos da coleção, o que pode ocasionar *overfitting*.

O LDA (BLEI; NG; JORDAN, 2003) estende o modelo PLSA incorporando um modelo generativo onde cada tópico obedece à distribuição multivariada de *Dirichlet* o que o torna menos propenso ao *overfitting* e capaz de inferir tópicos a documentos ainda não observados. É referenciado na literatura como estado-da-arte sobre modelos probabilísticos de extração de tópicos e influencia uma grande quantidade de trabalhos, tornando-se base para novos modelos.

O LDA utiliza a distribuição de Dirichlet para amostrar a distribuição dos tópicos. O modelo aloca os tópicos latentes que são distribuídos conforme a distribuição de Dirichlet. A função de densidade dessa distribuição é dada por:

$$Dir(z, \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} z_k^{\alpha_k - 1}, \tag{2.24}$$

onde $z = (z_1, \ldots, z_K)$ e $\alpha = (\alpha_1, \ldots, \alpha_K)$ são variáveis K-dimensionais e $B(\alpha)$ é a função Beta dada por:

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \tag{2.25}$$

No modelo LDA, o processo de geração de palavras é o resultado da amostragem da Dirichlet é usado para atribuir as palavras de diferentes tópicos e que irão compor os documentos. Os tópicos são entendidos como distribuições probabilísticas sobre um vocabulário de palavras. Enquanto que os documentos, surgem da escolha aleatória das palavras presentes a uma distribuição de tópicos. O processo gerador de um documento d_j no modelo LDA pode ser detalhado como a seguir:

1. Crie as distribuições $\phi_k \backsim Dir(\phi_k, \beta)$ para cada tópico k;

2. Crie uma distribuição $\theta_j \backsim Dir(theta_j, \alpha)$ para d_j;

3. Escolha elemento i a compor o documento d_j,

   a) Atribua aleatoriamente um tópico $z_{j,i} \backsim Multinomial(\theta_j)$;

   b) Atribua aleatoriamente uma palavra $w_{j,i}$ com probabilidade $p(w_{j,i}|\phi_{z_{j,i}})$.

O processo gerador representa as distribuições por meio de duas variáveis. A variável n-dimensional ϕ em que n é o número fixo de palavras do vocabulário, e a variável K-dimensional θ. Essas variáveis são geradas por Dir com seus respectivos parâmetros β e α.

Após as variáveis ϕ_k e θ_j serem inicializadas, gera-se por fim o documento d_j. Assim cada documento é associado a múltiplos tópicos com proporções distintas e cada palavra do documento é obtida de um tópico específico que foi anteriormente obtido a partir da distribuição de tópicos do documento. Isso permite ao modelo LDA atribuir, para cada documento, múltiplos tópicos com proporções distintas (BLEI, 2012; FALEIROS, 2016).

Outro modelo empregado na tarefa de extração de tópicos é o *K-Means*. O *K-Means* é um dos algoritmos de agrupamento particional mais usados, tendo se tornado bastante popular em tarefas de recuperação de informação (MANNING; RAGHAVAN; SCHüTZE, 2008). O agrupamento de documentos de uma coleção se dá inicialmente pela codificação de cada documento em um vetores em que seus elementos o representam. Em seguida usa-se medidas de distância para medir as similaridades entre os documentos da coleção. A ideia por traz do *K-Means* é definir k documentos aleatórios para representar os grupos, chamados centroides, e atribuir cada documento ao centroide mais próximo. Os centroides são recalculados os grupos rearranjados iterativamente até que não haja mudanças significativas. O funcionamento do *K-Means* para agrupar um coleção de documentos em k grupos pode ser definido com a seguir:

1. Selecione k objetos iniciais, chamados centroides;

2. Atribua cada documento ao grupo mais próximo;

3. Calcule o novo centroide para cada grupo;

4. Repita os passos 2 e 3 até que os grupos não sejam alterados significativamente.

O critério de parada do *K-Means* é determinado por soma dos erros quadráticos, definida como:

$$E = \sum_{i=1}^{k} \sum_{x \in G_i} |x - l_i|^2 \ , \tag{2.26}$$

na qual E é a somatória dos erros quadráticos calculados para cada documento da coleção, x é o vetor de atributos que representa o documento, e li é o centroide que representa o grupo G_i.

No contexto de extração de tópicos o *K-Means* pode selecionar termos contidos na coleção a fim de descrever os grupos. Esses termos, chamados descritores podem ser selecionados pela frequência dos termos no centroide. Esses termos, chamados descritores podem ser atribuídos a cada grupo com base na frequência dos termos dos centroides. Verifica-se quais termos são mais comuns no documento que representa o centroide e que menos frequentes nos documentos mais distantes ao centroide (GURUNG; WAGH, 2017; BUI et al., 2017; ROSSI; REZENDE, 2011; SANTOS; CARVALHO; REZENDE, 2010; MANNING; RAGHAVAN; SCHüTZE, 2008).

Os modelos de extração de tópicos foram inicialmente propostos para utilização em Mineração de Texto onde são empregados na redução de dimensionalidade, extração de informações em textos, bem como na organização e recuperação de documentos, sendo utilizados para mensurar a relevância de um termo ou conjunto de termos para determinado assunto ou documento. Visto a popularidade nessas tarefas e flexibilidade dos modelos, logo notou-se sua utilidade em outros tipos de dados com atributos discretos como na genética, grafos e imagens.

2.5 Trabalhos Relacionados

Nesta seção são apresentados os principais trabalhos relacionados a proposta dessa dissertação. Os trabalhos a seguir abordam a Segmentação Textual, Extração de Tópicos e Recuperação de Informação e a intersecção entre as técnicas.

Nos últimos anos, a crescente disponibilidade de documentos e a necessidade de gerenciá-los de forma eficiente, incentivou a pesquisa por técnicas de aprendizado de máquina para agrupar e classificar coleções de documentos longos. A maioria dessas pesquisas consideram que um documento pertence a único tópico. Essa premissa é verdadeira em muitos casos, como postagens em redes sociais, *reviews* de produtos e e-mails. Contudo, isso raramente é válido para documentos longos que por vezes possuem mais de um tema. Um dos primeiros trabalhos a agrupar documentos compostos por múltiplos temas é conhecido como *Suffix Tree Clustering* (STC) proposto por (ZAMIR; ETZIONI, 1998). O STC usa frases para calcular a similaridades e criar grupos sobrepostos de documentos, em que um documento pode pertencer a mais de um grupo.

Outro trabalho pioneiro nesse sentido foi proposto em (MASAO; KôITI, 2000). Esse trabalho foca na sumarização de múltiplos documentos sobre múltiplos tópicos. Os autores propuseram um método baseado em *spreading activation* em uma base de documentos anotados semanticamente. O método extrai partes dos documentos consideradas

importantes para criar uma rede que os relaciona. Essa abordagem foi capaz de identificar sentenças relacionadas bem como os documentos. Contudo essa abordagem não utilizada métodos de segmentação textual, considerando cada sentença como nós da rede. Além disso, vale-se de rotulação manual para criar relações entre as entidades.

O algoritmo *MultiSeg*, proposto em (JEONG; TITOV, 2010) visa descobrir descobrir ligações entre segmentos semanticamente relacionados. Os autores apresentam um modelo Bayesiano não paramétrico para inferir relação e agrupar segmentos de documentos. Essa abordagem se propõe a ajudar usuário a encontrar segmentos relacionados e detectar informações complementares à pesquisa inicial. Segundo os autores, essas relações ainda podem revelar tendências em fontes de dados.

Ainda nesse contexto (NGUYEN, 2011) cria uma Estrutura Hierárquica de Tópicos (*Hierarchical Structure of Topic-information*) – HST utilizando uma metodologia baseada em segmentos para agrupar segmentos de documentos e identificar os grupos por meio de uma frase que reflete o conteúdo dos segmentos pertences ao grupo. Inicialmente o texto de cada documento é dividido em parte topicamente coerentes gerando uma coleção de segmentos. Em seguida, uma hierarquia de tópicos é construída por meio um método de agrupamento aglomerativo hierárquico. Por fim, cada grupo recebe um título, o qual é gerado por meio de algoritmos de sumarização e extração de palavras-chave.

Em seu trabalho, (TAGARELLI; KARYPIS, 2013) consideram como documento multi-topical aqueles que têm múltiplas intenções comunicativas que refletem diferentes necessidades de informação. Exemplos de documentos multi-topicais podem ser encontrados em discussões em forums, páginas de notícias, discursos e transcrições de conversas e reuniões. Nesse contexto, Tagarelli e Karypis, (2013) propuseram um *framework* de agrupamento para documentos multi-topicais. Inicialmente os documentos são modelados como um conjunto de segmentos de acordo com seus tópicos. Em seguida os segmentos são agrupados e os documentos originais são classificados. Por fim, um classificador foi induzido a partir dos grupos de segmentos. Os autores aplicaram sua metodologia a 3 *datasets*: 1) RCV1 com 6.588 documentos; 2) PubMed com 3.687 documentos; 3) CaseLaw com 2.550 documentos. O trabalho apresenta uma metodologia que utiliza segmentos de um determinado documento para facilitar a atribuição deste a mais de um grupo (onde cada grupo contém segmentos relevantes a um tópico). Para isso, utiliza os parágrafos do texto como estrutura para divisão de um documento, dispensado algoritmos de segmentação textual. Como principal contribuição, fornece uma analise sobre algoritmos de agrupamento de documentos com sobreposição (ZHAO; KARYPIS, 2004a; ZHAO; KARYPIS, 2004b; DHILLON; MODHA, 2001) e propõe variantes deste para adequação ao problema estudado.

3 Sistema de Recuperação de Informação em Documentos Multi-temáticos

O sistema proposto tem como objetivo recuperar informações em uma coleção de documentos em que cada documento contém assuntos diversos e relativamente independentes entre si. Esse sistema deve identificar os assuntos de cada documento e disponibilizá-los de forma que o usuário consiga consultar a coleção de documentos e obter todo o histórico de ocorrências de um determinado tema de forma que possa identificar onde esse tema foi mencionado, bem como se houve uma decisão relacionada ao tema. Os documentos constituídos por diversos assuntos, são aqui chamados de documentos multi-temáticos, em contraste com aqueles cujo assunto central é bem definido e constante ao logo do texto.

A proposta original deste trabalho contempla funcionalidades de classificação para identificar automaticamente o tipo de ocorrência onde um assunto é mencionado, o qual pode ser classificado como uma decisão, informe ou simplesmente uma menção ao assunto. Contudo essas funcionalidades configuram trabalhos futuros para continuação do sistema como concebido inicialmente. Assim, esse trabalho de mestrado está focado na segmentação de atas de reunião, no agrupamento desses segmentos em tópicos e na recuperação de trechos de atas relacionados ao assunto da pesquisa.

Esse capítulo está organizado da seguinte forma: primeiramente, na Seção 3.1 é apresentada uma visão geral do sistema proposto, seu funcionamento e como as técnicas de segmentação textual e extração de tópicos são empregadas para gerar uma base de dados que concentra as informações necessárias para identificar e agrupar os diversos assuntos distribuídos na coleção de documentos. Ainda nessa seção, é apresentada a utilização das técnicas de recuperação de informação empregadas para entregar os segmentos de acordo com a consulta do usuário bem como permitir a exploração de segmentos relacionados ao mesmo tema, os quais originalmente estão distribuídos na coleção de documentos. Na Seção 3.2 é apresentada a aplicação do sistema proposto utilizando como base de dados uma coleção de atas de reunião. Uma vez que as atas configuram um *corpus* com documentos multi-temáticos, o sistema as utilizará e as técnicas empregadas serão analisadas. Ainda nessa seção, será apresentada a preparação das atas, bem como a descrição dos algoritmos utilizados e suas configurações.

3.1 Sistema Proposto

Essa Seção apresenta as etapas de desenvolvimento do sistema de recuperação de atas proposto, bem como o seu funcionamento geral, desde a preparação dos documentos

até a entrega dos resultados ao usuário. Na Figura 8 é mostrada a visão geral do sistema com suas principais entradas e saídas. Inicialmente o sistema recebe um conjunto inicial de documentos por meio do Módulo de preparação/manutenção cuja função é processar e manter esses textos para gerar uma base de dados interna que codifica os textos extraídos com seus respectivos tópicos. Essa base de dados fica disponível ao Módulo de Consulta recebe a entrada pela qual o usuário expressa um assunto de interesse. Em seguida, os trechos de texto que fazem menção ao esse assunto são exibidos ao usuário.

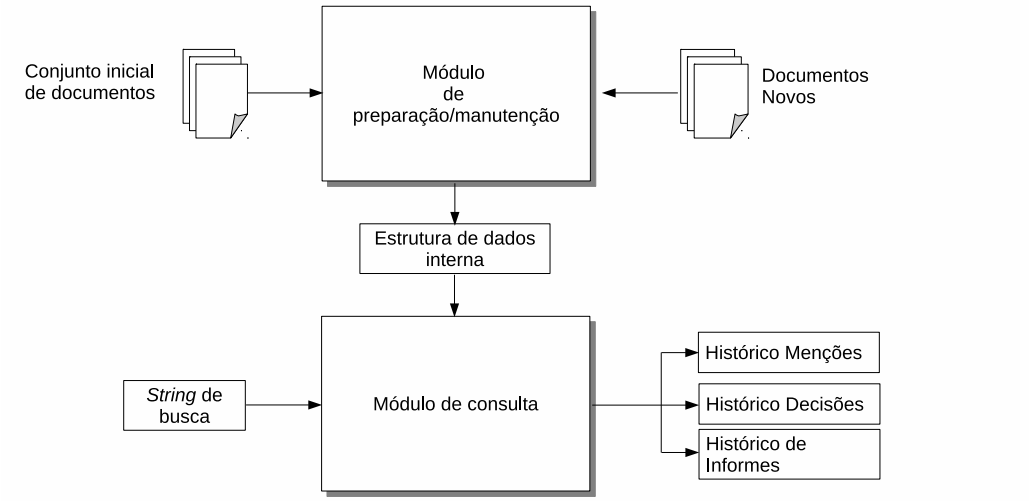


Figura 8 – Visão geral do sistema.

3.1.1 Módulo de Preparação e Manutenção

O módulo de preparação e manutenção tem como função principal manter uma base de dados necessária para os processos de Recuperação de Informação e Aprendizado de Máquina. Primeiramente, esse módulo deve receber um conjunto inicial de documentos e em seguida organizá-lo por meio das técnicas de segmentação textual e extração de tópicos, de forma que o conjunto resultante formado por segmentos de documentos e os dados obtidos pelo extrator de tópicos constitua um novo *corpus* estruturado e mais adequado às técnicas de Recuperação de Informação e Aprendizado de Máquina. Além disso, considera-se o crescimento da bases de documentos, assim, o sistema deve receber novos documentos a medida que são gerados. A seguir são detalhadas as etapas do Módulo de preparação e manutenção desde a preparação dos documento até a entrega da estrutura de dados interna ao Módulo de Consulta.

3.1.1.1 Preparação dos Documentos

Os documentos são normalmente armazenadas em arquivos binários do tipo *pdf*, *doc*, *docx* ou *odt*. A partir desses arquivos, os textos devem ser pré-processados e estruturadas para que possam ser aplicados métodos de Mineração de Texto e Recuperação de Informação.

Para isso, o texto plano[1], é extraído e passa por processos de transformação que incluem remoção de elementos considerados menos significativos e a identificação de sentenças. Esse processo é ilustrado na Figura 9 e descrito a seguir.

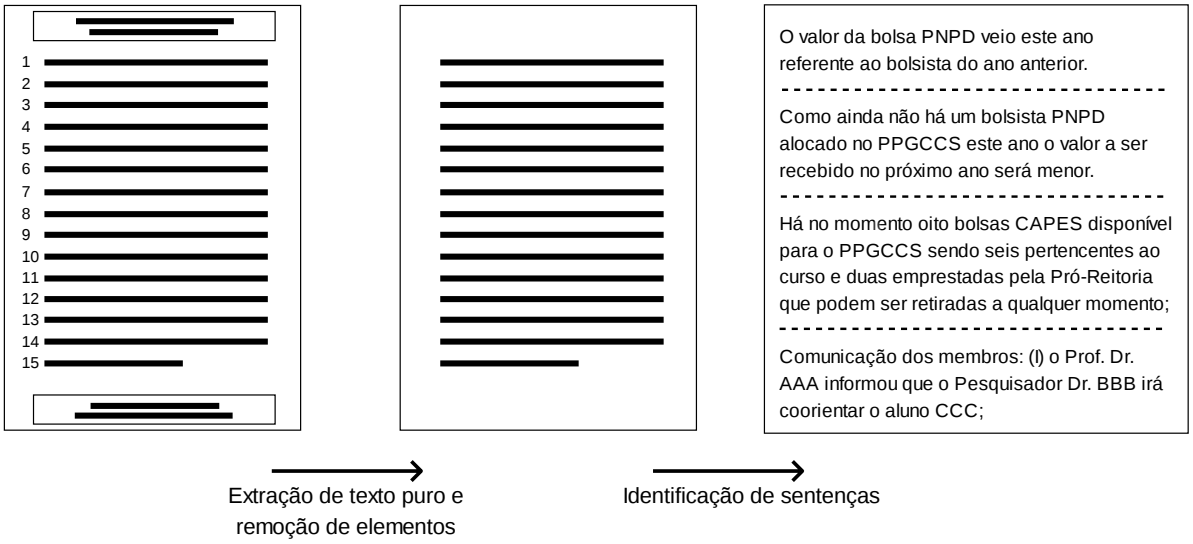


Figura 9 – Etapa de preparação de um documento que inclui da extração de texto puro, remoção de elementos menos significativos e a identificação de sentenças.

É frequente que alguns documentos, como as atas de reunião, contenham trechos que podem ser considerados pouco informativos e descartados durante o pré-processamento, como cabeçalhos e rodapés que se misturam aos tópicos tratados na reunião, podendo ser inseridos no meio de um tópico prejudicando tanto os algoritmos de Mineração de Texto e Recuperação de Informação, quanto a leitura do texto pelo usuário. Um cabeçalho é a porção de texto que inicia cada página do documento e, de forma semelhante, um rodapé e a porção que as encerra. Detecta-se os cabeçalhos e os rodapés sempre que há uma repetição das primeiras e últimas palavras do documento, comummente no início e final de cada página. Uma vez identificadas, essas repetições são consideradas irrelevantes e removidas.

Nesse trabalho considera-se as sentenças as menores unidades de informação a serem processadas pelos algoritmos de segmentação, por tanto, estas devem ser identificadas. Ao considerar intuitivamente que uma sentença seja uma sequência de palavras entre sinais de pontuação como “.”, “!” e “?”, alguns erros poderiam ocorrer quando esses tiverem outra função dentro do texto como em abreviações[2], endereços de internet e datas. Outro

[1] Texto plano é uma sequência de characteres sem informações sobre estilos, como tamanho, cor, sendo universalmente legível por computadores.

[2] As abreviações são identificadas por meio de uma lista com 234 abreviações conhecidas.

problema seriam frases curtas com poucas palavras e que não expressam um conceito completo, mas parte dele. Devido ao estilo de pontuação desses documentos, como encerrar sentenças usando um ";" e inserção de linhas extras, foram usadas as regras especiais para identificação de finais de sentença. No Algoritmo 1 é mostrado como cada *token*[3] é identificado como final de sentença.

Algoritmo 1: Identificação de finais de sentença.

Entrada : Texto
Saída : Texto com identificações de finais de sentença

para todo *token, marcá-lo como final de sentença se:*
    Terminar com um `!`
    Terminar com um `.` e não for uma abreviação
    Terminar em `.?;` e:
        For seguido de uma quebra de parágrafo ou tabulação
        O próximo *token* iniciar com `({["'`
        O próximo *token* iniciar com letra maiúscula
        O penúltimo caracter for `)}]"'`
fim

3.1.1.2 Pré-Processamento dos Documentos

Nessa etapa os documentos são pré-processados individualmente antes de serem recebidos pelos algoritmos de segmentação e extração de tópicos. Inicialmente, cada texto passa por um processo de transformação em que as letras são convertidas em caixa baixa e elimina-se sinais de pontuação, numerais e termos menores que três caracteres. Em seguida remove-se os termos que não contribuem para a etapa de segmentação, as quais são chamadas de *stop words*. Para identificá-las usa-se uma lista de 438 palavras conhecidas do idioma português, como preposições, artigos e verbos de estado e adicionalmente uma lista contendo palavras muito frequentes e pouco significantes desse domínio com como *"universidade"*, *"computação"* e nomes de pessoas as quais formam listadas manualmente. Em seguida, extrai-se o radical de cada palavra por meio da técnica *stemming* implementada no algoritmo de Porter (PORTER, 1997). A etapa de pré-processamento cria uma estrutura de dados intermediária, da qual foram removidos os elementos mencionados. Essa estrutura é utilizada para representar os textos, contudo, a estrutura de dados interna preserva os elementos removidos para apresentação adequada ao usuário.

3.1.1.3 Segmentação

Como já mencionado, uma ata registra a sucessão de assuntos discutidos em uma reunião, porém apresenta-se com poucas quebras de parágrafo e sem marcações de estrutura,

[3] Nesse trabalho um *token* é qualquer conjunto de símbolos entre sinais não visuais como espaços, quebras de linha e tabulações.

como capítulos, seções ou quaisquer indicações sobre o assunto do texto. Portanto, faz-se necessário descobrir quando há uma mudança de assunto no texto da ata. Para essa tarefa, as técnicas de segmentação de texto recebem uma lista de sentenças, a qual considera cada ponto entre duas sentenças como candidato a limite entre segmentos, ou seja, um ponto onde há transição entre assuntos (BOKAEI; SAMETI; LIU, 2016). Como resultado, para cada documento tem-se um conjunto de segmentos com um assunto relativamente independente que constituem o texto do documento original, contudo, sem indicações do teor dos segmentos. Em seguida, esses segmentos são processados por um extrator de tópicos que irá agrupá-los por assuntos e eleger palavras para descrever esses tópicos.

3.1.1.4 Extração de Tópicos

Após a segmentação da coleção e identificação das transições de assuntos em cada documento da coleção, o sistema inicia a etapa de extração de padrões por meio das técnicas de extração de tópicos, ou seja, uma vez identificado onde há uma mudança de assunto, o próximo passo é obter o tópico ao qual o assunto pertence.

O resultado do processo de extração de tópicos é a representação dos documentos e seus tópicos em uma matriz documento-tópico que atribui um peso a cada tópico para cada documento e uma matriz termo-tópico que pode representar a probabilidade de ocorrência do termo quando um tópico ocorre em um documento ou a frequência esperada desse termo. Nesse sistema, essas representações são utilizadas para melhorar as tarefas de recuperação de informação e agrupamento dos documentos. O agrupamento por tópicos e seus descritores são utilizados para ajudar o usuário a analisar e identificar os segmentos conforme sua consulta, bem como encontrar resultados similares. Após a etapa de extração de tópicos aplicada a um conjunto de segmentos, cria-se uma base de dados derivada do *corpus* original, a qual é descrita na seção a seguir.

3.1.2 Estrutura de Dados Interna

A estrutura de dados interna é o resultado dos processos de segmentação e extração de tópicos. Constituída por: (1) documentos originais para referência, (2) segmentos de texto contendo um assunto relativamente independente, (3) matrizes documento-tópico e termo-tópico. Na Figura 10 é apresentado a visão geral da estrutura de dados interna onde ilustra-se os arquivos da coleção de documentos (D) que contêm assuntos diversos (representados pelos círculos, quadrados e triângulos) ficando disponíveis para visualização e fonte original das informações.

A partir do texto extraído dos documentos originais é gerado o conjunto de segmentos de texto (S), armazenados em arquivos de texto plano que contém os textos em formato legível a serem exibidos ao usuário pelo Módulo de Consulta. Para que se

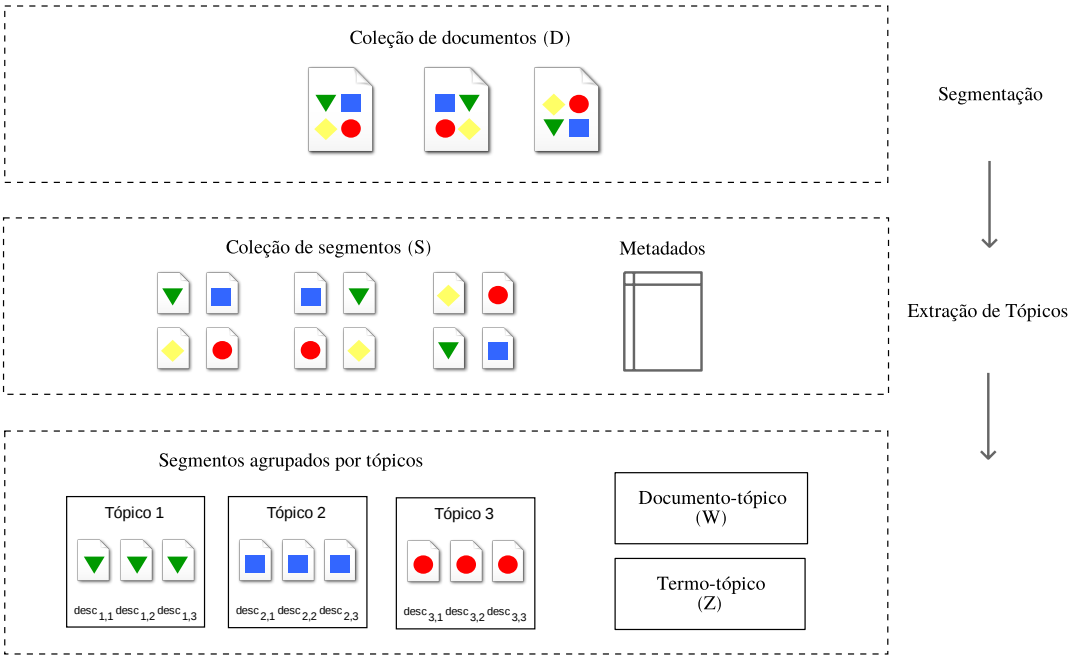


Figura 10 – Visão geral da estrutura de dados interna e seu processo de geração.

conheça o arquivo que deu origem a cada segmento, o sistema gera metadados que mantém a ligação entre os documentos e seus segmentos.

Os segmentos são então tratados como sub-documentos pelo sistema em que a partir deles o extrator de tópicos irá construir as matrizes documento-tópico (W) e termo-tópico (Z). Uma vez construídas, essas matrizes são armazenadas em arquivos e usadas para agrupar os segmentos com o mesmo tópico bem como encontrar os melhores termos da coleção para descrever cada tópico.

3.1.3 Módulo de Consulta

Uma vez que a Estrutura de Dados Interna contém os assuntos abordados na coleção de documentos e o trecho onde se encontram, tem-se um *corpus* derivado do original o qual é melhor organizado e acrescido de informação. A estrutura de dados gerada pelos processos anteriores é então utilizada pelo Módulo de Consulta, cuja responsabilidade é receber a *string* de consulta do usuário, resgatar os dados desejados e apresentá-los em ordem de relevância, dando condições para o usuário acessar os segmentos encontrados, bem como os documentos originais.

3.1.3.1 Ranqueamento

Os dados contidos na estrutura de dados interna devem ser acessados por meio de consultas do usuário. Deseja-se que somente os segmentos relevantes sejam exibidos e ordenados por sua similaridade com a consulta. Nesse trabalho, a seleção dos segmentos

é feita por meio buscas em um espaço vetorial (já apresentado na seção 2.1.2) onde cada tópico é representado pelos descritores obtidos no processo de extração de tópicos. Essa técnica permite utilizar as variáveis latentes, e fazer com que termos diferentes mas com mesmo significado (polissemia) sejam relacionados e considerados no processo de busca. Inicialmente, cria-se um *ranking* dos tópicos considerando a similaridade entre seus descritores e a consulta. Em seguida, seleciona-se os tópicos cujos descritores possuem maior relevância com a consulta do usuário e exibe-se os segmentos atribuídos ao primeiro tópico do *ranking*, ficando os demais disponíveis para exploração pelo usuário. Os segmentos de cada tópico são ordenados de acordo com o peso atribuído à relação entre o segmento e tópico. Como resultado, tem-se grupos de segmentos (tópicos) ordenados pela similaridade com a consulta, pelos quais se exibe os segmentos ordenados por sua relevância ao tópico.

3.1.3.2 Interface

O sistema deve dar acesso ao usuário para que busque e explore os dados concentrados na estrutura de dados interna, permitindo encontrar resultados relevantes à consulta bem como explorar a base de dados utilizando a organização dos resultados em grupos de segmentos com tópicos relacionados. Deve permitir ao usuário identificar resultados relativante suficientes para compreensão do conteúdo, evitando a leitura de documentos inteiros. Portanto, os textos apresentados devem ser suficientes para compreensão do assunto mencionado, sem necessidade de visualizar o documento original.

A interface desenvolvida para este trabalho foi construída com a finalidade de analisar os resultados de diferentes técnicas de segmentação textual e extração de tópicos. Melhorias nas funcionalidades e usabilidade da interface finais constituem trabalhos futuros relacionados a área de experiência de usuário.

A visualização é divida em 3 painéis principais: (1) campo de busca, onde se insere palavras chave para consulta para um assunto; (2) tópicos encontrados na base de dados; (3) segmentos atribuídos ao tópico selecionado. Na Figura 11 é apresenta a tela principal do sistema onde observa-se à esquerda os segmentos agrupados por tópicos. Os tópicos são representados por um componente para visualização em árvore e identificados por seus descritores. Uma pasta quando expandida revela os segmentos agrupados no tópico que representa. À direita, como resultado da consulta, são apresentados os segmentos selecionados pela relevância com a *string* de busca do usuário.

O usuário pode reconhecer um tópico por meio dos descritores que ajudam a identificar o assunto pelo qual os segmentos estão relacionados. A qualidade dos descritores e dos grupos extraídos está relacionada ás técnicas de segmentação textual e extração de tópicos empregadas. Nesse sistema, como forma de análise é possível obter resultados de três modelos de extração de tópicos, *K-Means*, *LDA* e *PLSA*, bem como configurá-los manualmente. Após selecionado um tópico, o sistema apresenta o texto contido em cada

segmento a ele atribuído e um *link* para visualizar o arquivo original, como referência. Além disso, ao selecionar um segmento, o sistema destaca outros tópicos aos quais um segmento selecionado está atribuído, permitindo uma forma de exploração à base dados que aproveita o agrupamento de segmentos por assunto.

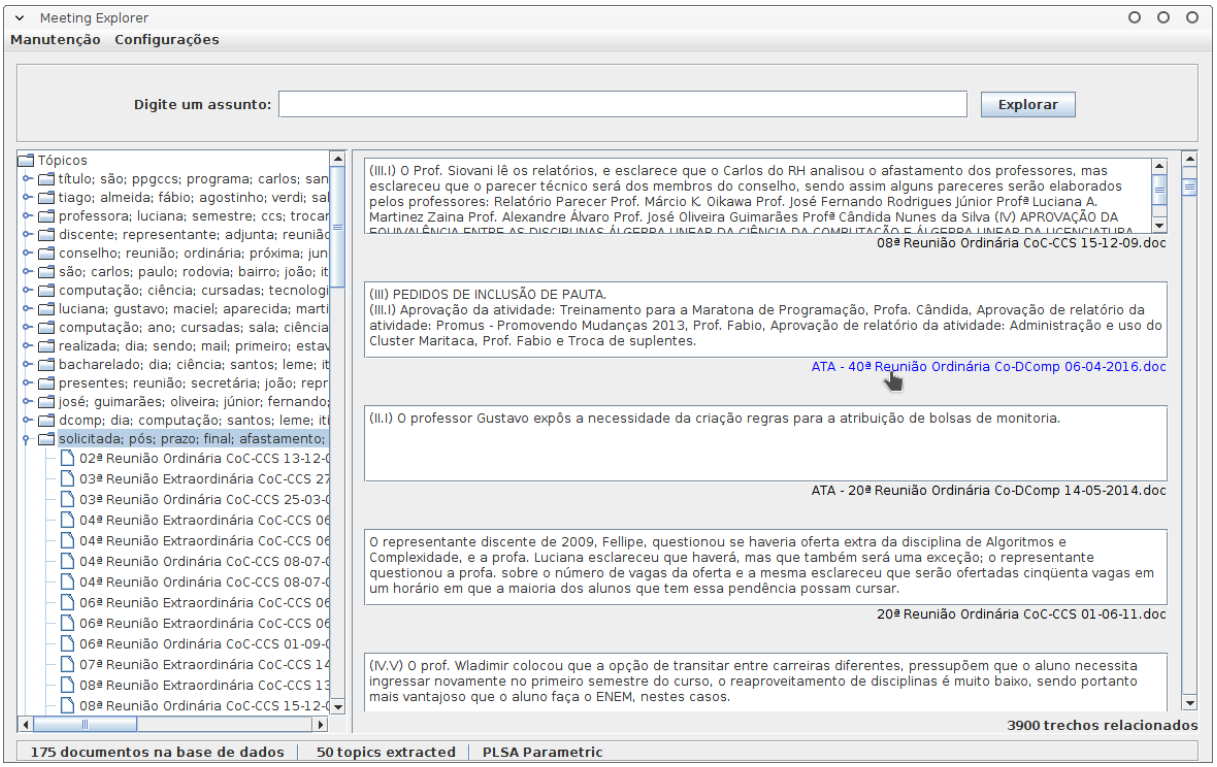


Figura 11 – Tela principal do sistema.

Como parte da proposta, o sistema oferece como opção a ordenação cronológica dos resultados a fim de apresentar cada resultado dentro de um histórico de menções. Dessa forma o usuário tem acesso uma interface que lhe fornece uma visão temporal das menções ao tema pesquisado. Outras funcionalidades podem ser desejáveis para um sistema voltado a usuários finais, como alternativas à buscas tracionais por palavras-chave, visto que os descritores são uma forma de representação que resume e relaciona o conteúdo dos segmentos, contudo buscas por termos exatos podem ser complementares.

3.2 Análise de um *Corpus* de Atas de Reunião Utilizando Ferramentas do Sistema

O foco principal deste projeto de mestrado é a exploração e recuperação de informação em atas de reunião. As atas de reunião, em geral, apresentam como característica textos relativamente curtos, em comparação com outros documentos como notícias, artigos, sites da *web*; o estilo de escrita formal em que o redator evita repetições de temos e conceitos em benefício da estética do texto; Multiplicidade de assuntos contidos em uma mesma

ata, na qual é difícil determinar um assunto central, mas diversos assuntos independentes que foram tratados durante a reunião. A literatura pesquisada apresenta poucos trabalhos voltados à essas caraterísticas, sobre tudo para o idioma português. Assim, escolheu-se um *corpus* de atas de reunião com propósito principal de contribuir com ferramentas e conhecimentos nesse sentido.

A seguir, será descrito o conjunto de atas de reunião utilizado como base de dados e os resultados obtidos pela aplicação das técnicas são observados para compreensão e análise do *corpus* estudado.

3.2.1 Composição do *Corpus*

O *corpus* abordado nesse trabalho foi formado por atas de reunião coletadas da Universidade Federal de São Carlos - Campus Sorocaba. Coletou-se 175 atas públicas das quais são 66 do Conselho do Departamento de Computação, sendo 55 referentes a reuniões ordinárias e 11 extraordinárias; 73 do Conselho do Curso de Bacharelado em Ciência da Computação, sendo 42 referentes a reuniões ordinárias e 31 extraordinárias e 36 da Comissão do Curso de Pós-Graduação em Ciência da Computação, sendo 31 referentes a reuniões ordinárias e 5 extraordinárias. As referentes a reuniões ordinárias têm em média 827 *tokens* enquanto as extraordinárias têm 667 *tokens* em média.

As atas de reunião diferem dos textos comumente estudados em outros trabalhos em alguns pontos. Frequentemente atas de reunião têm a característica de apresentar um texto com poucas quebras de parágrafo e sem marcações de estrutura, como capítulos, seções ou quaisquer indicações sobre o tema do texto. Além disso, possuem estilo de escrita bastante sucinto, em que o redator evita repetições de palavras em favor da estética do texto. O estilo de escrita formal mais compacto, pode dificultar processos de mineração de texto (CHOI; WIEMER-HASTINGS; MOORE, 2001).

3.2.2 Exploração e Observação do *Corpus*

Como retorno, diferentes modelos de extração de tópicos apresentam resultados distintos em relação aos descritores e segmentos atribuídos a cada tópico. Na Tabela 3 é apresentado um resumo dos tópicos extraídos do *corpus* por cada modelo. Os dados foram gerados configurando cada modelo para extrair um total de 70 tópicos da base da coleção, onde se observa 3 descritores extraídos para cada tópico. Os tópicos estão ordenados pela quantidade de segmentos atribuídos dos quais são exibidos os 45 primeiros tópicos. Os resultados completos dessa tabela, com 70 tópicos e 5 descritores podem ser vistos no Apêndice C.

K-Means		**LDA**		**PLSA**	
Descritores	*#Seg*	*Descritores*	*#Seg*	*Descritores*	*#Seg*
dia; realizada; chamada;	116	disciplinas; cursadas; fichas;	107	docentes; presidente; dia;	76
informado; compra; ofício;	106	colocar; deve; poderia;	94	disciplinas; álgebra; linear;	75
computação; conselho; aprovado;	102	docentes; presidente; técnica;	91	computação; acordo; levada;	62
docentes; técnica; administrativo;	72	dia; aprovado; aprovação;	85	aprovado; aprovação; unanimidade;	57
representante; discente; presidente;	55	representante; técnica; administrativo;	79	representante; discente; piccoli;	51
cursadas; conselho; coordenação;	45	conselho; junto; assina;	69	técnica; administrativo; representante;	45
aprovado; aprovação; atividades;	44	seguintes; chamada; conselho;	67	comunicação; presidência; informado;	38
computação; cursadas; conselho;	37	presença; realizada; cidade;	55	bacharelado; coordenação; cursadas;	37
disciplinas; cursadas; libras;	36	presidência; comunicação; informado;	54	afastamento; aprovado; aprovação;	35
professores; colocar; regras;	30	atividades; extensão; relatórios;	52	dia; ordem; anterior;	35
pedido; informado; substituí;	30	havendo; lavra; iniciou;	46	discente; representante; lúcio;	34
aprovado; trocar; pedido;	28	verba; compra; pagamento;	46	informado; compra; ofício;	34
dia; ordem; concurso;	27	afastamento; aprovado; aprovação;	44	presidente; presentes; lavra;	30
representante; administrativo; técnica;	26	coordenação; deliberar; restrito;	39	dia; seguintes; presidente;	28
afastamento; aprovado; referentes;	26	discente; representante; presidente;	34	relatórios; aprovado; lido;	28
extensão; atividades; coordenadores;	25	computador; tópicos; disciplinas;	30	dia; realizada; gestão;	27
compra; informado; verba;	23	disciplinas; calculada; diferentes;	27	dcomp; semestre; calendário;	26
aprovado; conselho; orientada;	22	gestão; conhecimento; conselho;	26	título; suplente; computação;	25
dia; ordem; aprovação;	22	processo; semestre; seletivo;	26	fichas; caracterização; obrigatório;	20
presidente; secretária; associado;	20	laboratório; máquina; técnica;	22	chamada; terceiro; dia;	19
unanimidade; aprovado; conselho;	19	aprovado; defesa; pedido;	17	dia; realizada; estado;	17
foram; aprovado; lido;	19	conselho; cursadas; sala;	16	cursadas; disciplinas; coordenação;	17
comunicação; presidência; presidente;	18	concurso; dados; bancos;	15	explicou; enviar; aprovação;	17
processo; seletivo; semestre;	18	pauta; inclusão; pedido;	14	extensão; atividades; coordenadores;	17
secretária; representante; presentes;	17	condicionado; informado; compra;	13	computação; cursadas; conselho;	16
fichas; caracterização; disciplinas;	16	discussão; decidido; regras;	12	computador; sistema; software;	16
aprovação; aprovado; política;	16	próxima; trazido; tomadas;	12	atividades; extensão; processo;	16
computação; teoria; paralela;	13	aprovado; aprovação; referentes;	10	pedido; deve; informado;	16
candidatos; concurso; lista;	12	pedido; atendida; compra;	10	cursadas; recurso; dcomp;	15
semestre; conceito; fronteiras;	12	implantação; serviços; horária;	10	orientada; prazo; meses;	15
aprovação; realizada; laboratório;	11	votação; votaram; equipe;	9	dia; conceito; laboratório;	14
redigida; lavra; presidente;	10	foram; material; estado;	8	extensão; programa; coordenadores;	14
deve; normalizado; assunto;	10	foram; conselho; aprovado;	7	projeto; comissão; esclarecido;	13
havendo; legal; número;	10	site; informado; dcomp;	7	ficou; novo; colocar;	13
realizada; pagamento; apresentação;	9	deve; laboratório; aprovado;	6	planos; ensino; foram;	13
aprovação; anterior; máquina;	9	learning; the; and;	6	provas; candidatos; presidente;	12
pauta; inclusão; pedido;	8	projeto; extensão; mudanças;	6	professores; cursadas; justificativa;	12
presentes; lavra; junto;	8	aprovado; dcomp; proposta;	3	área; concurso; problemas;	12
presidente; docentes; dia;	7	sugeriu; mail; poderia;	2	valor; compra; empenho;	12
ausência; justificativa; solicitação;	7	informado; aprovado; comissão;	1	graduação; pós; min;	11
técnica; administrativo; docentes;	7	informado; aprovado; dia;	1	vaga; transferência; foram;	11
presidência; comunicações; iniciou;	6	aprovado; informática; informado;	1	bancos; dados; aprovado;	11
comunicou; comunicação; conselho;	6	deve; lista; informado;	1	demanda; compra; pedido;	10
dados; bancos; ccs;	6	informado; deve; aprovado;	1	verba; cursadas; pagamento;	10
informática; sociedade; docentes;	6	deve; informática; conselho;	1	laboratório; manutenção; suplente;	10

Tabela 3 – Descritores e número de segmentos no tópico selecionado pelos Algoritmos

De maneira geral, os retornos do sistema apresentados na Tabela 3 oferecem uma perspectiva, segundo os modelos utilizados, dos principais assuntos abordados no *corpus*. Por exemplo, o termo *"aprovado"* e *"aprovação"* aparecem como descritores de vários dos tópicos mais numerosos, o que indica que grande parte dos segmentos, abordam assuntos relacionados a aprovações pelos membros da reunião. De forma semelhante, as frequências de termos como *"compra"* e *"verba"*, podem mostrar a importância desses termos para o ambiente onde se deram as reuniões.

Observou-se também que alguns tópicos concentram segmentos considerados pouco relevantes em termos de conteúdo, como a primeira parte introdutória da ata, onde se registra informações como data, local, membros e departamento. Esses segmentos são identificados com os termos *"dia; realizada; chamada;"* pelo K-Means com 116 segmentos; *"seguintes; chamada; conselho;"* pelo LDA com 67 segmentos e *"dia; realizada; gestão;"* e *"dia; seguintes; presidente;"* pelo PLSA com 28 e 27 segmentos. Vale salientar que os segmentos possuem texto similar, porém não idênticos e que os cabeçalhos e rodapés não estão presentes uma vez que foram removidos na etapa de preparação dos documentos. Como parte da proposta desse trabalho, esse agrupamento ajuda identificar textos com pouca relevância em termos de assuntos abordados na reunião. Os resultados dos sistema são melhores analisados no capítulos 4 e 5 onde as técnicas de segmentação textual e extração de tópicos são avaliadas no contexto das atas de reunião.

Os dados obtidos pela aplicação das técnicas permitem analisar o *corpus* pela distribuição dos tópicos ao longo da coleção de documentos identificando os assuntos em cada segmento de ata individualmente, gerando assim uma perspectiva ampla dos assuntos contidos na coleção de documentos. Além disso, essa metodologia pode dar uma visão da distribuição dos tópicos em cada um dos documentos. Na Figura 12 é exibido graficamente a distribuição de 6 tópicos extraídos de uma ata da coleção.

A ata exibida, foi segmentada utilizando algoritmo *BayesSeg* e os tópicos da coleção foram extraídos com o *K-Means*. Como já mencionado, as primeiras sentenças referem-se a introdução e apresentação da própria reunião e seus membros, as quais o extrator atribuiu a um grupo com 116 segmentos com os descritores *"dia; realizada; chamada; estado; conselho;"*. De forma semelhante, a região da ata reservada à assinatura dos membros foi atribuída a um grupo com 55 segmentos identificados pelos termos *"representante; discente; presidente; secretária; turma;"*.

Figura 12 – Distribuição de tópicos em uma ata real. Cada tópico é representado por uma região colorida. Abaixo estão os descritores identificados pela cor do respectivo tópico. Os nomes de pessoas foram ocultados por não expressarem significado nesse trabalho.

3.3 Considerações Finais

Nessa seção foi apresentado o sistema desenvolvido desde sua proposta original, as principais técnicas que o compõe, bem como sua aplicação em um *corpus* formado por atas de reunião. O sistema apresentado tem como escopo inicial a análise das técnicas de segmentação textual e modelos de extração de tópicos com documentos formados por múltiplos assuntos sem meta informações.

O sistema mostrou-se capaz de criar e manter uma base de dados estruturada a partir de documentos textuais não estruturados e utilizá-la para incorporar e recuperar

informações úteis. Usando técnicas de Recuperação de Informação foi possível localizar grupos e segmentos relevantes às consultas bem como estabelecer rankings com resultados mais pertinentes.

Os agrupamentos e descritores atribuídos aos segmentos permitem ao usuário visualizar os principais assuntos contidos na coleção de documentos, além de perceber relações latentes entre esses assuntos. As estrutura criada fornece uma representação que pode descrever o conteúdo dos segmentos das atas para exploração do usuário, bem como para técnicas de Recuperação de Informação. Mais análises voltadas aos resultados do sistema são discutidas no Capítulo 4 e no Capítulo 5. Trabalhos futuros relacionados ao sistema e à metodologia empregada são abordados no Capítulo 6.

4 Avaliação dos Segmentadores

Nesse capítulo é apresentada uma análise dos algoritmos de segmentação textual com objetivo de iniciar uma discussão sobre a potencialidade das técnicas a serem utilizadas em um *corpus* constituído de atas de reuniões conforme já mencionado na Seção 3.2.1. Foram escolhidos algoritmos comumente utilizados na literatura com diferentes características. Selecionou-se os algoritmos baseados em coesão léxica, *TextTiling* e *C99*, o *BayesSeg* e *TextSeg* por trazerem abordagens probabilísticas e o *MinCutSeg* o qual é baseado em particionamento de grafos. Além disso, incluiu-se também um algoritmo que simplesmente atribui um segmento a cada sentença, chamado aqui de *PseudoSeg*, para servir como um *baseline*.

A avaliação objetiva de um segmentador automático de textos exige uma referência, isto é, um conjunto de textos com os limites entre os segmentos conhecidos. Essa referência, deve ser confiável, sendo uma segmentação legítima que é capaz de dividir o texto em porções relativamente independentes, ou seja, uma segmentação ideal. Os resultados dos algoritmos foram avaliados por sua conformidade com uma segmentação de referência construída com base na metodologia de anotações em segmentos proposta em (HOVY; LAVID, 2010). Em seguida escolheu-se o modelo que apresenta melhores resultados para ser utilizado em conjunto com técnicas de extração de tópicos. Por meio de um experimento avaliou-se a performance de ambas as técnicas junto a profissionais com afinidade com atas de reunião. Os anotadores, forneceram suas percepções em relação aos resultados do segmentador empregado nesse trabalho. Os dados obtidos dos experimentos serviram de base para as análises dos algoritmos e de sua aplicação no contexto das atas de reuniões. Optou-se por criar um *corpus* de atas de reunião anotadas a fim de produzir-se uma segmentação de referência para avaliações dos algoritmos, bem como sua publicação e utilização em outros trabalhos voltados a esse domínio. Detalhes da criação da segmentação de referência utilizada nesse trabalho podem ser vistos na Seção 4.1.

Nas próximas seções o processo de anotação e criação da segmentação de referência será detalhado bem como serão avaliados os segmentadores abordados nesse trabalho. Inicialmente é descrita a preparação do *copus* e em seguida são mostradas as configurações utilizadas pelos algoritmos, bem como os critérios considerados para avaliá-los. Por fim, os resultados são apresentados e discutidos.

4.1 Preparação de um *Corpus* de Referência

A preparação desse *corpus* teve como guia a metodologia introduzida em (HOVY; LAVID, 2010) a qual indica sete passos para obter um *corpus* anotado. Essa metodologia

está descrita na da Seção 2.3.3 e explicada em (CARDOSO; PARDO; TABOADA, 2017).

O processo de anotação iniciou-se com a seleção de 12 atas de reuniões do departamento de computação da Universidade Federal de São Carlos - Campus Sorocaba, as quais são disponibilizadas publicamente[1]. Foram coletadas 6 atas de reuniões da Comissão do Curso de Pós-Graduação em Ciência da Computação e 6 atas de reuniões do Conselho do Curso de Bacharelado em Ciência da Computação, sendo 5 referentes a reuniões ordinárias e 1 extraordinária de cada setor.

O trabalho de anotação visa identificar como usuários entendem as transições de assunto no documento, visto a ausência de marcações e meta informações no texto. Além disso, as anotações desse trabalho devem registrar o entendimento dos usuários em relação aos assuntos de cada segmento identificado e classificá-los quanto ao tipo de menção dada ao assunto. Maiores detalhes sobre a segmentação e rotulação manual serão fornecidos mais adiante nessa seção.

Após a delimitação do fenômeno a ser explicado e escolha do *corpus* apropriado, selecionou-se um grupo de anotadores para analisar e coletar dados referentes a cada ata. Um grupo de 9 anotadores foi formado por profissionais com alguma afinidade com atas de reunião, como profissionais administrativos, professores e coordenadores de curso. A fim de facilitar o trabalho de anotação e diminuir eventuais erros, optou-se por desenvolver um *software*[2] como ferramenta para a coleta dos dados, conforme sugerido em (HOVY; LAVID, 2010). Essa ferramenta foi modelada para permitir aos anotadores visualizar os documentos e indicar livremente as divisões entre segmentos, bem como rotulá-los.

Nesse processo, cada documento original foi apresentado ao anotador em formato de texto plano. O processo de anotação de uma ata se iniciava com a leitura e identificação dos trechos de texto que tratavam de um único assunto. Um trecho após identificado, era selecionado e movido do texto integral para um painel onde a interface proveu controles para anotação. O processo se repetia até que todas as atas fossem segmentadas e anotadas.

O trabalho de rotulação do assunto de cada seguimento, constituiu-se em 3 passos:

1. Classificação quanto ao tipo de comunicação, onde se especificou uma entre as classes *"decisão"*, *"informe"* e *"irrelevante"*.

2. Classificação quanto ao contexto onde se gerou o assunto, onde se especificou entre as classes *"discussão"*, *"orientação"* e *"solicitação"*, podendo o anotador indicá-las simultaneamente. Para os passos 1 e 2 havia a opção *"outro"* em que o anotador poderia especificar livremente o tipo e o contexto do assunto.

[1] Acessível em <http://www.ppgccs.net/?page_id=1150>

[2] Códigos fontes disponíveis em: <https://github.com/ovidio-francisco/UFSCar/tree/master/codes/TextSegmentationTool>

3. Descrição do assunto, onde o anotador apontou até 5 palavras contidas no texto para representar o teor do segmento. A lista de palavras foi gerada a partir das palavras contidas no texto, excluindo-se as *stop words*, sem repetições, e apresentadas na ordem que ocorrem no texto.

Esses rótulos tem como propósito ajudar a descrever os segmentos de atas por meio de classes e descritores, podendo ser utilizados na etapa de treinamento de classificadores e avaliação de extratores de tópicos. Na Figura 13 é mostrada a interface da ferramenta utilizada para as anotações.

De acordo com a literatura a respeito de anotações em segmentos, é recomendável fornecer treinamento aos participantes. Nesse trabalho os anotadores receberam apenas informações básicas sobre o objetivo da pesquisa e instruções de como operar o *software* por meio de um vídeo[3], deixando o entendimento sobre a segmentação e rotulação sob responsabilidade de cada anotador. Assim, nenhum critério foi estabelecido para o procedimento, ficando os anotadores orientados apenas pela interface da ferramenta.

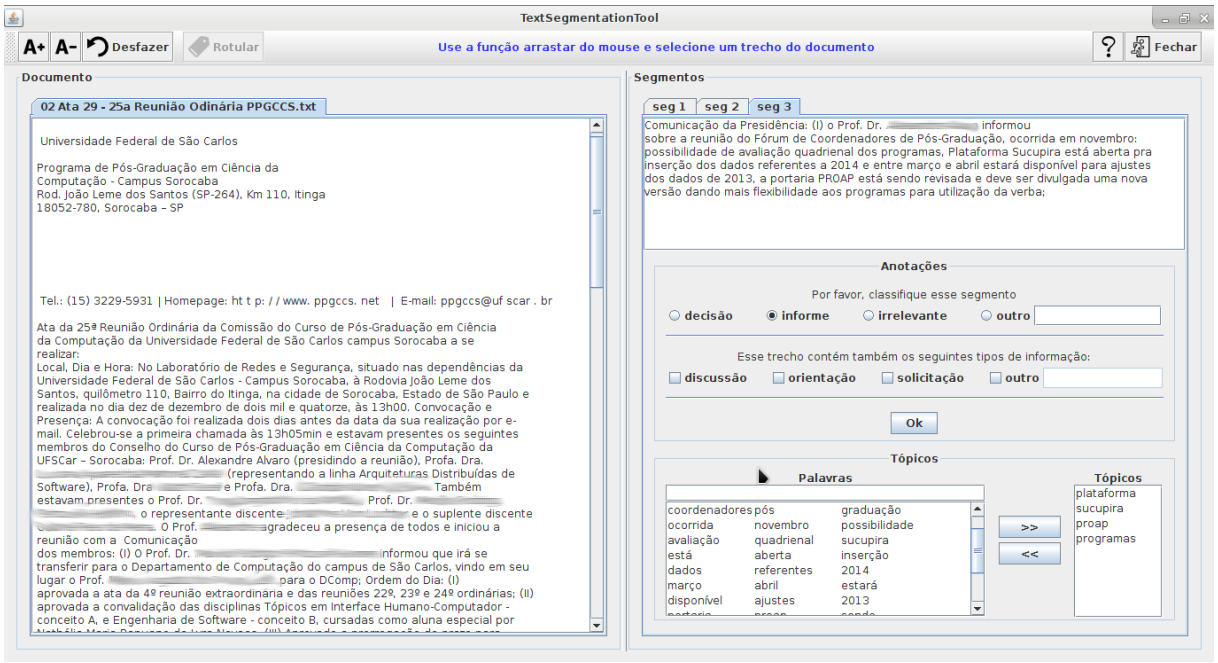


Figura 13 – Interface da ferramenta utilizada para anotações onde o texto a ser segmentado é exibido no painel a esquerda e os controles para anotação estão disponíveis a direita.

O procedimento de anotação deu-se remotamente. Os anotadores receberam por e-mail o acesso ao *software* e individualmente realizaram as anotações. A tarefa de segmentar e rotular manualmente demandou em torno de 4 a 6 horas de atenção dos anotadores. Em razão desse esforço, o *software* permitiu que a tarefa fosse pausada e reiniciada sempre que houvesse necessidade, sem prejuízos ao trabalho.

[3] Acessível em <https://youtu.be/zCv3U7Z8kKo>

O *software* foi desenhado para padronizar o trabalho e otimizar o tempo dos anotadores. Ao mesmo tempo, dava dava liberdade para segmentar o texto após qualquer character do texto. Um segmento sempre deve terminar com um sentença completa, uma vez que as sentenças são a menor unidade de informação para os segmentadores, conforme explicado na seção 3.2 e Algoritmo 1. Como previsto, houve casos onde o anotador omitiu ou adicionou characteres e palavras (quase sempre ocasionado por imprecisão durante a seleção do texto). Para garantir que todo segmento termine com uma sentença completa, aqueles que apresentavam sentenças incompletas tiveram seu incio ou final movidos para se adequar à saída dos algoritmos de segmentação. Após o processo, 12 atas foram segmentadas manualmente por 9 anotadores, gerando um conjunto de 108 anotações. Na Tabela 4, é mostrado para cada ata a quantidade de sentenças do documento original e de segmentos identificados por cada anotador.

Ata	**#Sent.**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
Ata 1	25	7	4	11	6	16	8	8	15	16
Ata 2	17	4	4	8	6	11	6	6	15	14
Ata 3	26	6	6	8	4	15	9	10	18	14
Ata 4	26	5	5	10	6	14	17	7	11	12
Ata 5	33	4	4	6	5	17	22	9	18	16
Ata 6	11	3	4	6	4	9	9	4	7	5
Ata 7	20	3	7	5	4	11	14	5	5	4
Ata 8	35	4	8	3	8	12	17	5	11	9
Ata 9	24	3	5	3	6	11	11	3	9	9
Ata 10	50	4	5	4	7	31	29	5	9	8
Ata 11	43	4	7	5	7	29	19	5	9	12
Ata 12	56	3	10	4	16	33	25	4	13	11
Total	**366**	**50**	**69**	**73**	**79**	**209**	**186**	**71**	**140**	**130**

Tabela 4 – Descrição dos resultados obtidos com anotadores. Na segunda coluna #Sent., é mostrada a quantidade de sentenças de cada ata. Nas colunas A1-A9 é mostrado as quantidades de segmentos informados pelos anotadores.

O *corpus* anotado deve ser constituído a partir dos textos originais e da resultante das percepções dos anotadores a cerca do fenômeno a ser explicado. Assim, o texto de cada uma das 12 atas selecionadas será acrescido de informações geradas a partir dos dados coletados. Após o processo de anotação, a segmentação de referência foi criada utilizando o critério de maior concordância, como já relatado em outros trabalhos (HEARST, 1997; CARDOSO; PARDO; TABOADA, 2017; KAZANTSEVA; SZPAKOWICZ, 2012; PASSONNEAU; LITMAN, 1997; GALLEY et al., 2003). Assim, considerou-se que ocorre um limite entre segmentos quando a maioria dos anotadores (metade mais um) concordou que a mesma sentença é um final de segmento. Gerou-se então 12 documentos derivados das atas originais os quais constituem a segmentação de referência utilizada nesse trabalho.

Na Figura 14 é mostrado um exemplo de criação de uma segmentação de referência

por meio da concordância entre anotadores. As primeiras linhas representam segmentações fornecidas por anotadores e a última linha representa a segmentação resultante da concordância entre a maioria dos segmentadores e portanto mais confiável.

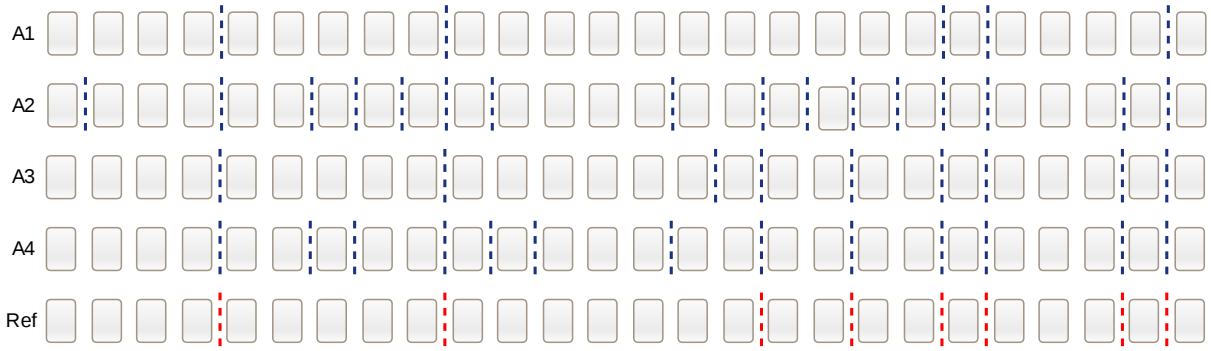


Figura 14 – Exemplo uma segmentação de referência criada a partir da concordância entre segmentações manuais.

Na Tabela 5 é mostrado um exemplo em que 6 dos 9 anotadores concordaram a respeito de um segmento. A tabela mostra quatro segmentos extraídos da segmentação de referência onde cada linha contém um segmento e os índices a esquerda indicam uma sentença. Junto à cada segmento é mostrada a classe e descritores rotulados por um dos anotadores. Vale ressaltar que esses rótulos não foram utilizados no processo de segmentação e não têm nenhuma influência sobre a segmentação de referência.

[7]	(II) Encerrada a etapa de inscrição para o processo seletivo como aluno regular para o segundo semestre de 2015: foram quarenta e nove inscrições on-line e dezoito candidatos entregaram a documentação; *<informe> <processo;seletivo>*
[8]	(III) O Prof. Dr. AAA informou que a Pró-Reitora comunicou a oferta de mais uma bolsa pela cota da Pró-Reitoria, mas não havia aluno disponível para alocação da bolsa.
[9]	(III) O Prof. Dr. AAA informou que a Pró-Reitora comunicou a oferta de mais uma bolsa pela cota da Pró-Reitoria, mas não havia aluno disponível para alocação da bolsa.
[10]	O Prof. Dr. BBB informou que havia uma aluna interessada, mas não informada durante o processo de elaboração do ranking no início do semestre.
[11]	Ficou decidido enviar e-mail aos docentes solicitando que comuniquem permanentemente interesse de alunos em bolsa pra atualização do ranking; *<informe> <solicitação;bolsa;cota;ranking;alunos>*
[12]	(IV) Com a mudança do Prof. Dr. DDD para o campus de São Carlos, o Prof. Dr. BBB assume o posto de suplente da linha Teoria Aplicada à Computação na CPG; *<informe> <mudança;suplente;teoria;aplicada;computação>*
[13]	Comunicação dos membros: Não houve; *<irrelevante>*

Tabela 5 – Exemplo de segmentação de referência com rotulação de um anotador

A qualidade da segmentação de referência está ligada principalmente a conformidade das anotações quanto a atribuição de final de segmento a cada sentença. Para

mensurar a concordância entre anotadores, a medida *kappa (k)* (CARLETTA, 1996) é frequentemente utilizada (GRUENSTEIN; NIEKRASZ; PURVER, 2007; CARDOSO; PARDO; TABOADA, 2017; HEARST, 1997). Ela mostra como os anotadores compreendem os textos analisados e o nível de confiabilidade da segmentação de referência. Essa medida retorna um valor no intervalo de 0 até 1, onde 1 significa uma concordância perfeita e 0 que não houve concordância. As medidas *WindowDiff* e P_k informam a probabilidade de duas sentenças escolhidas aleatoriamente estarem no mesmo segmento. Elas refletem a similaridade entre duas segmentações, assim também podem ser utilizadas para avaliar a concordância entre os anotadores. As formulações dessas medidas são abordadas em mais detalhes na Seção 2.3.2. Nesse trabalho calculou-se as medidas *WindowDiff*, P_k e *Kappa* para cada par de anotação referente a cada ata, ou seja, para cada uma das 12 atas calculou-se a concordância entre os anotadores. A Tabela 6 contém, para cada ata, a quantidade de sentenças e a quantidade de segmentos identificadas pelos participantes e as médias de *WindowDiff*, P_k e *Kappa*. Uma vez que P_k e *WindowDiff*, são medidas de dissimilaridade, são reportados os seus complementos.

Referência	**#Seg**	***Kappa*** Média (des. pad.)	$\boldsymbol{P_k}$ Média (des. pad.)	***WinDiff*** Média (des. pad.)
Ref. 01	15	0,344 (0,190)	0,433 (0,170)	0,631 (0,409)
Ref. 02	13	0,266 (0,246)	0,439 (0,190)	0,565 (0,379)
Ref. 03	15	0,328 (0,183)	0,442 (0,165)	0,590 (0,353)
Ref. 04	15	0,364 (0,241)	0,364 (0,161)	0,562 (0,377)
Ref. 05	19	0,315 (0,217)	0,458 (0,223)	0,889 (0,640)
Ref. 06	9	0,314 (0,218)	0,404 (0,163)	0,463 (0,266)
Ref. 07	8	0,235 (0,208)	0,343 (0,192)	0,507 (0,401)
Ref. 08	12	0,211 (0,225)	0,421 (0,186)	0,629 (0,479)
Ref. 09	12	0,234 (0,258)	0,472 (0,203)	0,660 (0,427)
Ref. 10	13	0,170 (0,206)	0,428 (0,227)	0,937 (1,050)
Ref. 11	12	0,209 (0,236)	0,368 (0,203)	0,704 (0,654)
Ref. 12	21	0,222 (0,195)	0,452 (0,200)	0,113 (1,202)
Total	**13.666**	**0,267 (0,218)**	**0,418 (0,190)**	**0,604 (0,553)**

Tabela 6 – Medidas de concordância entre os anotadores sobre cada ata segmentada manualmente.

Os valores de *Kappa* reforçam que a tarefa de segmentação é bastante subjetiva e indicam que, nesse contexto, a segmentação de referência resultante é pouco confiável, visto os baixos valores de concordância entre os anotadores. Embora (CARLETTA, 1996) afirme que valores de $k > 0,8$ indicam que os dados são confiáveis, visto a subjetividade da tarefa de segmentação textual, medidas menores podem ser aceitáveis, como reportado em (HEARST, 1997) que obteve $k = 0,64$ e (CARDOSO; PARDO; TABOADA, 2017), que obteve $k = 0,56$. Além dos baixos valores para as medidas nota-se também que os anotadores divergem na quantidade de segmentos, o que sugere que diferentes anotadores tem percepções distintas quanto a granularidade de assuntos, ou seja, trechos com pequenas

mudanças de assunto são entendidas como segmentos diferentes por alguns anotadores, enquanto outros entendem como pertencentes ao mesmo assunto.

Após a coleta dos dados das anotações e construção da segmentação de referência, o *corpus* anotado serviu de como base para avaliação dos segmentadores bem como aspectos da configuração e avaliação dos extratores de tópicos no Capítulo 5. O produto resultante do processo de anotação, constituído por segmentos de atas anotados está disponibilizado[4] para outros trabalhos voltados a esse domínio.

4.2 Configuração Experimental

Para avaliação objetiva dos algoritmos, faz-se necessário encontrar os valores de parâmetros que melhor configuram cada algoritmo. O *TextTiling* permite ajustar dois parâmetros, sendo o tamanho da janela e o passo. Por meio de testes empíricos escolheu-se os valores os valores 20, 40 e 60 para o tamanho da janela e 3, 6, 9 e 12 para o passo, gerando ao final 20 configurações, resultantes da combinação desses dois parâmetros.

O *C99* possui três parâmetros: a quantidade de segmentos desejados, o tamanho do quadro utilizado para gerar a matriz de ranking e a representação dos atributos das sentenças. Para o primeiro parâmetro, referente à quantidade segmentos desejados, uma vez que não se conhece o número ideal de segmentos, configurou-se a quantidade de segmentos com base em uma proporção dos candidatos a limite. Para isso atribuiu-se os valores 0,2; 0,4; 0,6; 0,8. Para o segundo parâmetro, referente ao tamanho do quadro utilizado para gerar a matriz de ranking, atribuiu-se os valores 9 e 11, sendo 11 o valor padrão apresentado pelo autor. O algoritmo permite ainda indicar se as sentenças serão representadas por vetores contendo a frequência ou o peso de cada termo. Ambas as representações foram utilizadas. Considerando todos os parâmetros, foram geradas 16 configurações para o algoritmo *C99*.

Os algoritmos tradicionais baseados em coesão léxica como o *TextTiling* e *C99* são fortemente afetados pela distribuição das palavras no texto, pois a maioria das medidas de similaridade, como a cosseno utilizada nesses algoritmos, baseiam-se na frequência das palavras. Para esses, a remoção de termos menos significativos na etapa de pré-processamento pode influenciar o desempenho. Para outras abordagens como *MinCutSeg* e *BayesSeg*, usou-se as configurações fornecidas por (EISENSTEIN; BARZILAY, 2008), onde essas técnicas foram utilizadas como *base line*. Para o *TextSeg* é requerido apenas a configuração da quantidade de segmentos a serem identificados. Há ainda outras estratégias passíveis de aplicação, como a utilização de fontes externas, por exemplo *thesaurus* e palavras pista, como discutido em (NAILI; CHAIBI; GHEZALA, 2016; GUTIERREZ

[4] Acessível em <https://github.com/ovidio-francisco/UFSCar/>

et al., 2016; FERRET, 2009). Nesse trabalho, essas estratégias não são utilizadas para manter uma abordagem não supervisionada e independente de domínio.

Calculou-se as medidas configurando cada algoritmo a fim de conhecer o impacto do pré-processamento nos algoritmos. Esses foram testados em duas etapas: com o texto integral, e com o texto pré-processado em que elementos menos significativos foram removidos, conforme mencionado na Seção 3.1.1.

4.3 Critérios de Avaliação

Para fins de avaliação desse trabalho, um bom método de segmentação é aquele cujo resultado melhor se aproxima da segmentação de referência, sem a obrigatoriedade de estar perfeitamente alinhado com tal. Ou seja, visto o contexto das atas de reunião, e a subjetividade da tarefa, não é necessário que os limites entre os segmentos (real e hipótese) sejam idênticos, mas que se assemelhem em localização e quantidade.

As segmentações obtidas com os algoritmos foram comparadas com a segmentação de referência obtida e calculou-se as medidas mais aplicadas à segmentação textual, P_k e *WindowDiff*. Além dessas, computou-se também as medidas tradicionais Acurácia e F^1 para análises que consideram a exatidão dessas técnicas.

4.4 Resultados

Obteve-se por meio dos testes apresentados as melhores configurações para as principais medidas de avaliação de segmentadores. Com essas configurações calculou-se a média de cada medida considerando o conjunto de documentos. Os algoritmos foram executados com as configurações apresentadas e discutidos mais adiante nessa seção.

A seguir são apresentados os resultados obtidos com os algoritmos baseados em coesão léxica considerando seus principais parâmetros e a aplicação do pré-processamento. Em seguida, são apresentados os resultados da avaliação geral de todos os algoritmos abordados nesse trabalho. Na Tabela 7 são apresentadas, as médias das medidas de desempenho obtidas com o *TextTiling* bem como as configurações utilizadas.

Observou-se que o *TextTiling* produz resultados muito similares em relação ao texto pré-processado e o texto integral. A quantidade de segmentos obtidos é influenciada por **Step** e **WS** (*Win Size*), em que menores valores desses parâmetros tendem a produzir mais segmentos, que por consequência também refletem em melhores valores para F^1, pois que essa medida é baseada na precisão, a qual é fortemente afetada por falsos positivos. Os melhores valores de *WinDiff* e P_k coincidem nas mesmas configurações que também apresentam resultados mais acurados. A relação entre essas medidas é discutida mais adiante nessa seção.

Step	WS	Com texto integral					Com texto pré-processado				
		WD	***P_k***	***Ac.***	***F^1***	**#Segs**	***WD***	***P_k***	***Ac.***	***F^1***	**#Segs**
20	30	0.513	0.490	0.538	0.334	8.500	0,461	0,444	0,581	**0,411**	8,833
	35	0.509	0.492	0.540	0.350	8.583	0,462	0,443	0,582	0,401	8,750
	40	0.517	0.495	0.532	0.342	8.583	0,485	0,466	0,562	0,378	8,250
	45	0.496	0.477	0.555	0.347	7.667	0,480	0,458	0,572	0,369	8,250
	50	0.481	0.465	0.569	**0.390**	8.750	0,523	0,503	0,528	0,327	8,417
	55	0.512	0.493	0.542	0.337	8.250	0,491	0,474	0,549	0,331	8,250
30	30	0.511	0.494	0.538	0.284	6.667	0,509	0,488	0,536	0,286	6,917
	35	0.517	0.500	0.536	0.285	6.583	0,500	0,479	0,551	0,318	7,167
	40	0.512	0.491	0.543	0.299	6.750	0,468	0,451	0,576	0,348	6,750
	45	0.502	0.483	0.555	0.320	6.917	**0,450**	**0,435**	**0,596**	0,373	6,417
	50	0.510	0.493	0.539	0.313	7.333	0,493	0,478	0,543	0,307	6,417
	55	0.498	0.480	0.543	0.328	7.250	0,481	0,463	0,558	0,346	7,083
40	30	0.493	0.477	0.555	0.248	4.917	0,475	0,460	0,566	0,306	5,833
	35	0.482	0.465	0.558	0.267	5.417	0,501	0,482	0,542	0,268	6,083
	40	0.476	0.459	0.565	0.275	5.500	0,499	0,478	0,548	0,293	6,083
	45	0.501	0.482	0.549	0.260	5.333	0,488	0,471	0,551	0,275	5,500
	50	0.498	0.481	0.551	0.266	5.333	0,495	0,474	0,552	0,280	5,833
	55	0.505	0.487	0.544	0.243	5.083	0,476	0,453	0,567	0,310	6,083
50	30	**0.474**	**0.455**	**0.579**	0.295	4.917	0,492	0,473	0,557	0,274	5,167
	35	0.528	0.511	0.531	0.202	4.583	0,504	0,484	0,549	0,268	5,583
	40	0.501	0.488	0.539	0.234	5.000	0,501	0,481	0,556	0,278	5,417
	45	0.489	0.476	0.558	0.275	5.167	0,508	0,484	0,549	0,264	5,500
	50	0.498	0.483	0.545	0.304	6.083	0,513	0,491	0,536	0,253	5,417
	55	0.490	0.470	0.556	0.303	5.583	0,509	0,487	0,543	0,276	5,833
60	30	0.499	0.486	0.557	0.234	4.417	0,481	0,462	0,564	0,267	4,917
	35	0.509	0.494	0.537	0.243	5.000	0,503	0,483	0,549	0,250	5,083
	40	0.501	0.486	0.545	0.182	3.833	0,497	0,481	0,554	0,242	4,750
	45	0.493	0.478	0.558	0.227	4.167	0,465	0,448	0,577	0,271	4,500
	50	0.495	0.478	0.562	0.225	4.083	0,478	0,459	0,569	0,250	4,333
	55	0.500	0.485	0.550	0.198	4.000	0,474	0,457	0,568	0,269	5,000

Tabela 7 – Resultados do TextTiling.

Na Tabela 8 são apresentadas, as médias obtidas com o *C99* bem como as configurações utilizadas, onde **SR** (*Segment Rate*) é a proporção de segmentos em relação a quantidade de candidatos, **RS** (*Ranking Size*) é o tamanho do quadro utilizado para criar a matriz de *rankings* e **W** indica se as sentenças são representadas por vetores contendo a frequência dos termos ou um peso que representa sua importância no documento.

Verificou-se que, entre os métodos baseados em coesão léxica, o *C99* obteve melhor desempenho em acurácia, precisão, F^1, P_k e *WindowDiff*, em relação ao *TextTiling*. De maneira geral, o algoritmo *C99* apresenta melhores resultados em relação ao *TextTiling*. O mesmo comportamento é observado para as demais técnicas *MinCutSeg*, *BayesSeg* e *TextSeg*, conforme pode ser verificado no Apêndice A que contém os resultados completos de todos algoritmos. Nesse trabalho, optou-se por aplicar o pré-processamento em todas as técnicas visto que além de apresentar pequena vantagem na segmentação textual, essa prática é amplamente aplicada em trabalhos nas áreas de Mineração de Texto pela capacidade de diminuição de dimensões e tempo de processamento.

A avaliação final foi feita pela comparação dos resultados dos algoritmos com a segmentação de referência usando as medidas *Pk* e *WindowDiff*. É apresentada também, para fins de comparação com outros trabalhos, as medidas tradicionais acurácia e F^1, entretanto, nesse contexto, essas medidas são menos significativa que P_k e *WindowDiff*, conforme já mencionado na Seção 2.3.2. Cada técnica foi executada variando seus principais

W	RS	SR	Com texto integral				Com texto pré-processado				#Segs
			WD	***P_k***	***Ac.***	***F^1***	***WD***	***P_k***	***Ac.***	***F^1***	
Sim	3	0,20	0,481	0,463	0,574	0,324	0,463	0,445	0,581	0,339	6,083
		0,30	0,457	0,437	0,596	0,447	**0,434**	**0,407**	0,607	0,457	9,250
		0,40	0,450	0,425	0,602	0,513	0,452	0,422	0,604	0,515	12,083
		0,50	**0,435**	**0,395**	**0,629**	0,594	0,499	0,458	0,577	0,539	15,500
		0,60	0,489	0,437	0,592	0,591	18,417	0,487	0,440	0,592	0,591
		0,70	0,482	0,420	0,602	**0,632**	0,485	0,431	0,602	**0,633**	21,417
	5	0,20	0,488	0,469	0,565	0,313	0,454	0,437	0,583	0,338	6,083
		0,30	0,476	0,458	0,571	0,426	0,454	0,434	0,595	0,446	9,250
		0,40	0,476	0,452	0,578	0,487	0,475	0,443	0,590	0,497	12,083
		0,50	0,463	0,425	0,605	0,566	0,460	0,421	**0,609**	0,571	15,500
		0,60	0,464	0,415	0,610	0,604	0,491	0,442	0,591	0,588	18,417
		0,70	0,504	0,435	0,589	0,619	0,525	0,449	0,576	0,609	21,417
	7	0,20	0,478	0,459	0,574	0,328	0,491	0,474	0,555	0,293	6,083
		0,30	0,481	0,462	0,570	0,418	0,486	0,469	0,565	0,395	9,250
		0,40	0,478	0,452	0,577	0,482	0,502	0,472	0,561	0,453	12,083
		0,50	0,471	0,427	0,604	0,563	0,460	0,421	0,604	0,561	15,500
		0,60	0,480	0,429	0,599	0,594	0,486	0,433	0,591	0,585	18,417
		0,70	0,516	0,444	0,579	0,611	0,547	0,470	0,551	0,586	21,417
Não	3	0,20	0,469	0,453	0,579	0,335	0,448	0,427	0,596	0,362	6,083
		0,30	0,441	0,421	0,608	0,463	0,454	0,426	0,594	0,445	9,250
		0,40	0,467	0,439	0,591	0,493	0,490	0,455	0,568	0,469	12,083
		0,50	0,483	0,442	0,593	0,554	0,529	0,481	0,543	0,503	15,500
		0,60	0,500	0,442	0,589	0,587	0,554	0,499	0,528	0,535	18,417
		0,70	0,492	0,423	0,602	0,632	0,565	0,496	0,526	0,570	21,417
	5	0,20	0,495	0,476	0,555	0,300	0,498	0,479	0,545	0,277	6,083
		0,30	0,503	0,485	0,549	0,386	0,505	0,482	0,540	0,369	9,250
		0,40	0,496	0,477	0,564	0,466	0,536	0,504	0,520	0,407	12,083
		0,50	0,488	0,452	0,574	0,533	0,540	0,490	0,529	0,485	15,500
		0,60	0,484	0,434	0,594	0,592	0,529	0,469	0,545	0,543	18,417
		0,70	0,522	0,451	0,574	0,609	0,542	0,464	0,549	0,584	21,417
	7	0,20	0,489	0,471	0,560	0,307	0,512	0,495	0,534	0,250	6,083
		0,30	0,498	0,479	0,554	0,394	0,527	0,506	0,522	0,336	9,250
		0,40	0,500	0,475	0,561	0,462	0,530	0,494	0,535	0,420	12,083
		0,50	0,479	0,441	0,592	0,551	0,503	0,454	0,571	0,523	15,500
		0,60	0,493	0,439	0,585	0,586	0,511	0,453	0,565	0,562	18,417
		0,70	0,506	0,430	0,590	0,621	0,559	0,476	0,535	0,572	21,417

Tabela 8 – Resultados do *C99*.

parâmetros a fim de verificar qual configuração melhor otimiza cada algoritmo. Todos os resultados analisados nessa avaliação final foram executados com o texto pré-processado, visto que essa etapa tem (embora pequena) influência positiva nos resultados, como apresentado nas tabelas 7 e 8 para os algoritmo baseados em coesão léxica e demais algoritmos no Apêndice A. A Tabela 9 contém a média dos dados obtidos onde os melhores resultados estão destacados. Vale lembrar que P_k e *WindowDiff* são medidas de dissimilaridade, ou seja, os valores menores significam melhores resultados.

De maneira geral, os algoritmos avaliados sobressaem ao obtido pelo *PseudoSeg* (que simplesmente atribui um segmento a todas sentenças). Observa-se também que os valores de P_k e *WindowDiff* são próximos devido a natureza similar dessas medidas. Pode-se notar também que as configurações que produzem os melhores valores de acurácia também registram melhores valores de P_k ou *WindowDiff* como se vê para *TextTiling*, *MinCutSeg* e *BayesSeg*, à exceção de *C99* que registra sua melhor acurácia muito próxima à obtida na configuração ótima para P_k e *WindowDiff*. Essa semelhança entre as medidas que toleram proximidade entre segmentações (P_k e *WindowDiff*) e a acurácia que apenas computa limites exatos, sugere que quando os anotadores concordam que dois blocos

Algoritmo		**Step**	**Win**	$\mathbf{P}_k$	**WD**	**Ac**	$\mathbf{F}^1$	**#Segs**
TextTiling-1		20	30	0.461	0.444	0.581	**0.411**	8.833
TextTiling-2		30	45	**0.450**	**0.435**	**0.596**	0.373	6.417
Algoritmo	**RS**	**W**	**SRate**	$\mathbf{P}_k$	**WD**	**Ac**	$\mathbf{F}^1$	**#Segs**
C99-1	3	true	0.300	**0.434**	**0.407**	0.607	0.457	9.250
C99-2	3	true	0.700	0.485	0.431	0.602	**0.633**	21.417
C99-3	5	true	0.500	0.460	0.421	**0.609**	0.571	15.500
Algoritmo		**Cut**	**SRate**	$\mathbf{P}_k$	**WD**	**Ac**	$\mathbf{F}^1$	**#Segs**
MinCutSeg-1		9	0.400	**0.444**	0.408	**0.614**	0.526	11.917
MinCutSeg-2		11	0.500	0.459	**0.407**	0.603	0.563	15.000
MinCutSeg-3		5	0.700	0.528	0.438	0.567	**0.599**	21.000
Algoritmo	**Prior**	**Disp.**	**SRate**	$\mathbf{P}_k$	**WD**	**Ac**	$\mathbf{F}^1$	**#Segs**
BayesSeg-1	0.080	0.500	Auto	**0.380**	**0.361**	**0.655**	0.551	10.000
BayesSeg-2	0.110	0.100	0.600	0.462	0.399	0.615	**0.619**	18.417
Algoritmo			**SRate**	$\mathbf{P}_k$	**WD**	**Ac**	$\mathbf{F}^1$	**#Segs**
TextSeg-1			Auto	**0.455**	0.439	0.585	0.368	6.417
TextSeg-2			0.500	0.475	**0.417**	**0.594**	0.566	15.500
TextSeg-3			0.900	0.604	0.484	0.524	**0.627**	27.500
Algoritmo			**SRate**	$\mathbf{P}_k$	**WD**	**Ac**	$\mathbf{F}^1$	**#Segs**
PseudoSeg			1.000	**0.640**	**0.490**	**0.506**	**0.638**	30.500

Tabela 9 – Resumo dos melhores resultados obtidos por cada configuração e medida de avaliação para cada algoritmos utilizado na configuração experimental.

de texto referem-se a assuntos diferentes, também concordam no ponto exato onde há transição de assunto. Esses resultados são ilustrados graficamente na Figura 15.

Durante a avaliação experimental analisou-se a influência das parâmetros na eficiência dos algoritmos, em que observou-se que a proporção de segmentos extraídos causaram maior impacto nas medidas de desempenho. Na Figura 15 é exibida a relação entre a taxa de segmentação e as medidas de desempenho as quais apresentam melhores valores entre 30% e 50% de sentenças marcadas como final de segmento.

A abordagem baseada em janelas deslisantes empregada no *TextTiling* não permite a configuração direta da quantidade de segmentos extraídos, possibilitando o ajuste do tamanho do passo e comprimento da janela que influenciam seu comportamento nesse aspecto, os quais são analisados a seguir. Uma vez que a coesão léxica é pressuposto de muitas abordagens em segmentação textual, fez-se uma análise desses documentos quanto a similaridade dos termos ao longo do texto. Verificou-se que a técnica de janelas deslizantes empregada pelo TextTiling encontra os vales que indicam transições entre segmentos. Contudo ao comparar esses vales com a segmentação de referência, nota-se que a maioria dos limites coincide ou estão próximos aos vales. Porém há casos onde a referência indica limites em trechos com alta coesão léxica e outros onde a queda da coesão, indicada por vales, não coincide com nenhum limite de referência.

Na Figura 16 é apresentado a variação da coesão léxica ao longo de uma ata e a

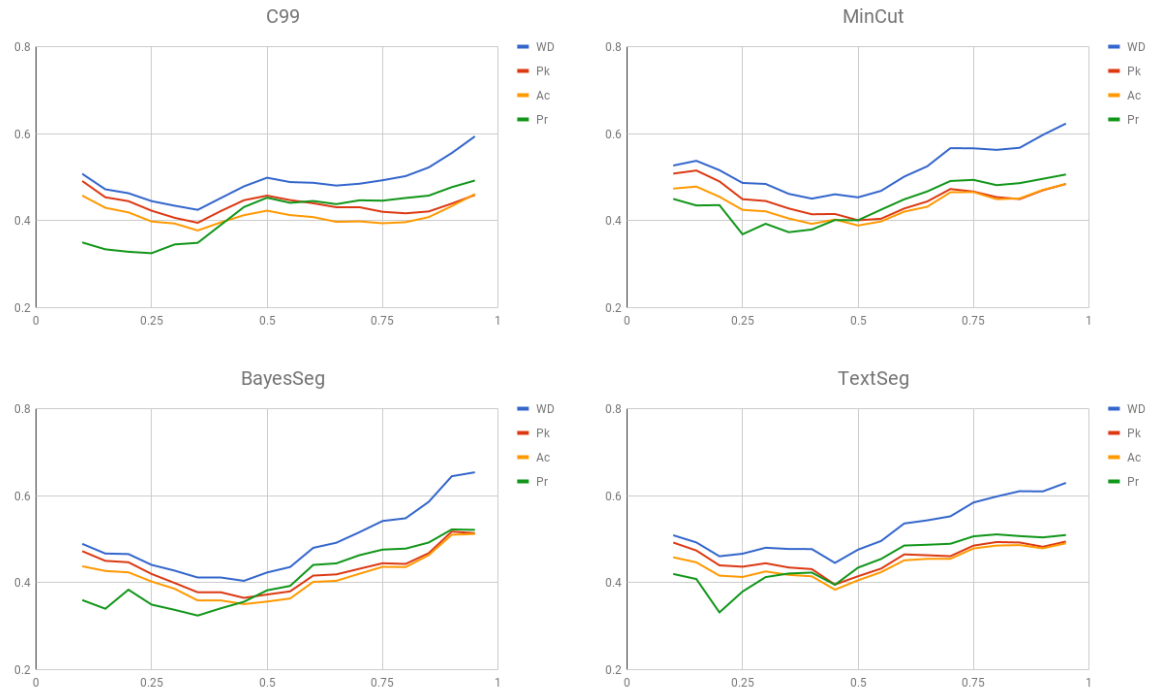


Figura 15 – Influência do taxa segmentos na eficiência dos algoritmos

segmentação obtida pelo *TextTiling* usando tamanho de janela igual a 50 e passo 9.

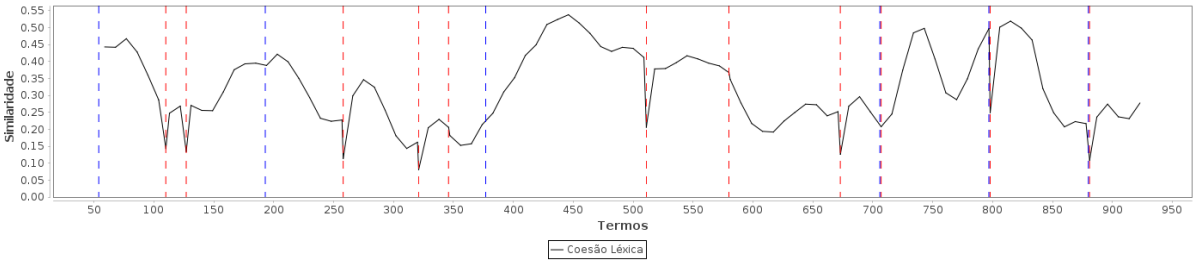


Figura 16 – Variação da coesão léxica ao longo de uma ata junto a uma segmentação automática em contraste com uma segmentação de referência. A linha horizontal representa a variação da coesão léxica e as linha verticais azuis e vermelhas representam os limites entre segmentos atribuídos pela referência e pelo algoritmo respectivamente.

Em termos de performance, o modelo implementado no algoritmo *BayesSeg* apresenta melhores resultados para as medidas P_k e *WindowDiff*, visto que essas são mais adequadas ao contexto analisado, este será empregado nos próximos experimentos. Contudo, dado a subjetividade da tarefa de segmentação textual, outros modelos podem ser utilizados satisfatoriamente dependendo do critério adotado.

Na Figura 17 é apresentada a performance dos algoritmos. Observa-se valores altos de F^1 para a segmentação por sentenças, pois é atribuído um limite a todo candidato a final de segmento, o que resulta no valor máximo para revocação. De maneira semelhante,

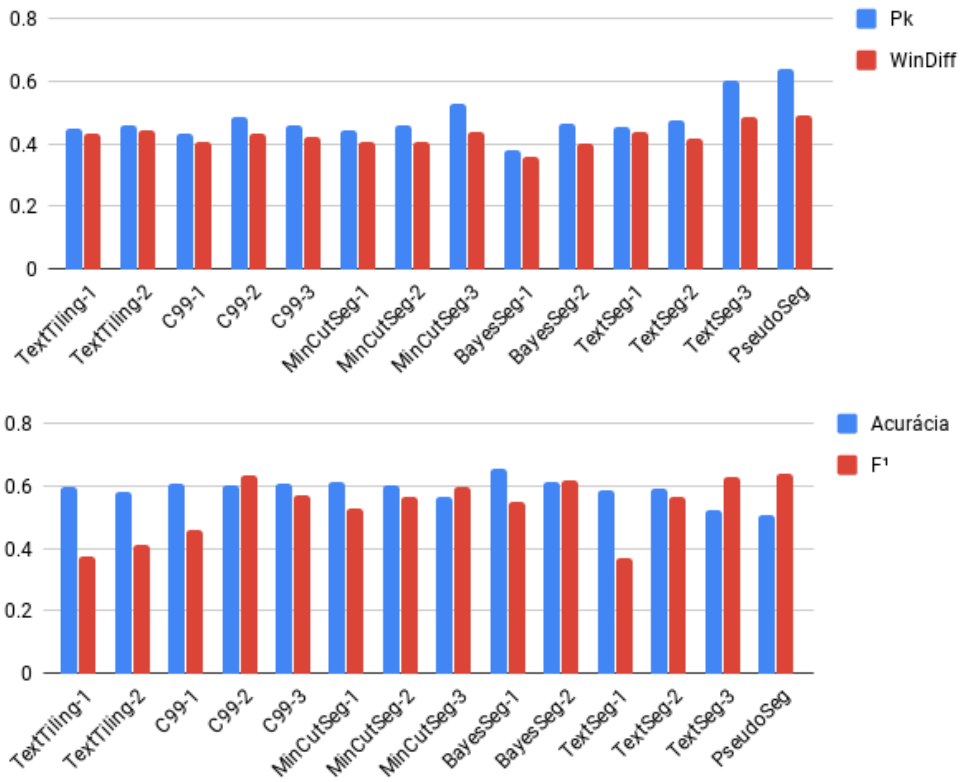


Figura 17 – Performance geral dos algoritmos de segmentação textual e as configurações que apresentam melhores resultados para cada medida de desempenho.

o comportamento do *TextTiling* gera menos segmentos em relação aos demais, e com isso tem-se valores menores de F^1, o que pode ser alterado pela configuração do algoritmo com passos mais curtos, ou ainda, sobre-escrevendo a função que calcula os *depth scores* para reconhecer vales menos profundos.

Após a análise dos métodos, selecionou-se o algoritmo *BayesSeg* para a tarefa de segmentação textual nesse trabalho. A escolha justifica-se por este algoritmo apresentar melhores resultados para as medidas P_k e *WindowDiff* visto a subjetividade da tarefa e a aderência dessas medias à acurácia e precisão. Mais análises quanto à segmentação textual e desempenho do *BayesSeg* serão abordadas no Capítulo 5. Em que realizou-se então uma experimento a fim de avaliar subjetivamente o segmentador escolhido em conjunto com extratores de tópicos.

4.5 Considerações Finais

Nessa seção os algoritmos de segmentação textual foram avaliados objetivamente. Inicialmente, criou-se um conjunto com segmentações manuais usados para gerar uma segmentação de referência, sob a qual os resultados dos algoritmos foram comparados. O produto desse processo também foi acrescido de rótulos e descrições sobre o assunto de

cada segmento. Embora a concordância sobre a segmentação entre anotadores tenha se mostrado abaixo do esperado, obteve-se, *corpus* anotado que ajuda a entender a coleção de atas original em termos de transição de assuntos entre segmentos e seus conteúdos. Contudo, mais experimentos são necessários para criação de uma segmentação de referência mais robusta, a qual pode ser obtida com maior número de anotadores, e aperfeiçoamento do processo de anotação. Após esses primeiros resultados, melhorias na ferramenta para segmentação manual e aplicação de um treinamento adequado dos participantes podem ser desenvolvidos com base nessa primeira experiência.

De maneira geral, os dados apontam que há pouca diferença de performance entre os algoritmos de segmentação textual. Contudo, o *BayesSeg* mostrou-se superior aos demais considerando-se a similaridades entre seus resultados e a segmentação de referência, as quais foram mensuradas pelas medidas *WindowDiff* e P_k. Uma análise subjetiva do segmentador escolhido é apresentada na Seção 5, em que observa-se resultados satisfatórios a repeito da qualidade dos segmentos obtidos.

O *corpus* criado a partir da coleção original de documentos foi disponibilizados para trabalhos futuros voltados a esse domínio bem como o *software* para segmentação manual e rotulação de documentos.

5 Avaliação dos Extratores de Tópicos

Nesse capítulo, as técnicas de extração de tópicos consideradas nesse projeto de mestrado são avaliadas. O objetivo é comparar os algoritmos de extração de tópicos em sua capacidade de representar os segmentos. Para essa análise, escolheu-se os modelos LDA, PLSA e K-Means devido a popularidade desses métodos os quais são amplamente utilizados (ZHU et al., 2012) e frequentemente referenciados em trabalhos voltados a organização de bases textuais (AGGARWAL, 2018; O'CALLAGHAN et al., 2015; STEYVERS; GRIFFITHS, 2007). Os algoritmos foram inicialmente configurados com base em avaliações internas (HASSANI; SEIDL, 2017) e observações empíricas nas quais escolheu-se os melhores valores para seus parâmetros. Os resultados desses modelos foram submetidos a uma avaliação subjetiva a fim de analisá-los junto a usuários com afinidade com atas de reuniões.

Nessa avaliação, utilizou-se o segmentador *BayesSeg* por apresentar melhores resultados quanto a segmentação das atas. Essa técnica também foi avaliada nesse experimento, uma vez que é a etapa anterior a extração de tópicos está diretamente ligada a os resultados apresentados ao avaliador bem como pode interferir no funcionamento dos modelos de extração de tópicos. Assim, a técnica de segmentação textual foi avaliada subjetivamente em complemento a análise apresentada no Capítulo 4.

A avaliação se deu por meio de questionários onde profissionais com afinidade com atas de reunião forneceram suas percepções em relação aos resultados dos modelos de extração de tópicos. Por fim, os dados obtidos dos experimentos serviram de base para as análises dos algoritmos e de sua aplicação no contexto das atas de reuniões.

5.1 Configuração Experimental

Durante os primeiros testes, a qualidade dos resultados mostrou-se sensível à quantidade de tópicos extraídos. Inicialmente, realizou-se um teste prévio utilizando uma versão não-paramétrica dos algorítimos a fim de automaticamente obter valores ótimos para esse parâmetro por meio da análise das medidas Silhueta e Coesão. Essa configuração automática resultou valores em torno de 20 tópicos. Contudo, os resultados apresentam grupos com muitos segmentos (em torno de 100). Com base em observações empíricas, verificou-se que valores abaixo de 60 tópicos geram grupos com muitos segmentos o que por consequência torna os grupos menos coesos por incluir segmentos com assuntos muito distantes, além de diminuir a capacidade representativa dos descritores. Por outro lado, valores acima de 80 geram tópicos com poucos segmentos, permitindo que assuntos próximos sejam atribuídos a grupos distintos. Nesse trabalho, optou-se por configurar os

algoritmos para extrair 70 tópicos da coleção de segmentos por apresentar uma distribuição mais coerente em termos de agrupamento na visão de usuário.

Outro fator importante é a quantidade de descritores selecionados para cada tópico. Com base no experimento de anotações de segmentos, descrito no Capítulo 4, os anotadores selecionaram em média 5 palavras para descrever os segmentos, sendo esse valor adotado para essa avaliação.

5.2 Critérios de Avaliação

Após a identificação da configuração mais adequada para cada algoritmo, cada um dos três modelos de extração de tópicos foi submetido a duas consultas[1]: "*compra de equipamentos*" e "*defesa de dissertação*" gerando 6 cenários distintos a serem analisados. Para cada cenário, o sistema seleciona o tópico com maior relevância com a consulta, conforme apresentado na Seção 3.1.3.1. Em seguida, o sistema exibe 5 segmentos desse tópico escolhidos aleatoriamente. Vale dizer que nessa avaliação as técnicas de ranqueamento dos resultados não são aplicadas para os segmentos contidos no tópico selecionado que estas não interfiram na avaliação dos extratores, contudo, o sistema final poderá ranquear também os segmentos com maior relevância de um ou mais tópicos por meio de técnicas de recuperação de informação. Os resultados desses cenários foram apresentados a um grupo de avaliadores que individualmente avaliaram a qualidade das técnicas de extração de tópicos.

O perfil dos avaliadores é de profissionais da area acadêmica/escolar devido à sua afinidade com o ambiente de gestão e conhecimentos de assuntos relacionados ao *corpus* estudado nesse trabalho. O grupo convidado a participar do experimento é formado por 24 profissionais da UFSCar - Campus Sorocaba, 13 profissionais de escolas técnicas e 3 profissionais de escolas do Ensino Fundamental, sendo 12 ocupantes de cargos de gestão como coordenadores de curso e diretores, 17 membros de conselhos, 5 profissionais administrativos e 3 professores, totalizando 41 avaliadores em que a maioria afirma ter afinidade com atas e reuniões. Apenas 3 declararam nenhuma afinidade com esses documentos, sendo esses últimos descartados por não se adequarem ao perfil desejado. Os avaliadores foram divididos em dois grupos, sendo o primeiro formado por 18 participantes que avaliaram a primeira consulta, "*compra de equipamentos*" e o segundo formado por 19 participantes que avaliaram a segunda consulta, "*defesa de dissertação*". Os grupos avaliaram as técnicas de extração tópicos a partir de uma consulta, ou seja, cada indivíduo avaliou 3 cenários distintos. A avaliação consistiu de um documento impresso contendo uma breve apresentação do trabalho, seguido de uma cópia dos resultados das técnicas de extração de tópicos. Para cada técnica, os avaliadores recebiam 4 questões sobre os

[1] O termo consulta significa um conjunto de palavras chave passadas à um sistema de busca.

resultados. Os documentos de ambas avaliações podem ser vistas no Apêndice B.

O questionário foi formado por questões envolvendo aspectos dos extratores de tópicos e questões referentes à técnica de segmentação textual empregada. Todas as respostas seguiram a escala *Likert* (NORMAN, 2010) com 5 alternativas.

As questões 1 e 2 estão relacionadas ao extrator de tópicos. A primeira, *"Todos os trechos apresentados compartilham um mesmo assunto."*, refere-se ao agrupamento dos segmentos pela qual foi avaliada a semelhança dos trechos em termos de assunto. A segunda questão, *"As palavras <descritores> resumem bem o assunto tratado nos trechos."*, diz respeito aos descritores selecionados, ao respondê-la o avaliador indicou o quão bem esses termos representam aquele grupo. As opções de resposta seguiram a escala *Likert* (NORMAN, 2010) com 5 alternativas, sendo para as questões 1 e 2: As questões 1 e 2, tiveram como alternativas: *"Discordo Totalmente"*; *"Discordo Parcialmente"*; *"Não Concordo, nem Discordo"*; *"Concordo Parcialmente"* e *"Concordo Totalmente"*.

As questões 3 e 4 estão ligadas à técnica de segmentação utilizada, o *BayesSeg* conforme já mencionado no Capítulo 4. A questão 3, *"Existem trechos que não tratam de um único assunto?"*, diz respeito à coesão de cada segmento, levando em conta a homogeneidade do texto em relação a um assunto. A questão 4, *"Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?"*, refere-se a completude dos segmentos, ou seja, o quão bem os segmentos podem ser bem compreendidos independentemente da leitura do documento integral. Para afastar a hipótese de que os resultados das técnicas fossem influenciados pela ordem apresentada, essas foram apresentadas aos avaliadores em ordem aleatórias. As questões 3 e 4, tiveram como alternativas: *"Nenhum"*; *"Poucos"*; *"Nem muitos, nem poucos"*; *"Muitos"* e *"Todos"*.

5.3 Avaliação dos Extratores Junto a Usuários

Inicialmente, a coleção de documentos utilizada como base nesse experimento foi constituída de 175 atas, dessas, foram extraídos 1272 segmentos dos quais os algoritmos atribuíram em média 19 segmentos a cada tópico. Na Tabela 5.3 é exibido para cada cenário a quantidade de segmentos atribuídos ao tópico selecionado na busca, bem como os descritores que os identificam.

Na primeira consulta, *"compra de equipamentos"*, os 3 algoritmos selecionaram um total de 22 segmentos diferentes, sendo que desses 12 foram concomitantemente selecionados por todos os algoritmos. O LDA e o PLSA compartilham 3 desses segmentos nos resultados. Na segunda consulta, *"defesa de dissertação"*, os 3 algoritmos selecionaram um total de 49 segmentos diferentes, sendo que desses 4 foram selecionados concomitantemente por todos os algoritmos. O K-Means compartilha 7 segmentos com o LDA e 3 com o PLSA. O LDA e o PLSA compartilham 2 segmentos nos resultados.

Consulta	Algoritmo	Descritores	#Seg
Consulta 1	KMeans	*compra, material, verba, permanente e valor*	13
	LDA	*verba, compra, pagamento, material e valor*	47
	PLSA	*verba, compra, pagamento, valor e realizada*	9
Consulta 2	KMeans	*orientada, meses, defesa, prazo e dissertação*	12
	LDA	*aprovado, defesa, pedido, dissertação e orientada*	19
	PLSA	*orientada, prazo, bolsa, meses e defesa*	19

Tabela 10 – Descritores e número de segmentos no tópico selecionado pelos Algoritmos

Nessa seção, os dados coletados das avaliações são apresentados e analisados. Os modelos de extração de tópicos discutidos nesse trabalho são comparados de acordo com os critérios mencionados anteriormente: (1) comparar algoritmos de extração de tópicos na tarefa de extração de padrões no contexto das atas de reunião, (2) analisar a qualidade dos descritores extraídos para recuperar os documentos dos grupos. Além disso, as questões referentes à segmentação, são analisadas a fim de validar a performance do segmentador empregado, como complemento ao experimento discutido na Seção 4.

Na Figura 18 é apresentado as frequência das respostas coletadas sobre a primeira questão, a qual refere-se a qualidade do agrupamento levando em conta a semelhança dos segmentos em termos de assunto. A afirmação foi dada como opções de resposta: *"Discordo Totalmente"*; *"Discordo Parcialmente"*; *"Não Concordo, nem Discordo"*; *"Concordo Parcialmente"* e *"Concordo Totalmente"*, representadas na figura como DT, DP, NCND e CT, respectivamente. Verifica-se que o K-Means tem resultados similares ao LDA enquanto o PLSA se mostrou menos eficiente nesse critério uma vez que mais avaliadores rejeitaram a afirmação de que todos os segmentos tratam de um único assunto em comum.

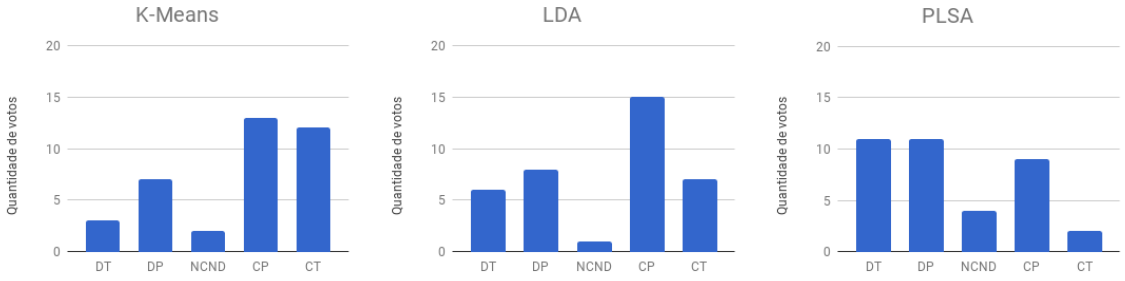


Figura 18 – Contagem de respostas referente a primeira questão cujo enunciado foi: *"Todos os trechos apresentados compartilham um mesmo assunto."*. O eixo vertical indica a frequência das alternativas representadas no eixo horizontal.

Outro ponto importante a ser analisado é a capacidade representativa dos descritores, ou seja, o quão bem os descritores podem representar o tópico ao qual os segmentos foram atribuídos. A Figura 19 contém a frequência as respostas referentes a segunda questão, onde os avaliadores tiveram as mesmas opções de respostas que a anterior. Observa-se na Figura 19 que no caso do K-Means a quantidade de usuários que concordam com a

afirmação de que os descritores extraídos são bons atributos para descrever o teor dos segmentos foi bem maior em relação ao que discordam.

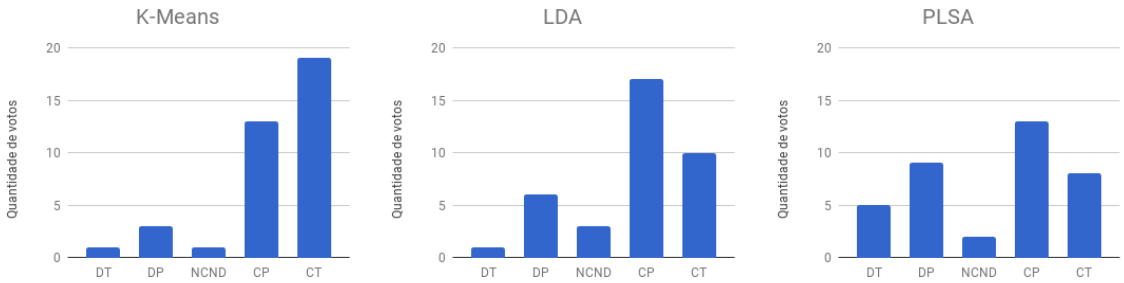


Figura 19 – Contagem de respostas referente a segunda questão cujo enunciado foi: *"As palavras <descritores> resumem bem o assunto tratado nos trechos."*. O eixo vertical indica a frequência das alternativas representadas no eixo horizontal.

Ao analisar os resultados, verifica-se que de maneira geral os modelos K-Means e LDA podem ser considerados satisfatórios na tarefa de agrupar e representar os segmentos das atas. Verifica-se também que para o modelo K-Means os avaliadores identificaram que, na maioria dos casos, houve resultados satisfatórios, principalmente quanto a representatividade dos descritores. Embora a avaliação aponte imperfeições, esse modelo apresenta maior constância em relação aos demais.

5.4 Validação do Segmentador

Nessa seção, os dados coletados no experimento são utilizados para validar o segmentador empregado nesse trabalho. O algoritmo *BayesSeg* é analisado quanto a sua capacidade de extrair segmentos coesos, isto é, que tratem de um único assunto central evitando trazer informações alheias ou irrelevantes. Outro critério avaliado é a completude dos segmentos, ou seja, o segmentos devem conter informações suficientes para o entendimento do texto sem necessidade dos textos adjacentes. Em outras palavras, usuário deve receber informações precisas sobre o tópico selecionado.

Os dados coletados de ambas consultas foram somadas, uma vez que esta é uma etapa anterior à extração de tópicos, a princípio, o modelo que selecionou os segmentos não interfere em sua avaliação. Assim, a Figura 20 mostra as respostas do avaliadores considerando todos os cenários. As respostas referentes a terceira questão, na qual se averígua a homogeneidade de cada segmento quanto ao seu assunto central, apontam que poucos segmentos contém mais de um assunto.

Ainda sobre a qualidade da segmentação, a Figura 21 mostra os resultados da quarta questão a qual investiga a integridade de cada segmento, isto é, sua capacidade de informar o usuário sobre o assunto que trata sem necessidade de se recorrer a leitura

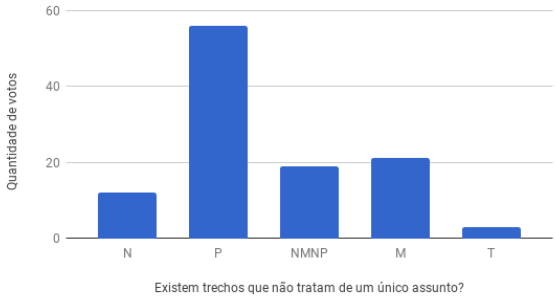


Figura 20 – Contagem de respostas referente a terceira questão cujo enunciado foi: *"Existem trechos que não tratam de um único assunto?"*. O eixo vertical indica a frequência das alternativas representadas no eixo horizontal.

do documento integral. Nesse critério, a maioria das avaliações indicam que nenhum ou poucos segmentos apresentam texto insuficiente para leitura. Uma análise mais detalhada das questões relacionadas a segmentação das atas foi discutida no Capítulo 4, ficando aqui análises de pontos onde a segmentação influencia os extratores e os resultados finais apresentados ao usuário.

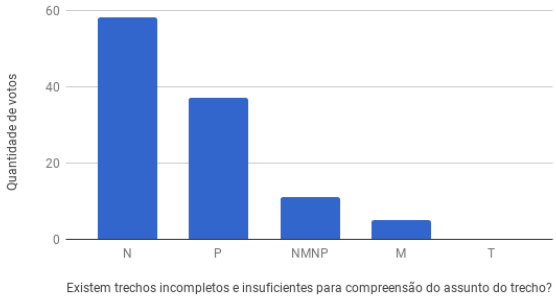


Figura 21 – Contagem de respostas referente a quarta questão cujo enunciado foi: *"Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?"*. O eixo vertical indica a frequência das alternativas representadas no eixo horizontal.

Outra questão analisada foi o comportamento dos modelos nas diferentes consultas. Ao se isolar as respostas das questões referentes a uma consulta específica, nota-se certa alteração nas respostas dos modelos. Os gráficos apresentados na Figura 22 mostram na linha superior as respostas para cada modelo considerando-se os segmentos extraídos na primeira consulta e na linha inferior aqueles referentes à segunda consulta. O K-Means apresenta uma diminuição considerável na segunda consulta em relação à primeira no que se refere a respostas afirmando que os todos os segmentos compartilham um único assunto, e um aumento de respostam indicando discordância total com a afirmação da questão. De forma semelhante, o PLSA apresenta diminuição de respostas positivas para essa afirmação na proporção que há um aumento de respostas negativas. Por outro lado o LDA mantém resultados semelhantes em ambas as consultas, nas quais apresenta resultados equilibrados

entre respostas positivas e negativas. Em outras palavras, os modelos K-Means e PLSA sofreram perda de desempenho enquanto o LDA manteve-se estável em ambas consultas.

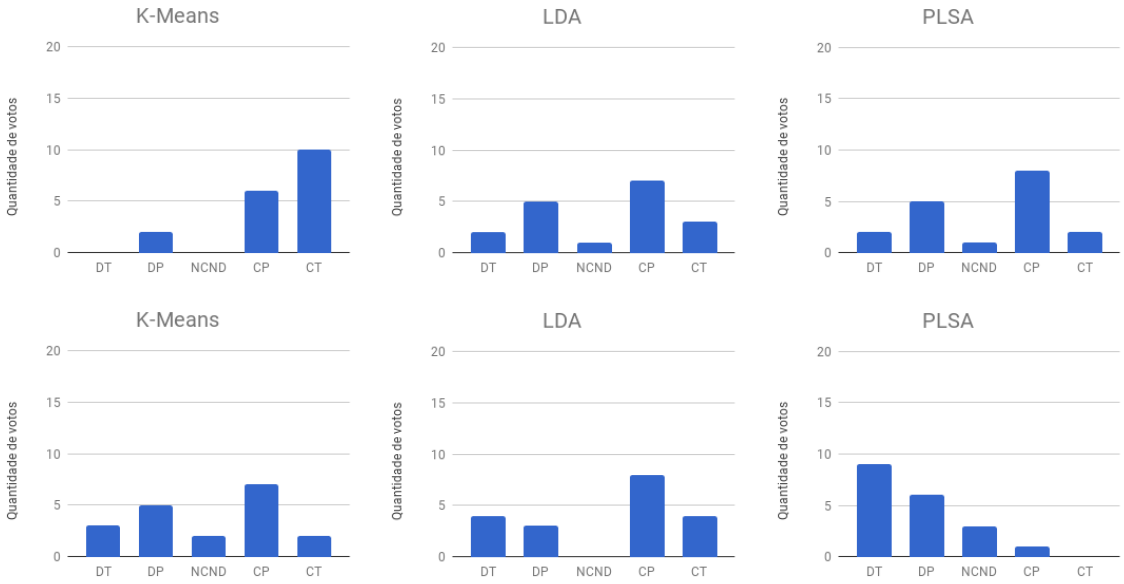


Figura 22 – Contagem das respostas referentes a Primeira questão. A primeira consulta, "*compra de equipamentos*", é mostrada na linha superior e a segunda consulta, "*defesa de dissertação*", na linha inferior.

Ao analisar separadamente a segunda questão, referente a representatividade dos descritores observa-se na Figura 23 que todos os modelos apresentam perda de performance na segunda consulta em relação à primeira, contudo, de forma acentuada no PLSA.

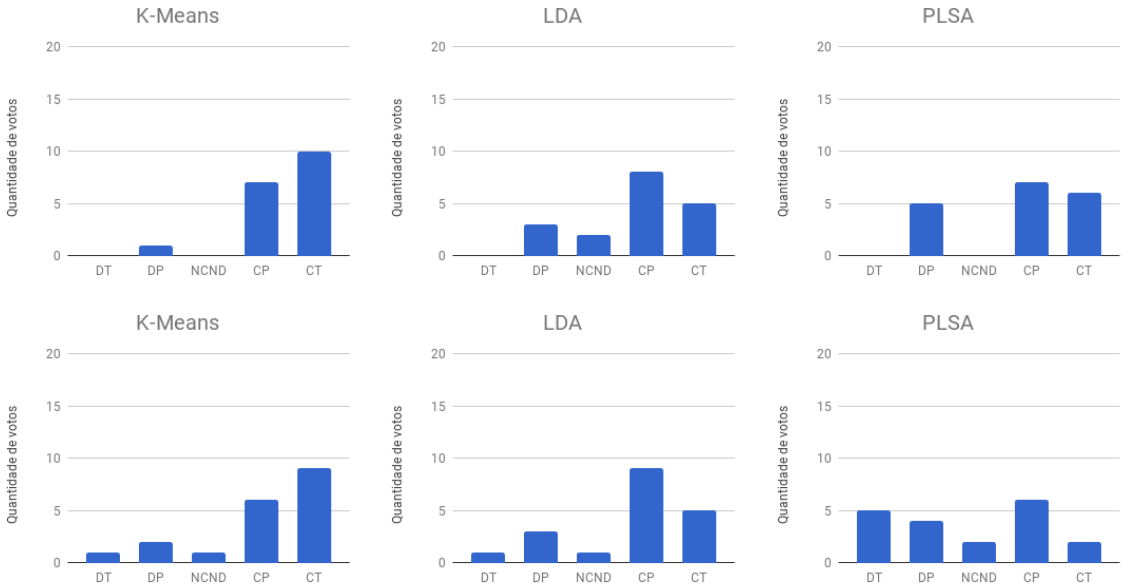


Figura 23 – Contagem das respostas referentes a Segunda questão. A primeira consulta, "*compra de equipamentos*", é mostrada na linha superior e a segunda consulta, "*defesa de dissertação*", na linha inferior.

5.5 Influência dos Extratores na Qualidade dos Segmentos

Analisou-se a influência das técnicas de extração de tópicos sobre a qualidade dos segmentos apresentados, quanto aos critérios de coesão e completude já mencionados. Vale lembrar que embora a extração de tópicos seja uma etapa posterior à extração, esta influencia na seleção dos segmentos apresentados ao usuário. A Figura 24 apresenta as contagens das respostas da terceira questão, *"Existem trechos que não tratam de um único assunto?"*, considerando-se cada extrator separadamente. De forma semelhante, a Figura 25, apresenta os dados da quarta resposta, *"Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?"*.

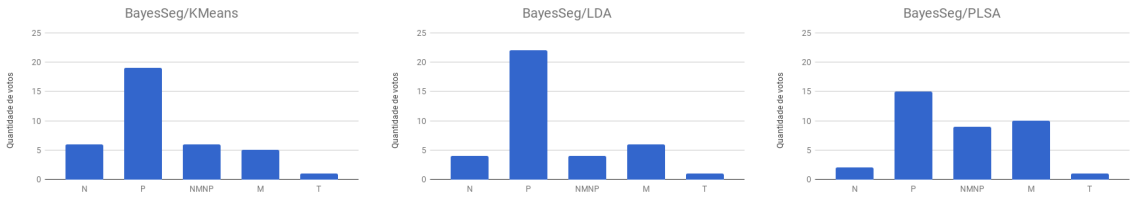


Figura 24 – Contagem das respostas referentes a terceira questão, isolando-se as técnicas de extração de tópicos.

Verifica-se que os algoritmos K-Means e LDA selecionam segmentos igualmente coesos e pouco diferem entre si e em relação ao apresentado por todos os extratores, conforme mostrado na Figura 20, em que a maioria dos avaliadores informou que poucos segmentos não tratam de um único assunto. Sob o mesmo critério, os segmentos selecionados pelo PLSA são considerados menos coesos em relação aos demais modelos.

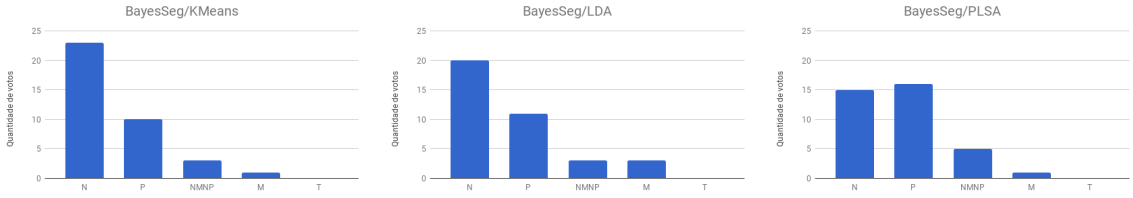


Figura 25 – Contagem das respostas referentes a quarta questão, isolando-se as técnicas de extração de tópicos.

Quanto a completude dos segmentos selecionados, observa-se na Figura 25 um comportamento similar entre os extratores, sendo o K-Means e LDA semelhantes e com pouca discrepância em relação aos resultados analisados na Figura 21. O PLSA tente a tolerar segmentos menos coesos e menos completos enquanto os demais favorecem aqueles com maior qualidade de acordo com os quesitos abordados nesse experimento.

5.6 Considerações Finais

Nessa seção, analisou-se a validação do BayesSeg como segmentador textual e as técnicas de extração de tópicos na criação de uma estrutura derivada do *corpus* original como uma representação estruturada da coleção de atas a qual foi organizada e acrescida de atributos para sua descrição. As análises sugerem que tais técnicas podem oferecer a sistemas de recuperação uma representação estruturada que preserva o conteúdo dos documentos ao mesmo tempo que cria atributos adicionais que incorporam informação à base de dados e podem ser inseridas no espaço de busca.

De maneira geral, os dados apontam que o uso do K-Means como extrator de tópicos se sobressai em relação ao LDA e PLSA, principalmente quanto a eficiência do modelo em identificar bons descritores para representar os grupos, mostrando-se assim uma boa eficiência na tarefa de agrupar e descrever os segmentos.

Na validação do segmentador empregado no sistema, o *BayesSeg*, consegue satisfazer os principais quesitos relacionados a um segmentador, contudo, o sistema apresenta resultados melhores quando os segmentos são selecionados por meio do K-Means e LDA, o que sugere que melhores configurações para o PLSA podem ser analisadas em outros experimentos.

Os dados coletados forneceram uma base para as análises onde se verificou os pontos principais de cada técnica bem como a relação entre os modelos de extração de tópicos na seleção dos segmentos. Outros critérios podem ser levados em conta visto a subjetividade das avaliações bem como mais experimentos são necessários para obter melhores resultados.

6 Conclusão

Com a grande disponibilidade de dados em formato textual, há um constante interesse no desenvolvimento de ferramentas computacionais para recuperação automática de informações úteis a partir desses dados. Dentre os textos usados para registros, identificou-se documentos multi-temáticos, os quais abordam assuntos diversos no mesmo texto, como exemplo, as atas de reuniões apresentam como característica a ausência de meta informações ou mesmo quebras de parágrafos que ajudariam a separar e identificar seus conteúdos. Entre as abordagens aplicáveis a este problema estão a segmentação textual e os modelos de extração de tópicos as quais em conjunto são capazes de separar, identificar e agrupar os assuntos contidos nessa categoria de documentos.

Dados os desafios apresentados acima, este trabalho de mestrado visou o desenvolvimento de um Sistema de Recuperação de Informação utilizando técnicas de Segmentação Textual e Extração de Tópicos para extração automática de conhecimento em uma base de dados composta por atas de reunião coletadas da Universidade Federal de São Carlos, campus Sorocaba. Desenvolveu-se uma metodologia que utiliza a segmentação textual para fragmentar as atas em porções de texto com um assunto relativamente independente os quais são agrupados e descritos semanticamente por um modelo de extração de tópicos, conforme apresentado no Capítulo 3.

Inicialmente, a fim de analisar as técnicas utilizadas, gerou-se um *corpus* anotado por 9 profissionais com afinidade com o domínio investigado. Os anotadores segmentaram e rotularam manualmente 12 atas de reunião. Para isso desenvolveu-se uma ferramenta para anotação manual de documentos, pela qual os anotadores forneceram segmentações e informações sobre os segmentos identificados. Os textos originais acrescidos das informações fornecidas pelos anotadores foram reunidas para formar um *corpus* derivado que ajuda a entender o *corpus* investigado em termos de sua distribuição de tópicos. Os dados referentes a segmentação manual foram utilizados para criar uma segmentação de referência para avaliação objetiva dos segmentadores.

A avaliação dos segmentadores considerou os algoritmos e seus principais parâmetros para encontrar o modelo que melhor otimiza a tarefa de segmentação do *corpus* investigado. Comparou-se os resultados obtidos com a segmentação de referência e verificou-se que o algoritmo *BayesSeg* apresenta resultados melhores em relação às demais técnicas analisadas. Além disso, a escolha escolheu-se o *BayesSeg* devido ao sua abordagem probabilísticas similar a modelos de extração de tópicos como o LDA. Os resultados mostram certa dificuldade devido as características das atas, como estilo de escrita e segmentos relativamente curtos, além da subjetividade intrínseca da tarefa. Embora os métodos de segmentação

textual tenham se mostrado suficientes, melhores resultados podem ser alcançados com o acréscimo das técnicas mencionadas e a construção de uma segmentação de referência mais robusta em termos concordância entre os anotadores que ajudaram a criá-la. Uma vez escolhido o segmentador, este foi utilizado para segmentar os textos de um conjunto de 175 atas, que gerou um conjunto de 1276 segmentos, os quais foram submetidos aos modelos de extração de tópicos. Detalhes sobre o *corpus* anotado e avaliação objetiva dos segmentadores foram analisados no Capítulo 4.

Após a avaliação dos segmentadores e segmentação do *corpus* inicial, o desenvolvimento prosseguiu com a avaliação dos extratores de tópicos. Os extratores foram executados para agrupar e extrair descritores do conjunto de segmentos. Cada extrator formou 70 grupos (tópicos) e para cada grupo foram extraídos 5 descritores.

A metodologia utilizada nesse trabalho conecta as técnicas de segmentação textual aos modelos de extração de tópicos a fim de gerar um estrutura derivada a partir de um *corpus* não estruturado. Essa nova estrutura concentra os textos originais acrescidos de informação latentes e organizados por sua semelhança semântica. Essa organização permite que técnicas de recuperação de informação expandam o espaço de busca além do conjunto de termos original de cada segmento, sendo assim, favorecidas quanto a identificação segmentos relacionados a consulta do usuário bem como permite a exploração dos grupos com assuntos relacionados.

Os modelos de extração de tópicos foram avaliados subjetivamente por meio de questionários em que 41 avaliadores responderam à questões referentes à qualidade dos trechos apresentados como resultado à consultas à base de dados criada com essa metodologia. A fim de avaliar 3 dos principais modelos de extração de tópicos e o segmentador escolhido (*BayesSeg*), o sistema foi submetido a duas consultas distintas, gerando 6 cenários a serem avaliados. Os questionários foram formulados para obter primeiramente a percepção dos avaliadores quanto a qualidade dos tópicos extraídos, levando em conta a similaridade dos segmentos em relação ao mesmo assunto e a representatividade dos descritores. Os avaliadores também responderam questões referentes a qualidade dos segmentos apresentados, uma vez que a segmentação, embora etapa anterior, é parte do processo de obtenção de conhecimento. Sobre a segmentação, considerou-se a completude dos segmentos a sua unidade em relação ao assunto. Os dados dos questionários após coletados e analisados, sugerem que a metodologia empregada é capaz de entregar ao usuário resultados satisfatórios. Verificou-se também que o K-Means traz melhores resultados em relação aos demais extratores avaliados. A análise detalhada dessas avaliações e a relação entre as técnicas empregadas nessa metodologia foram apresentadas no Capítulo 5.

De modo geral, o sistema desenvolvido nesse trabalho recebe uma base de dados constituída por documentos multi-temáticos não estruturada e produz uma estrutura de dados interna mais organizada e acrescida de informações latentes extraídas do próprio

corpus original de forma não supervisionada. A utilização de descritores para expandir a o espaço de busca pelas técnicas de recuperação de informação possibilita ganho em relação a busca indexada por termos presentes nos documentos. Além disso, o retorno procura exibir apenas os trechos relevantes à consulta do usuário, ao invés de documentos inteiros que podem conter trechos irrelevantes à consulta. A estrutura dada ao *corpus* original permite analisar seu conteúdo pela perspectiva da distribuição dos tópicos. Com isso, é possível entender a composição e recorrência dos assuntos discutidos bem como traçar um histórico dos assuntos abordados a fim de visualizar sua evolução ao longo do tempo.

6.1 Contribuições

Considera-se como principais contribuições deste trabalho o método apresentado para extração de conhecimento em documentos multi-temáticos, o *corpus* de atas anotadas, o sistema proposto e sua implementação, as avaliações dos segmentadores e dos extratores de tópicos e os resultados produzidos durante a execução da proposta desse sistema.

A metodologia proposta permite organizar e adicionar informação à coleção de documentos multi-temáticos, com a finalidade de extrair automaticamente conhecimento em atas de reunião. Essa metodologia pode ainda ser aproveitada para trazer avanços em estudos e aplicações voltadas à extração e recuperação de informação em outros domínios com as mesmas características.

O *corpus* anotado gerado neste projeto contém dados adicionados aos documentos originais os quais ajudam a entender a distribuição dos tópicos ao longo das atas. O *corpus* anotado está disponível para servir como base de dados para outros trabalhos, bem como a ferramenta desenvolvida para esse propósito, a qual pode ser utilizada em qualquer *corpus* a ser anotado com fins de pesquisa, e pode ser modificada livremente.

O sistema proposto e sua implementação podem ser utilizados para extração de conhecimento em bases de dados formadas por documentos multi-temáticos, em especial, conjuntos de atas de reunião. Outras domínios que lidam com bases de dados com as mesmas características podem aproveitar esse sistema bem como seu código fonte para eventuais adaptações ou expansões.

As avaliações dos segmentadores mostraram dados objetivos sobre o desempenho dessas técnicas quando aplicadas à coleção de documentos utilizada nesse trabalho. A avaliação dos extratores apresentou a percepção de profissionais sobre os resultados da combinação entre o segmentador empregado e os extratores de tópicos.

6.2 Trabalhos Futuros

Entre as continuações e futuras melhorias para este trabalho, pode-se citar implementações no próprio sistema proposto e ampliação das bases de dados bem como inclusão de novos *corpora*. Os primeiros experimentos desse trabalho geraram um artigo que foi submetido para publicação, porém, não foi aceito. Essa recusa exigiu novos experimentos com mais dados os quais são apresentados nessa dissertação. Assim, um novo artigo será submetido em complemento à esse trabalho. A proposta inicial desse trabalho contempla a classificação dos segmentos em relação ao tipo de menção ao assunto, como decisão, orientação e irrelevante. Os dados colhidos no experimento mencionado na Seção 2.3.3 referentes ao tipo de menção e contexto do assunto serão utilizados para o gerar um classificador para categorizar automaticamente um segmento nas categorias. Para isso, as anotações coletadas podem ser utilizadas na etapa de treinamento dos classificadores.

Outras melhorias do sistema podem ser alcançadas com testes voltados a experiência do usuário a fim de medir a satisfação dos resultados apresentados e guiar a implementação de uma interface gráfica para possibilitar ao usuário analisar os dados pela exploração interativa dos grupos. Além disso, a implementação de algoritmos de agrupamento incremental devem ser incorporadas ao sistema a fim de suportar adequadamente o crescimento da coleção de documentos.

Cita-se também a utilização de fontes externas para melhorar os métodos de segmentação textual. Recursos como *thesaurus* (dicionários de sinônimos) e *clue words* (palavras-pista) podem adicionar conhecimento externo ao sistema e com isso alcançar melhores resultados. Pretende-se ainda replicar os experimentos deste trabalho em novas bases de dados com características semelhantes às ata de reunião. Por exemplo, transcrições de conversas, diálogos em *chats,* discursos e atas de outras organizações como instituições públicas e governamentais. Inclui-se também a implementação de técnicas tradicionais de Recuperação de Informação, a como base lines a fim de melhor comparar a metodologia apresentada nesse projeto.

Referências

AGGARWAL, C. C. *Machine Learning for Text*. [S.l.]: Springer International Publishing, 2018. ISBN 978-3-319-73531-3. Citado 5 vezes nas páginas 11, 23, 30, 37 e 71.

AGGARWAL, C. C.; ZHAI, C. X. *Mining Text Data*. [S.l.]: Springer Publishing Company, Incorporated, 2012. ISBN 1461432227, 9781461432227. Citado 2 vezes nas páginas 9 e 10.

BEEFERMAN, D.; BERGER, A.; LAFFERTY, J. Statistical models for text segmentation. *Machine Learning*, v. 34, n. 1, p. 177–210, 1999. ISSN 1573-0565. Disponível em: <http://dx.doi.org/10.1023/A:1007506220214>. Citado 2 vezes nas páginas 31 e 32.

BLEI, D. M. Probabilistic topic models. *Commun. ACM*, ACM, New York, NY, USA, v. 55, n. 4, p. 77–84, abr. 2012. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2133806.2133826>. Citado 2 vezes nas páginas 15 e 40.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, JMLR.org, v. 3, p. 993–1022, mar. 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944937>. Citado 3 vezes nas páginas 28, 37 e 39.

BOKAEI, M. H.; SAMETI, H.; LIU, Y. Linear discourse segmentation of multi-party meetings based on local and global information. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, IEEE Press, Piscataway, NJ, USA, v. 23, n. 11, p. 1879–1891, nov. 2015. ISSN 2329-9290. Disponível em: <http://dx.doi.org/10.1109/TASLP.2015.2456430>. Citado 3 vezes nas páginas 11, 23 e 25.

BOKAEI, M. H.; SAMETI, H.; LIU, Y. Extractive summarization of multiparty meetings through discourse segmentation. *Natural Language Engineering*, Cambridge University Press, v. 22, n. 1, p. 41–72, 2016. Citado na página 47.

BUI, Q. V. et al. Combining latent dirichlet allocation and k-means for documents clustering: Effect of probabilistic based distance measures. In: NGUYEN, N. T. et al. (Ed.). *Intelligent Information and Database Systems*. Cham: Springer International Publishing, 2017. p. 248–257. ISBN 978-3-319-54472-4. Citado na página 41.

CAO, L. Data science: A comprehensive overview. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 50, n. 3, p. 43:1–43:42, jun. 2017. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/3076253>. Citado na página 9.

CARDOSO, P.; PARDO, T.; TABOADA, M. Subtopic annotation and automatic segmentation for news texts in brazilian portuguese. *Corpora*, Edinburgh University Press, v. 12, n. 1, p. 23–54, 2017. Citado 5 vezes nas páginas 25, 37, 58, 60 e 62.

CARLETTA, J. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 22, n. 2, p. 249–254, jun. 1996. ISSN 0891-2017. Disponível em: <http://dl.acm.org/citation.cfm?id=230386.230390>. Citado 2 vezes nas páginas 36 e 62.

CHAIBI, A. H.; NAILI, M.; SAMMOUD, S. Topic segmentation for textual document written in arabic language. *Procedia Computer Science*, v. 35, p. 437 – 446, 2014. ISSN 1877-0509. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050914010898>. Citado 2 vezes nas páginas 25 e 35.

CHENG, X. et al. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In: *SDM*. SIAM, 2013. p. 749–757. ISBN 978-1-61197-283-2. Disponível em: <http://dblp.uni-trier.de/db/conf/sdm/sdm2013.html#ChengGLWY13>. Citado na página 38.

CHOI, F. Y. Y. Advances in domain independent linear text segmentation. In: *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. p. 26–33. Disponível em: <http://dl.acm.org/citation.cfm?id=974305.974309>. Citado 3 vezes nas páginas 25, 26 e 35.

CHOI, F. Y. Y.; WIEMER-HASTINGS, P.; MOORE, J. Latent semantic analysis for text segmentation. In: *In Proceedings of EMNLP*. [S.l.: s.n.], 2001. p. 109–117. Citado 2 vezes nas páginas 23 e 51.

CROFT, B.; METZLER, D.; STROHMAN, T. *Search Engines: Information Retrieval in Practice*. 1st. ed. USA: Addison-Wesley Publishing Company, 2009. ISBN 0136072240, 9780136072249. Citado 4 vezes nas páginas 10, 17, 18 e 19.

DEERWESTER, S. et al. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, v. 41, n. 6, p. 391–407, 1990. Citado 2 vezes nas páginas 15 e 38.

DHILLON, I. S.; MODHA, D. S. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 42, n. 1-2, p. 143–175, jan. 2001. ISSN 0885-6125. Disponível em: <https://doi.org/10.1023/A:1007612920971>. Citado na página 42.

DIAS, G.; ALVES, E.; LOPES, J. G. P. Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In: *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2*. AAAI Press, 2007. (AAAI'07), p. 1334–1339. ISBN 978-1-57735-323-2. Disponível em: <http://dl.acm.org/citation.cfm?id=1619797.1619859>. Citado na página 25.

EISENSTEIN, J.; BARZILAY, R. Bayesian unsupervised topic segmentation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. (EMNLP '08), p. 334–343. Disponível em: <http://dl.acm.org/citation.cfm?id=1613715.1613760>. Citado 5 vezes nas páginas 11, 23, 26, 28 e 63.

FALEIROS, T. d. P. *Propagação em grafos bipartidos para extração de tópicos em fluxo de documentos textuais*. Tese (Doutorado), São Paulo, SP, Brasil, 2016. Citado 3 vezes nas páginas 11, 37 e 40.

FELDMAN, R.; SANGER, J. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York, NY, USA: Cambridge University Press, 2006. ISBN 0521836573, 9780521836579. Citado na página 23.

FERRET, O. Improving text segmentation by combining endogenous and exogenous methods. In: *International Conference Recent Advances in Natural Language Processing, RANLP*. [S.l.: s.n.], 2009. p. 88–93. Citado 2 vezes nas páginas 63 e 64.

GALLEY, M. et al. Discourse segmentation of multi-party conversation. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (ACL '03), p. 562–569. Disponível em: <http://dx.doi.org/10.3115/1075096.1075167>. Citado 3 vezes nas páginas 25, 34 e 60.

GRUENSTEIN, A.; NIEKRASZ, J.; PURVER, M. *MEETING STRUCTURE ANNOTATION – Annotations Collected with a General Purpose Toolkit*. [S.l.]: Springer, Dordrecht, 2007. Citado 2 vezes nas páginas 36 e 62.

GURUNG, P.; WAGH, R. A study on topic identification using k means clustering algorithm: Big vs. small documents. In: . [S.l.]: Research India Publications, 2017. p. 221–233. ISBN 0973-6107. Citado na página 41.

GUTIERREZ, F. et al. A hybrid ontology-based information extraction system. *J. Inf. Sci.*, Sage Publications, Inc., Thousand Oaks, CA, USA, v. 42, n. 6, p. 798–820, dez. 2016. ISSN 0165-5515. Disponível em: <https://doi.org/10.1177/0165551515610989>. Citado 4 vezes nas páginas 10, 16, 63 e 64.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, JMLR.org, v. 3, p. 1157–1182, mar. 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944968>. Citado na página 37.

HASSANI, M.; SEIDL, T. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. *Vietnam Journal of Computer Science*, v. 4, n. 3, p. 171–183, Aug 2017. ISSN 2196-8896. Disponível em: <https://doi.org/10.1007/s40595-016-0086-9>. Citado na página 71.

HEARST, M. A. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 23, n. 1, p. 33–64, mar. 1997. ISSN 0891-2017. Disponível em: <http://dl.acm.org/citation.cfm?id=972684.972687>. Citado 3 vezes nas páginas 34, 60 e 62.

HOFMANN, T. Probabilistic latent semantic indexing. In: *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 1999. (SIGIR '99), p. 50–57. ISBN 1-58113-096-1. Disponível em: <http://doi.acm.org/10.1145/312624.312649>. Citado 3 vezes nas páginas 15, 37 e 39.

HOVY, E.; LAVID, J. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. v. 22, p. 13–36, 01 2010. Citado 4 vezes nas páginas 34, 35, 57 e 58.

HUANG, X. et al. Applying machine learning to text segmentation for information retrieval. *Information Retrieval*, v. 6, n. 3, p. 333–362, Sep 2003. ISSN 1573-7659. Disponível em: <https://doi.org/10.1023/A:1026028229881>. Citado na página 12.

JANIN, A. et al. The icsi meeting corpus. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. [S.l.: s.n.], 2003. p. 364–367. Citado na página 34.

JEONG, M.; TITOV, I. Multi-document topic segmentation. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2010. (CIKM '10), p. 1119–1128. ISBN 978-1-4503-0099-5. Disponível em: <http://doi.acm.org/10.1145/1871437.1871579>. Citado 3 vezes nas páginas 9, 12 e 42.

KAZANTSEVA, A.; SZPAKOWICZ, S. Topical segmentation: A study of human performance and a new measure of quality. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. (NAACL HLT '12), p. 211–220. ISBN 978-1-937284-20-6. Disponível em: <http://dl.acm.org/citation.cfm?id=2382029.2382058>. Citado 4 vezes nas páginas 31, 34, 37 e 60.

KERN, R.; GRANITZER, M. Efficient linear text segmentation based on information retrieval techniques. *Proceedings of the International Conference on Management of Emergent Digital EcoSystems, MEDES '09*, p. 167–171, 2009. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-74549147972&doi=10.1145%2f1643823.1643854&partnerID=40&md5=1c6f73bc0e07446fcc178440e48bbc40>. Citado 2 vezes nas páginas 25 e 32.

KOZIMA, H. Text segmentation based on similarity between words. In: *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1993. (ACL '93), p. 286–288. Disponível em: <http://dx.doi.org/10.3115/981574.981616>. Citado na página 24.

LEE, D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. v. 401, p. 788–91, 11 1999. Citado 2 vezes nas páginas 15 e 38.

LEE, J.-K. et al. Two-step sentence extraction for summarization of meeting minutes. In: . [S.l.: s.n.], 2011. v. 39, p. 614–619. Citado na página 9.

LUHN, H. P. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, IBM Corp., Riverton, NJ, USA, v. 2, n. 2, p. 159–165, abr. 1958. ISSN 0018-8646. Disponível em: <http://dx.doi.org/10.1147/rd.22.0159>. Citado na página 20.

MALIOUTOV, I.; BARZILAY, R. Minimum cut model for spoken lecture segmentation. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (ACL-44), p. 25–32. Disponível em: <https://doi.org/10.3115/1220175.1220179>. Citado 3 vezes nas páginas 28, 29 e 30.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715. Citado 8 vezes nas páginas 9, 10, 16, 18, 19, 30, 40 e 41.

MARCACINI, R. M.; REZENDE, S. O. An introduction to variable and feature selection. *WFA'2010: IX Workshop de Ferramentas e Aplicações. Em conjunto com o XVI Simpósio*

Brasileiro de Sistemas Multimídia e Web(Webmedia), v. 3, p. 1–3, 2010. Disponível em: <http://sites.labic.icmc.usp.br/torch/>. Citado 5 vezes nas páginas 11, 12, 20, 23 e 37.

MASAO, U.; KôITI, H. Multi-topic multi-document summarization. In: *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000. (COLING '00), p. 892–898. Disponível em: <https://doi.org/10.3115/992730.992775>. Citado na página 41.

MISRA, H. et al. Text segmentation via topic modeling: An analytical study. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2009. (CIKM '09), p. 1553–1556. ISBN 978-1-60558-512-3. Disponível em: <http://doi.acm.org/10.1145/1645953.1646170>. Citado 3 vezes nas páginas 11, 23 e 25.

NAILI, M.; CHAIBI, A. H.; GHEZALA, H. H. B. Exogenous approach to improve topic segmentation. *International Journal of Intelligent Computing and Cybernetics*, v. 9, n. 2, p. 165–178, 2016. Disponível em: <https://doi.org/10.1108/IJICC-01-2016-0001>. Citado 3 vezes nas páginas 25, 63 e 64.

NGUYEN, V. C. *A Study on Statistical Generation of a Hierarchical Structure of Topic-information for Multi-documents*. Tese (Doutorado) — School of Information Science Japan Advanced Institute of Science and Technology, 2011. Disponível em: <http://hdl.handle.net/10119/12059>. Citado 2 vezes nas páginas 15 e 42.

NORMAN, G. R. Likert scales, levels of measurement and the "laws"of statistics. *Advances in health sciences education : theory and practice*, v. 15 5, p. 625–32, 2010. Citado na página 73.

O'CALLAGHAN, D. et al. An analysis of the coherence of descriptors in topic modeling. *Expert Syst. Appl.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 42, n. 13, p. 5645–5657, ago. 2015. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2015.02.055>. Citado 2 vezes nas páginas 37 e 71.

PASSONNEAU, R. J.; LITMAN, D. J. Discourse segmentation by human and automated means. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 23, n. 1, p. 103–139, mar. 1997. ISSN 0891-2017. Disponível em: <http://dl.acm.org/citation.cfm?id=972684.972689>. Citado 2 vezes nas páginas 34 e 60.

PEVZNER, L.; HEARST, M. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, v. 28, n. 1, p. 19–36, 2002. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-0037870455&doi=10.1162%2f089120102317341756&partnerID=40&md5=279abc4e76fcfc2c4a1896e76a245034>. Citado na página 33.

PORTER, M. F. Readings in information retrieval. In: JONES, K. S.; WILLETT, P. (Ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. cap. An Algorithm for Suffix Stripping, p. 313–316. ISBN 1-55860-454-5. Disponível em: <http://dl.acm.org/citation.cfm?id=275537.275705>. Citado na página 46.

PRINCE, V.; LABADIÉ, A. Text segmentation based on document understanding for information retrieval. In: KEDAD, Z. et al. (Ed.). *Natural Language Processing and Information Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 295–304. ISBN 978-3-540-73351-5. Citado na página 12.

REYNAR, J. C. *Topic Segmentation: Algorithms and Applications*. Tese (Doutorado), Philadelphia, PA, USA, 1998. AAI9829978. Citado 2 vezes nas páginas 24 e 27.

REZENDE, S. O. *Sistemas Inteligentes*. Barueri, SP: Manole, 2003. 337 - 270 p. Citado 3 vezes nas páginas 16, 21 e 22.

RIJSBERGEN, C. J. V. *Information Retrieval*. 2nd. ed. Newton, MA, USA: Butterworth-Heinemann, 1979. ISBN 0408709294. Citado 2 vezes nas páginas 18 e 19.

ROSSI, R. G.; REZENDE, S. O. Building a topic hierarchy using the bag-of-related-words representation. In: *Proceedings of the 11th ACM Symposium on Document Engineering*. New York, NY, USA: ACM, 2011. (DocEng '11), p. 195–204. ISBN 978-1-4503-0863-2. Disponível em: <http://doi.acm.org/10.1145/2034691.2034733>. Citado na página 41.

SAKAHARA, M.; OKADA, S.; NITTA, K. Domain-independent unsupervised text segmentation for data management. In: *2014 IEEE International Conference on Data Mining Workshop*. [S.l.: s.n.], 2014. p. 481–487. ISSN 2375-9232. Citado 2 vezes nas páginas 11 e 23.

SALTON, G.; ALLAN, J. Automatic text decomposition and structuring. In: *Intelligent Multimedia Information Retrieval Systems and Management - Volume 1*. Paris, France, France: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 1994. (RIAO '94), p. 6–20. Disponível em: <http://dl.acm.org/citation.cfm?id=2856823.2856826>. Citado na página 17.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 24, n. 5, p. 513–523, ago. 1988. ISSN 0306-4573. Disponível em: <http://dx.doi.org/10.1016/0306-4573(88)90021-0>. Citado na página 17.

SANTOS, F. F. dos; CARVALHO, V. O. de; REZENDE, S. O. Selecting candidate labels for hierarchical document clusters using association rules. In: *Advances in Soft Computing - Lecture Notes in Computer Science*. Berlin: Springer, 2010. ISBN 9783642167720. Disponível em: <http://www.springerlink.com/content/m83l718tk1162840>. Citado na página 41.

SCHWARTZ-ZIV, M.; WEISBACH, M. S. What do boards really do? evidence from minutes of board meetings. *Journal of Financial Economics*, v. 108, n. 13, p. 349–366, 2013. Citado na página 9.

SHAMSINEJADBABKI, P.; SARAEE, M. A new unsupervised feature selection method for text clustering based on genetic algorithms. *J. Intell. Inf. Syst.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 38, n. 3, p. 669–684, jun. 2012. ISSN 0925-9902. Disponível em: <http://dx.doi.org/10.1007/s10844-011-0172-5>. Citado na página 17.

SHI, J.; MALIK, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 8, p. 888–905, Aug 2000. ISSN 0162-8828. Citado na página 29.

SHROUT, P. E.; FLEISS, J. L. Intraclass correlations: Uses in assessing rater reliability. v. 86, p. 420–8, 04 1979. Citado na página 36.

SOARES, M. V. B.; PRATI, R. C.; MONARD, M. C. *PreTexT II: descrição da reestruturação da ferramenta de pré-processamento de textos*. [S.l.], 2008. Citado na página 20.

STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. In: LANDAUER, T.; MCNAMARA, S. D.; KINTSCH, W. (Ed.). *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2007. cap. Probabilistic topic models. Disponível em: <http://psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf>. Citado 2 vezes nas páginas 37 e 71.

TAGARELLI, A.; KARYPIS, G. A segment-based approach to clustering multi-topic documents. *Knowledge and Information Systems*, v. 34, n. 3, p. 563–595, Mar 2013. ISSN 0219-3116. Disponível em: <https://doi.org/10.1007/s10115-012-0556-z>. Citado 4 vezes nas páginas 9, 12, 15 e 42.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367. Citado 2 vezes nas páginas 18 e 23.

UTIYAMA, M.; ISAHARA, H. A statistical model for domain-independent text segmentation. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2001. (ACL '01), p. 499–506. Disponível em: <https://doi.org/10.3115/1073012.1073076>. Citado na página 27.

WEI, X. *TOPIC MODELS IN INFORMATION RETRIEVAL*. Tese (Doutorado) — University of Massachusetts Amherst, Massachusetts, MA, USA, 2007. Citado 2 vezes nas páginas 11 e 37.

WEI, X.; CROFT, W. B. Lda-based document models for ad-hoc retrieval. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2006. p. 178–185. ISBN 1-59593-369-7. Disponível em: <http://doi.acm.org/10.1145/1148170.1148204>. Citado 2 vezes nas páginas 10 e 12.

YI, X.; ALLAN, J. A comparative study of utilizing topic models for information retrieval. In: BOUGHANEM, M. et al. (Ed.). *Advances in Information Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 29–41. ISBN 978-3-642-00958-7. Citado 3 vezes nas páginas 11, 12 e 37.

ZAMIR, O.; ETZIONI, O. Web document clustering: A feasibility demonstration. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 1998. (SIGIR '98), p. 46–54. ISBN 1-58113-015-5. Disponível em: <http://doi.acm.org/10.1145/290941.290956>. Citado na página 41.

ZHAI, C. Probabilistic topic models for text data retrieval and analysis. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2017. (SIGIR '17), p. 1399–1401. ISBN 978-1-4503-5022-8. Disponível em: <http://doi.acm.org/10.1145/3077136.3082067>. Citado na página 12.

ZHAO, Y.; KARYPIS, G. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 55, n. 3, p. 311–331, jun. 2004. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/B:MACH.0000027785.44527.d6>. Citado na página 42.

ZHAO, Y.; KARYPIS, G. Soft clustering criterion functions for partitional document clustering: A summary of results. In: *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2004. (CIKM '04), p. 246–247. ISBN 1-58113-874-1. Disponível em: <http://doi.acm.org/10.1145/1031171.1031225>. Citado na página 42.

ZHU, D. et al. Intuitive topic discovery by incorporating word-pair's connection into lda. In: *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. [S.l.: s.n.], 2012. v. 1, p. 303–310. Citado 2 vezes nas páginas 37 e 71.

ZIPF, G. K. *Selective Studies and the Principle of Relative Frequency in Language*. [S.l.]: Harvard University Press, 1932. Citado na página 20.

APÊNDICE A – Tabelas para Análise de de Parâmetros para os algoritmos de Segmentação Textual

Nesse apêndice podem ser observadas tabelas com os valores de *Window Diff* P_k, Acurácia, e F^1 com as variações dos principais parâmetros dos segmentadores *TextTiling*, *C99*, *MinCutSeg*, *BayesSeg*, *TextSeg* e *PseudoSeg*.

Todos os gráficos apresentados foram analisados para escolha e configuração do algoritmo de Segmentação Textual utilizado na avaliação experimental apresentada no Capítulo 4. Nas tabelas, cada linha apresenta a variação dos parâmetros e a média dos valores obtidos por meio da segmentação de referência apresentada na Seção 2.3.3. Vale lembrar que todos os valores de *WindowDiff* e P_k, representam a dissimilaridade entre entre uma segmentação automática e uma referência.

TextTiling

Step	Win Size	$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
20	30	0.461	0.145	0.444	0.153	0.581	0.141	**0.411**	**0.161**	8.833	3.387
20	35	0.462	0.111	0.443	0.119	0.582	0.116	0.401	0.168	8.750	3.767
20	40	0.485	0.117	0.466	0.126	0.562	0.124	0.378	0.113	8.250	2.947
20	45	0.480	0.101	0.458	0.089	0.572	0.081	0.369	0.149	8.250	3.031
20	50	0.523	0.115	0.503	0.120	0.528	0.118	0.327	0.147	8.417	2.842
20	55	0.491	0.144	0.474	0.149	0.549	0.139	0.331	0.195	8.250	3.515
30	30	0.509	0.103	0.488	0.113	0.536	0.106	0.286	0.122	6.917	2.532
30	35	0.500	0.094	0.479	0.101	0.551	0.098	0.318	0.102	7.167	2.764
30	40	0.468	0.106	0.451	0.112	0.576	0.104	0.348	0.085	6.750	2.241
30	45	**0.450**	**0.103**	**0.435**	**0.109**	**0.596**	**0.110**	0.373	0.087	6.417	2.465
30	50	0.493	0.152	0.478	0.171	0.543	0.162	0.307	0.131	6.417	2.326
30	55	0.481	0.135	0.463	0.154	0.558	0.137	0.346	0.086	7.083	2.361
40	30	0.475	0.125	0.460	0.137	0.566	0.126	0.306	0.104	5.833	2.034
40	35	0.501	0.125	0.482	0.138	0.542	0.127	0.268	0.104	6.083	2.629
40	40	0.499	0.151	0.478	0.163	0.548	0.149	0.293	0.102	6.083	2.465
40	45	0.488	0.134	0.471	0.150	0.551	0.137	0.275	0.098	5.500	1.936
40	50	0.495	0.104	0.474	0.113	0.552	0.110	0.280	0.125	5.833	2.154
40	55	0.476	0.084	0.453	0.103	0.567	0.093	0.310	0.072	6.083	2.100
50	30	0.492	0.138	0.473	0.150	0.557	0.149	0.274	0.120	5.167	2.075
50	35	0.504	0.138	0.484	0.147	0.549	0.143	0.268	0.097	5.583	2.985
50	40	0.501	0.102	0.481	0.115	0.556	0.122	0.278	0.070	5.417	2.139
50	45	0.508	0.092	0.484	0.107	0.549	0.111	0.264	0.089	5.500	1.803
50	50	0.513	0.162	0.491	0.175	0.536	0.162	0.253	0.149	5.417	2.253
50	55	0.509	0.143	0.487	0.156	0.543	0.150	0.276	0.130	5.833	2.511
60	30	0.481	0.105	0.462	0.124	0.564	0.121	0.267	0.082	4.917	2.019
60	35	0.503	0.120	0.483	0.136	0.549	0.139	0.250	0.118	5.083	1.935
60	40	0.497	0.104	0.481	0.119	0.554	0.127	0.242	0.124	4.750	1.738
60	45	0.465	0.108	0.448	0.127	0.577	0.121	0.271	0.134	4.500	1.658
60	50	0.478	0.116	0.459	0.129	0.569	0.128	0.250	0.129	4.333	1.434
60	55	0.474	0.101	0.457	0.116	0.568	0.111	0.269	0.121	5.000	1.871

Tabela 11 – Valores das medidas de desempenho para análise do algoritmo *TextTiling*, utilizando o texto pré-processado.

Seg Rate	Raking Size	Weitght	$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
0.200	3	true	0.463	0.130	0.445	0.140	0.581	0.131	0.339	0.091	6.083	2.660
0.300	3	true	**0.434**	**0.089**	**0.407**	**0.101**	0.607	0.084	0.457	0.070	9.250	3.961
0.400	3	true	0.452	0.114	0.422	0.092	0.604	0.087	0.515	0.091	12.083	5.123
0.500	3	true	0.499	0.162	0.458	0.098	0.577	0.085	0.539	0.112	15.500	6.397
0.600	3	true	0.487	0.194	0.440	0.105	0.592	0.084	0.591	0.120	18.417	7.794
0.700	3	true	0.485	0.225	0.431	0.130	0.602	0.111	**0.633**	**0.134**	21.417	8.949
0.200	5	true	0.454	0.130	0.437	0.143	0.583	0.125	0.338	0.092	6.083	2.660
0.300	5	true	0.454	0.121	0.434	0.116	0.595	0.111	0.446	0.093	9.250	3.961
0.400	5	true	0.475	0.119	0.443	0.087	0.590	0.080	0.497	0.082	12.083	5.123
0.500	5	true	0.460	0.147	0.421	0.091	**0.609**	**0.079**	0.571	0.107	15.500	6.397
0.600	5	true	0.491	0.186	0.442	0.098	0.591	0.081	0.588	0.121	18.417	7.794
0.700	5	true	0.525	0.251	0.449	0.106	0.576	0.094	0.609	0.132	21.417	8.949
0.200	7	true	0.491	0.121	0.474	0.133	0.555	0.129	0.293	0.099	6.083	2.660
0.300	7	true	0.486	0.097	0.469	0.097	0.565	0.098	0.395	0.117	9.250	3.961
0.400	7	true	0.502	0.119	0.472	0.086	0.561	0.082	0.453	0.133	12.083	5.123
0.500	7	true	0.460	0.143	0.421	0.085	0.604	0.078	0.561	0.125	15.500	6.397
0.600	7	true	0.486	0.198	0.433	0.113	0.591	0.104	0.585	0.143	18.417	7.794
0.700	7	true	0.547	0.248	0.470	0.113	0.551	0.108	0.586	0.141	21.417	8.949
0.200	3	false	0.448	0.128	0.427	0.145	0.596	0.129	0.362	0.093	6.083	2.660
0.300	3	false	0.454	0.125	0.426	0.127	0.594	0.111	0.445	0.098	9.250	3.961
0.400	3	false	0.490	0.116	0.455	0.098	0.568	0.089	0.469	0.100	12.083	5.123
0.500	3	false	0.529	0.145	0.481	0.091	0.543	0.083	0.503	0.104	15.500	6.397
0.600	3	false	0.554	0.167	0.499	0.095	0.528	0.084	0.535	0.094	18.417	7.794
0.700	3	false	0.565	0.204	0.496	0.075	0.526	0.070	0.570	0.103	21.417	8.949
0.200	5	false	0.498	0.159	0.479	0.170	0.545	0.151	0.277	0.128	6.083	2.660
0.300	5	false	0.505	0.136	0.482	0.139	0.540	0.123	0.369	0.110	9.250	3.961
0.400	5	false	0.536	0.130	0.504	0.106	0.520	0.096	0.407	0.118	12.083	5.123
0.500	5	false	0.540	0.161	0.490	0.091	0.529	0.082	0.485	0.121	15.500	6.397
0.600	5	false	0.529	0.187	0.469	0.086	0.545	0.087	0.543	0.135	18.417	7.794
0.700	5	false	0.542	0.245	0.464	0.101	0.549	0.108	0.584	0.147	21.417	8.949
0.200	7	false	0.512	0.099	0.495	0.107	0.534	0.104	0.250	0.074	6.083	2.660
0.300	7	false	0.527	0.093	0.506	0.095	0.522	0.090	0.336	0.090	9.250	3.961
0.400	7	false	0.530	0.099	0.494	0.043	0.535	0.038	0.420	0.095	12.083	5.123
0.500	7	false	0.503	0.164	0.454	0.076	0.571	0.068	0.523	0.132	15.500	6.397

0.600	7	false	0.511	0.178	0.453	0.070	0.565	0.070	0.562	0.124	18.417	7.794
0.700	7	false	0.559	0.239	0.476	0.087	0.535	0.096	0.572	0.138	21.417	8.949

Tabela 12 – Valores das medidas de desempenho para análise do algoritmo *C99*, utilizando o texto pré-processado.

MinCutSeg

Seg Rate	LenCutoff	$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
0.200	5	0.523	0.127	0.499	0.136	0.530	0.130	0.241	0.087	5.833	2.609
0.200	7	0.516	0.121	0.490	0.132	0.544	0.131	0.263	0.094	5.833	2.609
0.200	9	0.516	0.107	0.490	0.121	0.545	0.127	0.268	0.091	5.833	2.609
0.200	11	0.493	0.114	0.467	0.132	0.561	0.128	0.296	0.091	5.833	2.609
0.200	13	0.491	0.111	0.464	0.124	0.564	0.119	0.296	0.079	5.833	2.609
0.200	15	0.490	0.117	0.458	0.140	0.568	0.132	0.311	0.100	5.833	2.609
0.300	5	0.478	0.096	0.450	0.123	0.575	0.121	0.410	0.091	8.667	3.771
0.300	7	0.486	0.093	0.449	0.112	0.574	0.104	0.401	0.073	8.667	3.771
0.300	9	0.484	0.104	0.445	0.116	0.579	0.112	0.409	0.108	8.667	3.771
0.300	11	0.474	0.090	0.439	0.119	0.581	0.109	0.412	0.095	8.667	3.771
0.300	13	0.457	0.095	0.427	0.119	0.594	0.112	0.433	0.099	8.667	3.771
0.300	15	0.483	0.108	0.448	0.112	0.575	0.106	0.402	0.107	8.667	3.771
0.400	5	0.484	0.077	0.447	0.120	0.571	0.108	0.477	0.096	11.917	5.251
0.400	7	0.477	0.084	0.430	0.082	0.589	0.079	0.491	0.082	11.917	5.251
0.400	9	**0.444**	**0.084**	0.408	0.093	**0.614**	**0.093**	0.526	0.084	11.917	5.251
0.400	11	0.450	0.086	0.412	0.117	0.601	0.102	0.512	0.087	11.917	5.251
0.400	13	0.462	0.089	0.422	0.131	0.589	0.112	0.499	0.103	11.917	5.251
0.400	15	0.471	0.085	0.432	0.139	0.580	0.119	0.490	0.107	11.917	5.251
0.500	5	0.493	0.112	0.435	0.098	0.578	0.088	0.535	0.091	15.000	6.519
0.500	7	0.481	0.106	0.428	0.099	0.587	0.093	0.546	0.093	15.000	6.519
0.500	9	0.467	0.107	0.412	0.098	0.600	0.090	0.560	0.094	15.000	6.519
0.500	11	0.459	0.100	**0.407**	**0.098**	0.603	0.087	0.563	0.088	15.000	6.519
0.500	13	0.500	0.112	0.444	0.096	0.572	0.088	0.528	0.092	15.000	6.519
0.500	15	0.494	0.111	0.435	0.100	0.578	0.090	0.534	0.096	15.000	6.519
0.600	5	0.520	0.140	0.449	0.077	0.564	0.073	0.559	0.096	17.917	7.719
0.600	7	0.497	0.161	0.425	0.117	0.584	0.108	0.583	0.113	17.917	7.719
0.600	9	0.501	0.173	0.428	0.110	0.579	0.103	0.577	0.114	17.917	7.719
0.600	11	0.511	0.173	0.438	0.116	0.570	0.109	0.567	0.125	17.917	7.719
0.600	13	0.502	0.168	0.428	0.118	0.579	0.110	0.576	0.124	17.917	7.719
0.600	15	0.500	0.166	0.427	0.120	0.580	0.111	0.577	0.125	17.917	7.719
0.700	5	0.528	0.219	0.438	0.122	0.567	0.120	**0.599**	**0.135**	21.000	9.211
0.700	7	0.540	0.235	0.446	0.107	0.559	0.109	0.592	0.124	21.000	9.211
0.700	9	0.567	0.218	0.473	0.094	0.535	0.093	0.570	0.117	21.000	9.211
0.700	11	0.561	0.192	0.469	0.081	0.537	0.076	0.575	0.095	21.000	9.211

0.700	13	0.564	0.192	0.472	0.083	0.534	0.078	0.572	0.097	21.000	9.211
0.700	15	0.551	0.197	0.459	0.080	0.546	0.077	0.583	0.097	21.000	9.211

Tabela 13 – Valores das medidas de desempenho para análise do algoritmo *MinCutSeg*, utilizando o texto pré-processado.

BayesSeg

#SegsKnown	Seg Rate	Prior	Dispertion	$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
false	Auto	0.0800	0.1000	0.395	0.084	0.377	0.105	0.640	0.092	0.528	0.087	9.667	1.748
false	Auto	0.0900	0.1000	0.402	0.078	0.383	0.096	0.636	0.088	0.515	0.077	9.333	1.650
false	Auto	0.1000	0.1000	0.395	0.074	0.376	0.092	0.642	0.083	0.518	0.077	9.167	1.572
false	Auto	0.1100	0.1000	0.402	0.081	0.383	0.099	0.636	0.090	0.508	0.075	9.000	1.414
false	Auto	0.0800	0.3000	**0.380**	**0.086**	**0.361**	**0.104**	**0.655**	**0.091**	0.551	0.100	10.000	1.780
false	Auto	0.0900	0.3000	0.393	0.081	0.374	0.097	0.645	0.088	0.529	0.092	9.583	1.754
false	Auto	0.1000	0.3000	0.393	0.071	0.374	0.089	0.644	0.081	0.520	0.083	9.167	1.404
false	Auto	0.1100	0.3000	0.390	0.070	0.371	0.088	0.647	0.079	0.522	0.084	9.083	1.382
false	Auto	0.0800	0.5000	**0.380**	**0.086**	**0.361**	**0.104**	**0.655**	**0.091**	0.551	0.100	10.000	1.780
false	Auto	0.0900	0.5000	0.398	0.082	0.379	0.099	0.640	0.090	0.523	0.095	9.583	1.656
false	Auto	0.1000	0.5000	0.397	0.074	0.378	0.092	0.641	0.084	0.518	0.084	9.250	1.479
false	Auto	0.1100	0.5000	0.388	0.072	0.370	0.089	0.649	0.080	0.523	0.083	9.000	1.225
false	Auto	0.0800	0.7000	0.385	0.073	0.366	0.089	0.652	0.081	0.546	0.095	10.000	1.683
false	Auto	0.0900	0.7000	0.393	0.077	0.374	0.094	0.645	0.086	0.528	0.101	9.667	1.650
false	Auto	0.1000	0.7000	0.395	0.076	0.376	0.094	0.642	0.085	0.519	0.083	9.167	1.344
false	Auto	0.1100	0.7000	0.388	0.072	0.370	0.089	0.649	0.080	0.523	0.083	9.000	1.225
true	0.300	0.0800	0.1000	0.428	0.150	0.398	0.171	0.617	0.154	0.491	0.122	9.250	3.961
true	0.300	0.0900	0.1000	0.428	0.150	0.398	0.171	0.617	0.154	0.491	0.122	9.250	3.961
true	0.300	0.1000	0.1000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.300	0.1100	0.1000	0.427	0.150	0.398	0.174	0.615	0.155	0.487	0.129	9.250	3.961
true	0.300	0.0800	0.3000	0.428	0.150	0.398	0.171	0.617	0.154	0.491	0.122	9.250	3.961
true	0.300	0.0900	0.3000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.300	0.1000	0.3000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.300	0.1100	0.3000	0.424	0.152	0.395	0.176	0.618	0.156	0.492	0.130	9.250	3.961
true	0.300	0.0800	0.5000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.300	0.0900	0.5000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.300	0.1000	0.5000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.300	0.1100	0.5000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.300	0.0800	0.7000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.300	0.0900	0.7000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.300	0.1000	0.7000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.300	0.1100	0.7000	0.428	0.150	0.399	0.170	0.614	0.151	0.485	0.121	9.250	3.961
true	0.600	0.0800	0.1000	0.480	0.133	0.416	0.056	0.598	0.052	0.601	0.079	18.417	7.794
true	0.600	0.0900	0.1000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794

true	0.600	0.1000	0.1000	0.467	0.139	0.404	0.056	0.611	0.052	0.613	0.079	18.417	7.794
true	0.600	0.1100	0.1000	0.462	0.141	0.399	0.055	0.615	0.051	**0.619**	**0.074**	18.417	7.794
true	0.600	0.0800	0.3000	0.480	0.133	0.416	0.056	0.598	0.052	0.601	0.079	18.417	7.794
true	0.600	0.0900	0.3000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794
true	0.600	0.1000	0.3000	0.467	0.139	0.404	0.056	0.611	0.052	0.613	0.079	18.417	7.794
true	0.600	0.1100	0.3000	0.462	0.141	0.399	0.055	0.615	0.051	**0.619**	**0.074**	18.417	7.794
true	0.600	0.0800	0.5000	0.480	0.133	0.416	0.056	0.598	0.052	0.601	0.079	18.417	7.794
true	0.600	0.0900	0.5000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794
true	0.600	0.1000	0.5000	0.467	0.139	0.404	0.056	0.611	0.052	0.613	0.079	18.417	7.794
true	0.600	0.1100	0.5000	0.462	0.141	0.399	0.055	0.615	0.051	**0.619**	**0.074**	18.417	7.794
true	0.600	0.0800	0.7000	0.480	0.133	0.416	0.056	0.598	0.052	0.601	0.079	18.417	7.794
true	0.600	0.0900	0.7000	0.480	0.133	0.416	0.056	0.598	0.052	0.601	0.079	18.417	7.794
true	0.600	0.1000	0.7000	0.467	0.139	0.404	0.056	0.611	0.052	0.613	0.079	18.417	7.794
true	0.600	0.1100	0.7000	0.462	0.141	0.399	0.055	0.615	0.051	**0.619**	**0.074**	18.417	7.794
true	0.900	0.0800	0.1000	0.645	0.352	0.517	0.131	0.490	0.142	0.600	0.148	27.500	11.601
true	0.900	0.0900	0.1000	0.645	0.352	0.517	0.131	0.490	0.142	0.600	0.148	27.500	11.601
true	0.900	0.1000	0.1000	0.651	0.348	0.524	0.127	0.483	0.138	0.596	0.145	27.500	11.601
true	0.900	0.1100	0.1000	0.651	0.348	0.524	0.127	0.483	0.138	0.596	0.145	27.500	11.601
true	0.900	0.0800	0.3000	0.645	0.352	0.517	0.131	0.490	0.142	0.600	0.148	27.500	11.601
true	0.900	0.0900	0.3000	0.645	0.352	0.517	0.131	0.490	0.142	0.600	0.148	27.500	11.601
true	0.900	0.1000	0.3000	0.651	0.348	0.524	0.127	0.483	0.138	0.596	0.145	27.500	11.601
true	0.900	0.1100	0.3000	0.651	0.348	0.524	0.127	0.483	0.138	0.596	0.145	27.500	11.601
true	0.900	0.0800	0.5000	0.645	0.352	0.517	0.131	0.490	0.142	0.600	0.148	27.500	11.601
true	0.900	0.0900	0.5000	0.645	0.352	0.517	0.131	0.490	0.142	0.600	0.148	27.500	11.601
true	0.900	0.1000	0.5000	0.651	0.348	0.524	0.127	0.483	0.138	0.596	0.145	27.500	11.601
true	0.900	0.1100	0.5000	0.651	0.348	0.524	0.127	0.483	0.138	0.596	0.145	27.500	11.601
true	0.900	0.0800	0.7000	0.645	0.352	0.517	0.131	0.490	0.142	0.600	0.148	27.500	11.601
true	0.900	0.0900	0.7000	0.645	0.352	0.517	0.131	0.490	0.142	0.600	0.148	27.500	11.601
true	0.900	0.1000	0.7000	0.651	0.348	0.524	0.127	0.483	0.138	0.596	0.145	27.500	11.601
true	0.900	0.1100	0.7000	0.651	0.348	0.524	0.127	0.483	0.138	0.596	0.145	27.500	11.601

Tabela 14 – Valores das medidas de desempenho para análise do algoritmo *BayesSeg*, utilizando o texto pré-processado.

TextSeg

Seg Rate	$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
Auto	**0.455**	**0.130**	0.439	0.142	0.585	0.132	0.368	0.130	6.417	0.954
0.100	0.502	0.146	0.486	0.160	0.548	0.158	0.163	0.122	3.167	1.344
0.200	0.473	0.160	0.452	0.175	0.569	0.159	0.320	0.166	6.083	2.660
0.300	0.496	0.159	0.460	0.180	0.560	0.165	0.406	0.150	9.250	3.961
0.400	0.484	0.119	0.444	0.142	0.575	0.125	0.487	0.111	12.083	5.123
0.500	0.475	0.107	**0.417**	**0.108**	**0.594**	**0.087**	0.566	0.073	15.500	6.397
0.600	0.504	0.124	0.439	0.087	0.571	0.067	0.582	0.054	18.417	7.794
0.700	0.531	0.173	0.447	0.074	0.562	0.066	0.605	0.083	21.417	8.949
0.800	0.579	0.259	0.478	0.103	0.531	0.109	0.605	0.126	24.417	10.259
0.900	0.604	0.340	0.484	0.130	0.524	0.140	**0.627**	**0.142**	27.500	11.601

Tabela 15 – Valores das medidas de desempenho para análise do algoritmo *TextSeg*, utilizando o texto pré-processado.

TextTiling

Step	Win Size	$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
20	30	0.513	0.138	0.490	0.144	0.538	0.138	0.334	0.173	8.500	3.571
20	35	0.509	0.127	0.492	0.126	0.540	0.121	0.350	0.135	8.583	2.871
20	40	0.517	0.132	0.495	0.144	0.532	0.137	0.342	0.142	8.583	3.148
20	45	0.496	0.114	0.477	0.122	0.555	0.117	0.347	0.117	7.667	2.528
20	50	0.481	0.140	0.465	0.138	0.569	0.134	**0.390**	**0.178**	8.750	3.467
20	55	0.512	0.133	0.493	0.135	0.542	0.132	0.337	0.156	8.250	3.295
30	30	0.511	0.130	0.494	0.130	0.538	0.128	0.284	0.145	6.667	2.173
30	35	0.517	0.100	0.500	0.109	0.536	0.113	0.285	0.099	6.583	2.019
30	40	0.512	0.128	0.491	0.131	0.543	0.121	0.299	0.082	6.750	2.586
30	45	0.502	0.112	0.483	0.108	0.555	0.106	0.320	0.087	6.917	2.499
30	50	0.510	0.107	0.493	0.117	0.539	0.117	0.313	0.112	7.333	2.560
30	55	0.498	0.146	0.480	0.162	0.543	0.146	0.328	0.115	7.250	2.350
40	30	0.493	0.132	0.477	0.141	0.555	0.134	0.248	0.071	4.917	2.060
40	35	0.482	0.121	0.465	0.132	0.558	0.123	0.267	0.106	5.417	2.178
40	40	0.476	0.112	0.459	0.120	0.565	0.114	0.275	0.120	5.500	2.566
40	45	0.501	0.134	0.482	0.144	0.549	0.143	0.260	0.120	5.333	2.095
40	50	0.498	0.123	0.481	0.135	0.551	0.134	0.266	0.087	5.333	2.285
40	55	0.505	0.116	0.487	0.131	0.544	0.131	0.243	0.077	5.083	1.706
50	30	**0.474**	**0.135**	**0.455**	**0.138**	**0.579**	**0.132**	0.295	0.106	4.917	1.552
50	35	0.528	0.126	0.511	0.137	0.531	0.146	0.202	0.088	4.583	1.706
50	40	0.501	0.103	0.488	0.121	0.539	0.122	0.234	0.108	5.000	1.683
50	45	0.489	0.112	0.476	0.125	0.558	0.135	0.275	0.092	5.167	2.034
50	50	0.498	0.158	0.483	0.171	0.545	0.162	0.304	0.100	6.083	1.891
50	55	0.490	0.151	0.470	0.167	0.556	0.157	0.303	0.123	5.583	2.178
60	30	0.499	0.092	0.486	0.103	0.557	0.123	0.234	0.098	4.417	1.754
60	35	0.509	0.143	0.494	0.164	0.537	0.159	0.243	0.111	5.000	1.472
60	40	0.501	0.113	0.486	0.128	0.545	0.129	0.182	0.108	3.833	1.462
60	45	0.493	0.118	0.478	0.129	0.558	0.136	0.227	0.136	4.167	1.462
60	50	0.495	0.110	0.478	0.118	0.562	0.127	0.225	0.081	4.083	1.656
60	55	0.500	0.104	0.485	0.114	0.550	0.120	0.198	0.075	4.000	1.155

Tabela 16 – Valores das medidas de desempenho para análise do algoritmo *TextTiling*, utilizando o texto o texto integral.

Seg Rate	Raking Size	Weitght	$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
0.200	3	true	0.481	0.118	0.463	0.121	0.574	0.122	0.324	0.094	6.083	2.660
0.300	3	true	0.457	0.109	0.437	0.104	0.596	0.105	0.447	0.091	9.250	3.961
0.400	3	true	0.450	0.153	0.425	0.142	0.602	0.123	0.513	0.143	12.083	5.123
0.500	3	true	**0.435**	**0.155**	**0.395**	**0.106**	**0.629**	**0.095**	0.594	0.123	15.500	6.397
0.600	3	true	0.489	0.194	0.437	0.091	0.592	0.075	0.591	0.119	18.417	7.794
0.700	3	true	0.482	0.232	0.420	0.111	0.602	0.107	**0.632**	**0.139**	21.417	8.949
0.200	5	true	0.488	0.122	0.469	0.133	0.565	0.135	0.313	0.106	6.083	2.660
0.300	5	true	0.476	0.166	0.458	0.175	0.571	0.166	0.426	0.151	9.250	3.961
0.400	5	true	0.476	0.127	0.452	0.127	0.578	0.121	0.487	0.113	12.083	5.123
0.500	5	true	0.463	0.142	0.425	0.095	0.605	0.087	0.566	0.119	15.500	6.397
0.600	5	true	0.464	0.187	0.415	0.110	0.610	0.100	0.604	0.141	18.417	7.794
0.700	5	true	0.504	0.244	0.435	0.117	0.589	0.108	0.619	0.142	21.417	8.949
0.200	7	true	0.478	0.124	0.459	0.133	0.574	0.135	0.328	0.108	6.083	2.660
0.300	7	true	0.481	0.145	0.462	0.150	0.570	0.141	0.418	0.115	9.250	3.961
0.400	7	true	0.478	0.129	0.452	0.125	0.577	0.118	0.482	0.127	12.083	5.123
0.500	7	true	0.471	0.171	0.427	0.108	0.604	0.093	0.563	0.131	15.500	6.397
0.600	7	true	0.480	0.186	0.429	0.104	0.599	0.094	0.594	0.134	18.417	7.794
0.700	7	true	0.516	0.241	0.444	0.106	0.579	0.100	0.611	0.133	21.417	8.949
0.200	3	false	0.469	0.119	0.453	0.129	0.579	0.130	0.335	0.107	6.083	2.660
0.300	3	false	0.441	0.073	0.421	0.086	0.608	0.089	0.463	0.056	9.250	3.961
0.400	3	false	0.467	0.062	0.439	0.057	0.591	0.067	0.493	0.092	12.083	5.123
0.500	3	false	0.483	0.137	0.442	0.082	0.593	0.078	0.554	0.108	15.500	6.397
0.600	3	false	0.500	0.199	0.442	0.099	0.589	0.085	0.587	0.120	18.417	7.794
0.700	3	false	0.492	0.244	0.423	0.115	0.602	0.103	0.632	0.133	21.417	8.949
0.200	5	false	0.495	0.161	0.476	0.170	0.555	0.160	0.300	0.128	6.083	2.660
0.300	5	false	0.503	0.134	0.485	0.143	0.549	0.141	0.386	0.123	9.250	3.961
0.400	5	false	0.496	0.110	0.477	0.104	0.564	0.108	0.466	0.109	12.083	5.123
0.500	5	false	0.488	0.114	0.452	0.072	0.574	0.067	0.533	0.104	15.500	6.397
0.600	5	false	0.484	0.171	0.434	0.077	0.594	0.065	0.592	0.108	18.417	7.794
0.700	5	false	0.522	0.235	0.451	0.105	0.574	0.095	0.609	0.122	21.417	8.949
0.200	7	false	0.489	0.162	0.471	0.170	0.560	0.159	0.307	0.132	6.083	2.660
0.300	7	false	0.498	0.146	0.479	0.153	0.554	0.149	0.394	0.132	9.250	3.961
0.400	7	false	0.500	0.119	0.475	0.111	0.561	0.108	0.462	0.110	12.083	5.123
0.500	7	false	0.479	0.145	0.441	0.089	0.592	0.080	0.551	0.115	15.500	6.397

0.600	7	false	0.493	0.172	0.439	0.080	0.585	0.073	0.586	0.106	18.417	7.794
0.700	7	false	0.506	0.261	0.430	0.131	0.590	0.126	0.621	0.149	21.417	8.949

Tabela 17 – Valores das medidas de desempenho para análise do algoritmo *C99*, utilizando o texto o texto integral.

MinCutSeg

Seg Rate	LenCutoff	$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
0.200	5	0.513	0.132	0.489	0.143	0.539	0.137	0.257	0.118	5.833	2.609
0.200	7	0.510	0.128	0.486	0.135	0.545	0.132	0.267	0.098	5.833	2.609
0.200	9	0.498	0.111	0.474	0.130	0.553	0.127	0.282	0.097	5.833	2.609
0.200	11	0.487	0.115	0.459	0.135	0.566	0.128	0.302	0.103	5.833	2.609
0.200	13	0.473	0.124	0.445	0.135	0.580	0.126	0.324	0.093	5.833	2.609
0.200	15	0.467	0.128	0.443	0.145	0.581	0.137	0.333	0.109	5.833	2.609
0.300	5	0.483	0.082	0.451	0.110	0.573	0.104	0.402	0.062	8.667	3.771
0.300	7	0.474	0.110	0.437	0.121	0.585	0.113	0.421	0.085	8.667	3.771
0.300	9	0.480	0.099	0.441	0.118	0.579	0.107	0.410	0.093	8.667	3.771
0.300	11	0.454	0.098	0.418	0.119	0.601	0.109	0.442	0.092	8.667	3.771
0.300	13	0.460	0.097	0.423	0.124	0.594	0.111	0.434	0.091	8.667	3.771
0.300	15	0.455	0.100	0.417	0.125	0.599	0.111	0.440	0.096	8.667	3.771
0.400	5	0.444	0.082	0.407	0.117	0.609	0.107	0.523	0.104	11.917	5.251
0.400	7	0.455	0.095	0.410	0.104	0.606	0.093	0.513	0.098	11.917	5.251
0.400	9	0.465	0.130	0.418	0.135	0.601	0.123	0.514	0.112	11.917	5.251
0.400	11	0.442	0.137	0.404	0.156	0.613	0.142	0.533	0.136	11.917	5.251
0.400	13	0.434	0.144	0.400	0.162	**0.620**	**0.152**	0.543	0.148	11.917	5.251
0.400	15	**0.430**	**0.150**	**0.397**	**0.172**	0.620	0.156	0.543	0.152	11.917	5.251
0.500	5	0.484	0.128	0.426	0.112	0.587	0.099	0.550	0.085	15.000	6.519
0.500	7	0.472	0.162	0.412	0.127	0.602	0.121	0.563	0.133	15.000	6.519
0.500	9	0.466	0.147	0.411	0.140	0.602	0.128	0.567	0.127	15.000	6.519
0.500	11	0.465	0.141	0.413	0.141	0.598	0.127	0.564	0.122	15.000	6.519
0.500	13	0.451	0.146	0.399	0.149	0.612	0.134	0.578	0.130	15.000	6.519
0.500	15	0.462	0.154	0.405	0.148	0.606	0.134	0.570	0.134	15.000	6.519
0.600	5	0.500	0.154	0.431	0.099	0.581	0.088	0.581	0.091	17.917	7.719
0.600	7	0.498	0.143	0.427	0.110	0.579	0.096	0.579	0.104	17.917	7.719
0.600	9	0.492	0.153	0.423	0.107	0.588	0.098	0.591	0.095	17.917	7.719
0.600	11	0.482	0.161	0.412	0.112	0.598	0.102	0.600	0.102	17.917	7.719
0.600	13	0.474	0.150	0.404	0.121	0.602	0.105	0.605	0.102	17.917	7.719
0.600	15	0.482	0.161	0.410	0.113	0.598	0.102	0.600	0.102	17.917	7.719
0.700	5	0.512	0.193	0.424	0.076	0.579	0.076	**0.612**	**0.097**	21.000	9.211
0.700	7	0.522	0.194	0.433	0.089	0.570	0.085	0.603	0.105	21.000	9.211
0.700	9	0.528	0.205	0.438	0.098	0.565	0.091	0.602	0.097	21.000	9.211
0.700	11	0.532	0.220	0.440	0.093	0.568	0.088	0.605	0.094	21.000	9.211

0.700	13	0.537	0.210	0.445	0.095	0.560	0.088	0.598	0.094	21.000	9.211
0.700	15	0.530	0.208	0.438	0.085	0.567	0.080	0.604	0.087	21.000	9.211

Tabela 18 – Valores das medidas de desempenho para análise do algoritmo *MinCutSeg*, utilizando o texto o texto integral.

BayesSeg

#SegsKnown	Seg Rate	Prior	Dispertion	$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
false	Auto	0.0800	0.1000	0.399	0.087	0.380	0.108	0.637	0.095	0.526	0.088	9.750	1.785
false	Auto	0.0900	0.1000	0.405	0.080	0.386	0.099	0.633	0.091	0.513	0.077	9.417	1.706
false	Auto	0.1000	0.1000	0.399	0.077	0.380	0.095	0.639	0.087	0.517	0.078	9.250	1.639
false	Auto	0.1100	0.1000	0.405	0.083	0.387	0.102	0.633	0.093	0.506	0.075	9.083	1.498
false	Auto	0.0800	0.3000	**0.383**	**0.089**	**0.364**	**0.107**	**0.652**	**0.094**	0.549	0.101	10.083	1.801
false	Auto	0.0900	0.3000	0.396	0.084	0.377	0.100	0.642	0.091	0.527	0.093	9.667	1.795
false	Auto	0.1000	0.3000	0.397	0.074	0.378	0.092	0.641	0.084	0.518	0.084	9.250	1.479
false	Auto	0.1100	0.3000	0.393	0.073	0.374	0.091	0.644	0.082	0.520	0.084	9.167	1.462
false	Auto	0.0800	0.5000	**0.383**	**0.089**	**0.364**	**0.107**	**0.652**	**0.094**	0.549	0.101	10.083	1.801
false	Auto	0.0900	0.5000	0.401	0.084	0.382	0.102	0.637	0.093	0.521	0.096	9.667	1.700
false	Auto	0.1000	0.5000	0.400	0.077	0.381	0.095	0.638	0.087	0.516	0.084	9.333	1.546
false	Auto	0.1100	0.5000	0.392	0.075	0.373	0.092	0.646	0.083	0.521	0.083	9.083	1.320
false	Auto	0.0800	0.7000	0.388	0.077	0.369	0.093	0.649	0.085	0.545	0.096	10.083	1.706
false	Auto	0.0900	0.7000	0.396	0.080	0.377	0.097	0.642	0.089	0.526	0.102	9.750	1.689
false	Auto	0.1000	0.7000	0.398	0.079	0.380	0.097	0.639	0.088	0.517	0.083	9.250	1.422
false	Auto	0.1100	0.7000	0.392	0.075	0.373	0.092	0.646	0.083	0.521	0.083	9.083	1.320
true	0.300	0.0800	0.1000	0.421	0.144	0.391	0.165	0.624	0.147	0.499	0.110	9.250	3.961
true	0.300	0.0900	0.1000	0.421	0.144	0.391	0.165	0.624	0.147	0.499	0.110	9.250	3.961
true	0.300	0.1000	0.1000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.300	0.1100	0.1000	0.420	0.143	0.392	0.168	0.621	0.148	0.495	0.119	9.250	3.961
true	0.300	0.0800	0.3000	0.421	0.144	0.391	0.165	0.624	0.147	0.499	0.110	9.250	3.961
true	0.300	0.0900	0.3000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.300	0.1000	0.3000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.300	0.1100	0.3000	0.417	0.146	0.389	0.169	0.624	0.150	0.500	0.120	9.250	3.961
true	0.300	0.0800	0.5000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.300	0.0900	0.5000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.300	0.1000	0.5000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.300	0.1100	0.5000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.300	0.0800	0.7000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.300	0.0900	0.7000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.300	0.1000	0.7000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.300	0.1100	0.7000	0.421	0.144	0.393	0.163	0.620	0.145	0.493	0.110	9.250	3.961
true	0.600	0.0800	0.1000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794
true	0.600	0.0900	0.1000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794

true	0.600	0.1000	0.1000	0.467	0.139	0.404	0.056	0.611	0.052	0.613	0.079	18.417	7.794
true	0.600	0.1100	0.1000	0.462	0.141	0.399	0.055	0.615	0.051	**0.619**	**0.074**	18.417	7.794
true	0.600	0.0800	0.3000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794
true	0.600	0.0900	0.3000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794
true	0.600	0.1000	0.3000	0.467	0.139	0.404	0.056	0.611	0.052	0.613	0.079	18.417	7.794
true	0.600	0.1100	0.3000	0.462	0.141	0.399	0.055	0.615	0.051	**0.619**	**0.074**	18.417	7.794
true	0.600	0.0800	0.5000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794
true	0.600	0.0900	0.5000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794
true	0.600	0.1000	0.5000	0.467	0.139	0.404	0.056	0.611	0.052	0.613	0.079	18.417	7.794
true	0.600	0.1100	0.5000	0.462	0.141	0.399	0.055	0.615	0.051	**0.619**	**0.074**	18.417	7.794
true	0.600	0.0800	0.7000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794
true	0.600	0.0900	0.7000	0.473	0.137	0.410	0.057	0.605	0.054	0.607	0.083	18.417	7.794
true	0.600	0.1000	0.7000	0.467	0.139	0.404	0.056	0.611	0.052	0.613	0.079	18.417	7.794
true	0.600	0.1100	0.7000	0.462	0.141	0.399	0.055	0.615	0.051	**0.619**	**0.074**	18.417	7.794
true	0.900	0.0800	0.1000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.0900	0.1000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.1000	0.1000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.1100	0.1000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.0800	0.3000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.0900	0.3000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.1000	0.3000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.1100	0.3000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.0800	0.5000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.0900	0.5000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.1000	0.5000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.1100	0.5000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.0800	0.7000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.0900	0.7000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.1000	0.7000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601
true	0.900	0.1100	0.7000	0.638	0.357	0.511	0.139	0.496	0.149	0.605	0.153	27.500	11.601

Tabela 19 – Valores das medidas de desempenho para análise do algoritmo *BayesSeg*, utilizando o texto o texto integral.

TextSeg

Seg Rate	$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
Auto	**0.430**	**0.131**	**0.413**	**0.142**	**0.610**	**0.131**	0.397	0.133	6.083	0.862
0.100	0.493	0.172	0.476	0.185	0.558	0.181	0.191	0.155	3.167	1.344
0.200	0.456	0.135	0.435	0.155	0.585	0.141	0.347	0.115	6.083	2.660
0.300	0.483	0.135	0.451	0.168	0.567	0.151	0.419	0.125	9.250	3.961
0.400	0.469	0.140	0.426	0.167	0.586	0.145	0.507	0.122	12.083	5.123
0.500	0.476	0.127	0.417	0.108	0.593	0.093	0.563	0.082	15.500	6.397
0.600	0.496	0.150	0.425	0.071	0.587	0.058	0.593	0.070	18.417	7.794
0.700	0.551	0.210	0.463	0.065	0.550	0.064	0.591	0.097	21.417	8.949
0.800	0.593	0.279	0.488	0.101	0.522	0.108	0.595	0.134	24.417	10.259
0.900	0.620	0.342	0.495	0.115	0.511	0.130	**0.618**	**0.138**	27.500	11.601

Tabela 20 – Valores das medidas de desempenho para análise do algoritmo *TextSeg*, utilizando o texto o texto integral.

PseudoSeg

$WinDiff$	$\sigma WinDiff$	P_k	σP_k	Acurácia	σAcurácia	F^1	σF^1	#Segs	σ#Segs
0.640	**0.415**	**0.490**	**0.149**	**0.506**	**0.172**	**0.638**	**0.156**	30.500	12.907

Tabela 21 – Valores das medidas de desempenho para análise do pseudo algoritmo *PseudoSeg*, utilizando o texto o texto integral.

APÊNDICE B – Questionários Utilizados na Avaliação Subjetiva

Nesse apêndice podem ser observados os questionários utilizados na Avaliação Subjetiva dos Extratores de Tópicos e segmentador (*BayesSeg*) empregado nesse trabalho. O documento é composto por dois questionários sendo estes entregues a avaliadores divididos em dois grupos distintos. A descrição completa desses questionários bem como as análises dos dados obtidos podem ser vistas nos Capítulos 4 e 5.

Avaliação de extração de tópicos em atas de reuniões

Apresentação

Essa avaliação se refere a um sistema cujo objetivo é ajudar o usuário a fazer buscas por um assunto específico em uma coleção de atas de reunião. O sistema recebe uma consulta do usuário sobre um assunto e apresenta os trechos onde esse assunto é mencionado. Inicialmente, o sistema analisa todas atas e divide o texto de cada ata em trechos que contêm um assunto principal e relativamente independente. Ou seja, os diversos assuntos tratados em uma ata são separados em trechos com um único assunto. Em seguida, utiliza-se técnicas de inteligência artificial para identificar trechos com assuntos relacionados e agrupá-los. Cada grupo contém trechos de atas diferentes mas com assuntos relacionados. Além disso, o sistema seleciona um conjunto de palavras que indicam o assunto do grupo. Assim, espera-se que o agrupamento de trechos com assuntos similares extraídos de diferentes documentos facilite a navegação e busca por assuntos na coleção de atas.

Essa avaliação se refere a componentes internos do sistema e não em como seria o seu uso. Para isso, foi feita uma consulta ao sistema com os termos *"compra de equipamentos"* e foram extraídos alguns elementos para a avaliação dos componentes de interesse.

A seguir, você encontrará três grupos de cinco trechos de tamanhos variados. Para cada um dos grupos, há um conjunto de 4 questões simples, que você deve responder de forma a refletir a sua percepção quanto à qualidade do grupo e respectivos trechos. A avaliação deve ser feita na ordem em que os grupos e as questões foram apresentados, não importando a numeração de consulta e técnica fornecida (são apenas para controle interno).

Antes de iniciar a avaliação, identifique aqui a sua afinidade com atas de reuniões (por exemplo, presidente de conselho, membro de conselho, secretário, coordenador de curso):

Consulta 1 – Técnica 1

Grupo de trechos:

(III) COMPRAS COM VERBA DE MATERIAL DE CUSTEIO E PERMANENTE. (III.I) A professora AAA esclareceu que vamos continuar fazendo pedidos de compras e que chegaram novas demandas: compra de memória RAM para 32 máquinas (a pedido dos profs.
A professora AAA ficou responsável pela compra do material seguindo as sugestões dentro orçamento disponível.
Informes: O professor BBB informou que irá disponibilizar um link para abertura das solicitações de compras para 2015, e também que quando for realizada a reunião para distribuição das porcentagens da verba, irá solicitar o máximo possível para material permanente, visto que iremos precisar de verba permanente para equipar o prédio que será entregue o ano que vem.
(VII) DISCUSSÃO SOBRE A COMPRA DE MATERIAL PERMANENTE. (VII.I) Após discussão ficou acertado a compra de material e de informática e material para coffee break.
(IV) COMPRAS COM VERBA DE MATERIAL DE CUSTEIO E PERMANENTE (IV.I) A professora AAA forneceu os saldos parciais que ainda temos, em relação a capital para permanente temos em torno de R$ 2918,00 reais o qual será utilizada para compra de um PC novo para secretaria do DComp-So, e quando for liberado a segunda parcela da verba de capital a intenção é que se compre câmeras para os laboratórios, nobreak para servidor e também que seja atendida as prioridades 3 e 4 dos professores que já constam em planilha. Os professores aprovaram a distribuição sugerida. Já em relação ao custeio embora o professor CCC não tenha atualizado a planilha, fizemos uma estimativa e acreditamos que temos em torno de 7 mil reais. Foi pedido para professores apresentarem suas demandas que foram a seguinte: 14 pro labores de 300 reais; 32 memórias RAM; divisória antirruídos para criar um espaço dentro da sala do ATLab onde está cluster, nessa sala ficaria o técnico DDD; placas para identificação das portas dos laboratórios; HD de 1Tb para o servidor do LaSID; 2 HDs de 2,5 polegadas de 500Gb e também 51 ventoinhas. A professora EEE apresentou a demanda para pagamento de diária, inscrição e passagem para apresentação de artigo em congresso. Foram aprovados os pro labores e demanda de infraestrutura, não sendo possível atender a demanda de ajuda de custo da professora FFF. Serão dados os encaminhamentos para aquisição dos materiais aprovados, tais compras serão requisitadas na medida em que a verba for liberada.

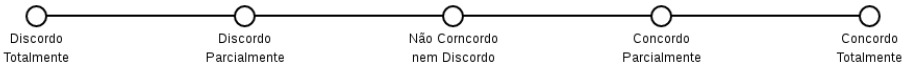
Consulta 1 – Técnica 1

Questões:

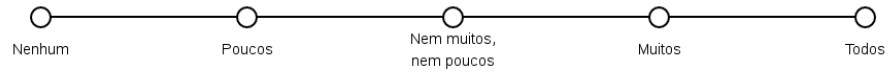
Com base na análise dos trechos, responda as questões a seguir:

1. Todos os trechos apresentados compartilham um mesmo assunto.

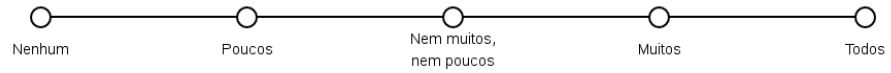
2. As palavras *compra, material, verba, permanente e valor* resumem bem o assunto tratado nos trechos.



3. Existem trechos que não tratam de um único assunto?



4. Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?



Consulta 1 – Técnica 2

Grupo de trechos:

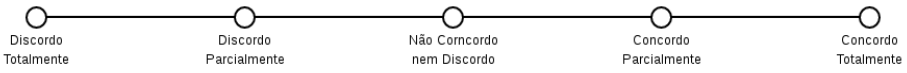
Sobre as câmeras ele fez um orçamento informal de um sistema que atenda as nossas necessidades e foi passado um valor de aproximadamente cinco mil reais, já o ar condicionado ele ainda não tem o valor.
Informes: O professor AAA Informou que já foi estabelecida as regras de como será realizada a divisão do orçamento, porém ainda não temos nem ideia de qual será o valor.
Informes: A professora BBB informou que o ofício a respeito do processo de seleção do PIBIC será enviado, no entanto o tom do ofício foi suavizado, informou também que tem um valor em auxílio estudante que precisará liquidar ainda este ano, pois provavelmente ele não virará o ano por ser de 2015, sendo assim está aceitando sugestões para gastos do referido dinheiro com apoio a projetos de disciplinas no valor máximo de até 800 reais e o que sobrar será enviado como ajuda custos para SeCoT 2017.
(VI) DELIBERAÇÃO SOBRE ORÇAMENTO PARA 2015 - COMPRA DE MATERIAIS DA VERBA DE CUSTEIO. (VI.I) A Profa. BBB colocou que temos um total de $R25.339,71 emais R$ 727,38 que foi destinado à pós, perfazendo um montante de R$ 26.607,12 para serem gastos com custeio e aulas práticas, relatou que alguns itens já foram pedidos como toners para a impressora do departamento, cabos para a SeCoT, auxílio estudante para maratona e hoje iremos deliberar sobre a visita técnica pedida pela coordenação do curso de BCCS, o qual deve ficar entre 1500 a 2000 reais, pagamento de pró-labores para banca da pós-graduação (8 de aproximadamente 300 reais) e auxílio estudante para congresso também da pós-graduação (5 de 500 reais).
(V.I) O prof. CCC explicou que este ano foi a primeira vez que recebemos verbas para todas as alíneas, foram R$2.365,52 para aulas práticas, R$ 19.180,57 para custeio, R$ 4.718,19 para material permanente e tínhamos segundo informações de São Carlos R$ 1.386,72 na conta do repasse FAI. A verba de material permanente e repasse FAI, foi empenhada para compra de cadeiras para os laboratórios, R$ 4.720,38 (verba material permanente) para compra 24 cadeiras e R$ 1.180,10 (repasse FAI) para compra de mais 6 cadeiras. O professor CCC explicou que nossa maior verba era para custeio a qual poder ser usada com pró-labore, diárias, auxilio estudante e outros diversos empenhos inclusive auxilio ao pesquisador e também em serviços, o que não foi muito fácil, pois ainda não temos experiência neste tipo de requisição. A verba de custeio e de aula prática foi unificada e ficamos com o montante de R$21.546,09, deste valor foi requisitado em material físico o total de R$ 7.700,97, alguns itens foram requisitados em quantidade maior do que precisamos no momento para que tenhamos um estoque, pois não sabemos como será dividida a verba no próximo ano. Já com serviços o total empenhado foi de R$ 1.888,49, sendo que R$ 1.111,82 com assinatura do Dream Spark (valor do dia da requisição, o que pode variar pelo fato de ser em dólar), e com a Arte Gráfica o total de R$ 776,67. Também empenhamos R$ 1.200,00 para diárias, R$ 528,00 para pró-labore e R$ 300,00 com auxílio estudante. Em relação à compra da assinatura do Dream Spark, a compra já teria acontecido se a empresa não estivesse com pendência no INSS, o setor de compras entrou com a referida empresa e a mesma informou que tal pendência será resolvida até dia 05 de outubro, caso a compra concretize o Departamento de Administração irá arcar com a metade de valor por que eles estão interessados em utilizar o MS Visio. O professor CCC colocou que ainda temos um montante que gira em torno de R$ 9.900,00 e dos quais mais ou menos R$ 1.500,00 será gasto com a gráfica, R$ 1.500,00 será permutado com o curso para compra de roteadores sem fio. Precisamos decidir onde empenharemos o restante, sendo que talvez tenhamos uma demanda de pagamento de passagens aéreas caso nossa equipe passe para a segunda fase da maratona de programação, (o total de 4 passagens). O conselho deliberou a favor de pagamento das passagens, porém sugeriu que as mesmas sejam pagas por meio de auxílio estudante, pois desta maneira o valor será bem menor. Aprovado. Também foi colocado o pedido da professora DDD para pagamento de taxa de inscrição no evento X IEEE International Conference on e Science, no valor de hoje de R$ 1.725,00. Aprovado. O professor CCC apresentou mais um demanda de um serviço que seria a manutenção do servidor e do site do DComp e está avaliando a possibilidade de contratar a empresa Júnior para cuidar destes serviços, mas que ainda é preciso conversar com os alunos para acertar detalhes e expor melhor ideia e decidir na próxima reunião.

Consulta 1 – Técnica 2

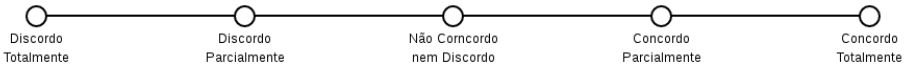
Questões:

Com base na análise dos trechos, responda as questões a seguir:

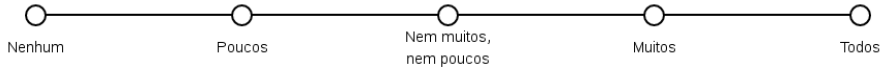
1. Todos os trechos apresentados compartilham um mesmo assunto.



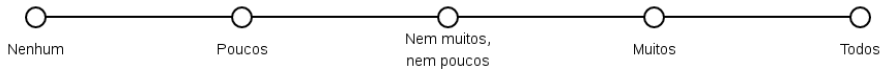
2. As palavras *verba, compra, pagamento, material e valor* resumem bem o assunto tratado nos trechos.



3. Existem trechos que não tratam de um único assunto?



4. Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?



Consulta 1 – Técnica 3

Grupo de trechos:

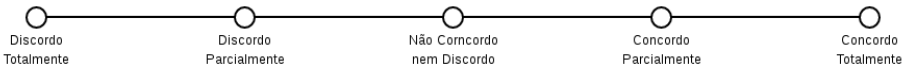
Informes: Compras: foi enviada a requisição da compra de aparelhos de ar condicionado para todas as salas do departamento e secretarias, utilizando toda verba de material permanente do departamento, com a verba do curso foram comprados dois aparelhos e um com repasse FAI e ainda ficamos com uma dívida de cerca de R$ 878,77 com o centro a qual será paga o ano que vem, ou pode vir a ser compensada após a licitação acontecer, caso consigamos comprar por um valor menor que o empenhado. As compras envolvendo mobiliário e computadores serão licitados de formas diferentes, será realizado um pedido para toda a universidade e serão comprados via ata, maiores informações de gastos verificar planilha anexa.
No momento o saldo que o departamento tem na FAI é de apenas R$ 264,91. Em relação aos ofícios que o conselho decidiu enviar a vários órgãos cobrando explicações sobre os cancelamentos das compras, a professora AAA informou que tal ofício já foi enviado ao CCGT, Setor de Compras Sorocaba, Reitoria, ProAd, ProGrad e Apoio ProAd.; no entanto só recebemos até agora a resposta do Setor de Compras Sorocaba. O ofício foi lido para todos os membros e diante da resposta do mesmo a sugestão é que para o ano que vem, 2016, todos os pedidos de compras sejam colocados no sistema o mais rápido possível. Foi informado que a impressora do prédio já foi instalada e como o custo da impressão por esta impressora é bem mais baixo foi pedido a todos que as impressões sejam feitas na impressora em questão; porém a conexão por rede ainda não está funcionando, então para imprimir ainda é necessário o uso do pen drive. O professor BBB colocou que quando é cobrado do departamento o valor é referente ao papel e a impressão em si, no entanto não tem papel disponível e cada professor está levando seu próprio papel, sendo assim o departamento irá pagar duas vezes pelo mesmo papel. Diante deste problema iremos contatar o centro para que o mesmo nos apresente uma solução. Sobre os pagamentos de pró-labores para membros de banca da pós, foi informado que para este ano temos apenas mais dois, o qual está sendo pago com saldo de empenho de 2014 e como provavelmente este empenho não irá virar para 2016 a solução encontrada para não perder dinheiro foi elevar um pouco o valor do pagamento dos pró-labores. Foi realizado um pedido de inclusão de pauta pela professora CCC, a qual pede ajuda financeira do departamento para divulgação da calourada 2016, o pedido foi aceito.
A estagiária foi contratada em 11 de junho e desde então tem auxiliado nas tarefas da secretaria dos cursos Ciência da Computação e Ciências Econômicas. 1.2.Verba: Os empenhos em Compra de materiais e pagamento estudantil estão em andamento, conforme definido na 30ª Reunião Ordinária do CoCCCS.
(II) VERBA DO CURSO: a Profa. DDD sugeriu que fosse destinada parte da verba para custear a participação dos alunos nas maratonas da computação. O aluno EEE (2009) informou que está sem patrocinadores para a realização da SECOT 2012, com isso, a Profa. DDD sugeriu destinar três empenhos de pró-labore, no valor de $R180,00 cada, com um total de R$ 540,00 para a SECOT, assim como o Prof. FFF se comprometeu em contatar algumas empresas para negociar patrocínio.
Os membros do conselho decidiu que esse dinheiro será utilizado para compra de material para infraestrutura dos laboratórios e departamento. Os membros apresentaram as seguintes sugestões: braçadeiras, alicates, expositor etiquetas, conectores de rede (fêmea), extensões para as bancadas do LEC e tomadas nos lugares que estão faltando para o LabRedes.

Consulta 1 – Técnica 3

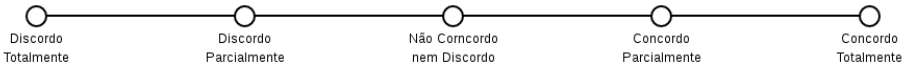
Questões:

Com base na análise dos trechos, responda as questões a seguir:

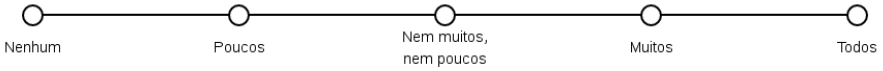
1. Todos os trechos apresentados compartilham um mesmo assunto.



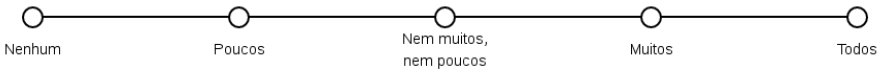
2. As palavras *verba, compra, pagamento, valor e realizada* resumem bem o assunto tratado nos trechos.



3. Existem trechos que não tratam de um único assunto?



4. Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?



Avaliação de extração de tópicos em atas de reuniões

Apresentação

Essa avaliação se refere a um sistema cujo objetivo é ajudar o usuário a fazer buscas por um assunto específico em uma coleção de atas de reunião. O sistema recebe uma consulta do usuário sobre um assunto e apresenta os trechos onde esse assunto é mencionado. Inicialmente, o sistema analisa todas atas e divide o texto de cada ata em trechos que contêm um assunto principal e relativamente independente. Ou seja, os diversos assuntos tratados em uma ata são separados em trechos com um único assunto. Em seguida, utiliza-se técnicas de inteligência artificial para identificar trechos com assuntos relacionados e agrupá-los. Cada grupo contém trechos de atas diferentes mas com assuntos relacionados. Além disso, o sistema seleciona um conjunto de palavras que indicam o assunto do grupo. Assim, espera-se que o agrupamento de trechos com assuntos similares extraídos de diferentes documentos facilite a navegação e busca por assuntos na coleção de atas.

Essa avaliação se refere a componentes internos do sistema e não em como seria o seu uso. Para isso, foi feita uma consulta ao sistema com os termos *"defesa de dissertação"* e foram extraídos alguns elementos para a avaliação dos componentes de interesse.

A seguir, você encontrará três grupos de cinco trechos de tamanhos variados. Para cada um dos grupos, há um conjunto de 4 questões simples, que você deve responder de forma a refletir a sua percepção quanto à qualidade do grupo e respectivos trechos. A avaliação deve ser feita na ordem em que os grupos e as questões foram apresentados, não importando a numeração de consulta e técnica fornecida (são apenas para controle interno).

Antes de iniciar a avaliação, identifique aqui a sua afinidade com atas de reuniões (por exemplo, presidente de conselho, membro de conselho, secretário, coordenador de curso):

Consulta 2 – Técnica 1

Grupo de trechos:

(III) Aprovado a prorrogação de prazo para defesa de dissertação dos alunos AAA, orientado pelo Prof. Dr. BBB, por 06 meses, CCC, orientado pela Prof. Dra. DDD, por 06 meses, EEE, orientado pela Profa. Dra. FFF, por 06 meses, GGG, orientado pelo Prof. Dr. BBB, por 06 meses, HHH, orientado pela Profa. Dra. DDD, por 06 meses; (IV) Homologado o relatório de defesa de dissertação do aluno III, orientado pela Profa. Dra. DDD;
(III) aprovado o pedido de prorrogação de prazo para defesa de dissertação do aluno AAA, orientado pelo Prof. Dr. BBB por mais 06 meses com prazo final para setembro de 2015;
Aprovados os requerimentos de prorrogação de prazo para defesa de dissertação: AAA pelo Prof. Dr. BBB por mais 06 meses com prazo final para janeiro de 2016, CCC filho orientado pelo Prof. Dr. DDD por mais 06 meses com prazo final para janeiro de 2016, EEE orientada pela Profa. Dra. FFF por mais 06 meses com prazo final para janeiro de 2016, GGG orientado pelo Prof. Dr. HHH por mais 06 meses com prazo final para janeiro de 2016, III orientada pela Profa. Dra. HHH por mais 06 meses com prazo final para fevereiro de 2016, JJJ orientado pelo Prof. Dr. KKK por mais 06 meses com prazo final para janeiro de 2016, LLL orientado pelo Prof. Dr. MMM por mais 06 meses com prazo final para janeiro de 2016, NNN orientada pelo Prof. Dr. OOO por mais 06 meses com prazo final para janeiro de 2016; (IX) Homologado o relatório de defesa de dissertação do aluno PPP, orientado pelo Prof. Dr. DDD, realizada em 08 de abril; Nada mais, Prof. Dr. QQQ Presidente desta Reunião Prof. Dr. HHH Professor Adjunto Prof. Dr. MMM Professor Associado Profa. Dra. FFF Professora Adjunta RRR Representante Discente 87
Também foi discutida a situação de alunos que recebem a bolsa no decorrer do curso e apresentam o exame de defesa antes do término do prazo da bolsa. A duração da bolsa oferecida pela CAPES é de 24 meses e foi apresentada a proposta do aluno continuar com a bolsa caso solicite prorrogação do prazo de defesa da dissertação, desde que mantenha os requisitos descritos na norma 5 e não tenha nenhum aluno na lista de espera do ranking de bolsas elaborado pela Comissão de Bolsas de Estudos do PPGCCS. Foi decidido retirar os seguintes requisitos da norma 5: 1. b) Não ter completado 12 (doze) meses corridos a contar da data de sua primeira matrícula no curso de Mestrado, exceto no caso de renovação. 3. d) ultrapassar 24 (vinte e quatro) meses no Programa, ou 3. e) não tiver cumprido a Norma 6 sobre obrigatoriedade de solicitação de Bolsa FAPESP.
(II) Aprovado o pedido de agendamento de exame de defesa de dissertação do aluno AAA, orientado pela Profa. Dra. BBB, para o dia 24 de agosto, com banca formada pela Profa. Dra. CCC, da USP e o Prof. Dr. DDD; (III) Aprovados os pedidos de prorrogação de prazo para defesa de dissertação do aluno EEE orientado pelo Prof. Dr. FFF por mais 06 meses com prazo final para fevereiro de 2016, GGG orientado pela Profa. Dra. HHH por mais 06 meses com prazo final para fevereiro de 2016;

Consulta 2 – Técnica 1

Grupo de trechos:

(III) Aprovado a prorrogação de prazo para defesa de dissertação dos alunos AAA, orientado pelo Prof. Dr. BBB, por 06 meses, CCC, orientado pela Prof. Dra. DDD, por 06 meses, EEE, orientado pela Profa. Dra. FFF, por 06 meses, GGG, orientado pelo Prof. Dr. BBB, por 06 meses, HHH, orientado pela Profa. Dra. DDD, por 06 meses; (IV) Homologado o relatório de defesa de dissertação do aluno III, orientado pela Profa. Dra. DDD;
(III) aprovado o pedido de prorrogação de prazo para defesa de dissertação do aluno AAA, orientado pelo Prof. Dr. BBB por mais 06 meses com prazo final para setembro de 2015;
Aprovados os requerimentos de prorrogação de prazo para defesa de dissertação: AAA pelo Prof. Dr. BBB por mais 06 meses com prazo final para janeiro de 2016, CCC filho orientado pelo Prof. Dr. DDD por mais 06 meses com prazo final para janeiro de 2016, EEE orientada pela Profa. Dra. FFF por mais 06 meses com prazo final para janeiro de 2016, GGG orientado pelo Prof. Dr. HHH por mais 06 meses com prazo final para janeiro de 2016, III orientada pela Profa. Dra. HHH por mais 06 meses com prazo final para fevereiro de 2016, JJJ orientado pelo Prof. Dr. KKK por mais 06 meses com prazo final para janeiro de 2016, LLL orientado pelo Prof. Dr. MMM por mais 06 meses com prazo final para janeiro de 2016, NNN orientada pelo Prof. Dr. OOO por mais 06 meses com prazo final para janeiro de 2016; (IX) Homologado o relatório de defesa de dissertação do aluno PPP, orientado pelo Prof. Dr. DDD, realizada em 08 de abril; Nada mais, Prof. Dr. QQQ Presidente desta Reunião Prof. Dr. HHH Professor Adjunto Prof. Dr. MMM Professor Associado Profa. Dra. FFF Professora Adjunta RRR Representante Discente 87
Também foi discutida a situação de alunos que recebem a bolsa no decorrer do curso e apresentam o exame de defesa antes do término do prazo da bolsa. A duração da bolsa oferecida pela CAPES é de 24 meses e foi apresentada a proposta do aluno continuar com a bolsa caso solicite prorrogação do prazo de defesa da dissertação, desde que mantenha os requisitos descritos na norma 5 e não tenha nenhum aluno na lista de espera do ranking de bolsas elaborado pela Comissão de Bolsas de Estudos do PPGCCS. Foi decidido retirar os seguintes requisitos da norma 5: 1. b) Não ter completado 12 (doze) meses corridos a contar da data de sua primeira matrícula no curso de Mestrado, exceto no caso de renovação. 3. d) ultrapassar 24 (vinte e quatro) meses no Programa, ou 3. e) não tiver cumprido a Norma 6 sobre obrigatoriedade de solicitação de Bolsa FAPESP.
(II) Aprovado o pedido de agendamento de exame de defesa de dissertação do aluno AAA, orientado pela Profa. Dra. BBB, para o dia 24 de agosto, com banca formada pela Profa. Dra. CCC, da USP e o Prof. Dr. DDD; (III) Aprovados os pedidos de prorrogação de prazo para defesa de dissertação do aluno EEE orientado pelo Prof. Dr. FFF por mais 06 meses com prazo final para fevereiro de 2016, GGG orientado pela Profa. Dra. HHH por mais 06 meses com prazo final para fevereiro de 2016;

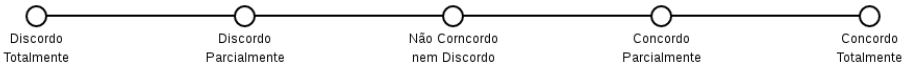
Consulta 2 – Técnica 1

Questões:

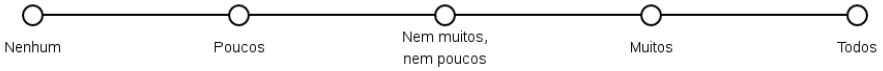
Com base na análise dos trechos, responda as questões a seguir:

1. Todos os trechos apresentados compartilham um mesmo assunto.

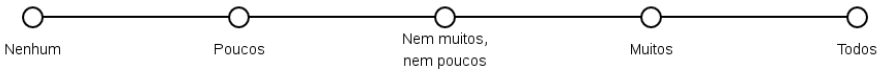
2. As palavras *orientada, meses, defesa, prazo e dissertação* resumem bem o assunto tratado nos trechos.



3. Existem trechos que não tratam de um único assunto?



4. Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?



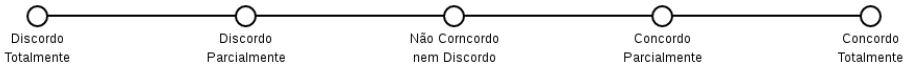
Consulta 2 – Técnica 2

Questões:

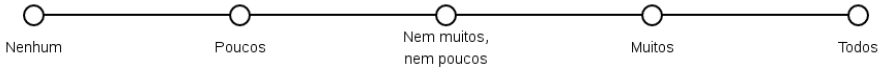
Com base na análise dos trechos, responda as questões a seguir:

1. Todos os trechos apresentados compartilham um mesmo assunto.

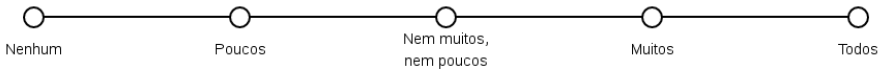
2. As palavras *aprovado, defesa, pedido, dissertação e orientada* resumem bem o assunto tratado nos trechos.



3. Existem trechos que não tratam de um único assunto?



4. Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?



Consulta 2 – Técnica 3

Grupo de trechos:

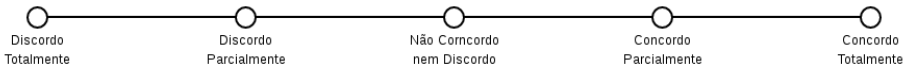
(III) Aprovado o pedido de prorrogação de prazo para defesa de dissertação do aluno AAA orientado pelo Prof. Dr. BBB por mais 06 meses com prazo final para fevereiro de 2016, CCC orientada pela Profa. Dra. DDD por mais 06 meses com prazo final para fevereiro de 2016, EEE orientado pela Profa. Dra. DDD por mais 06 meses com prazo final para fevereiro de 2016;
(II) Aprovado o pedido de prorrogação de prazo para defesa de dissertação do aluno AAA por mais 06 meses com prazo final para fevereiro de 2016. A partir de agora o formulário para solicitação de prorrogação de prazo para defesa de dissertação possui a opção de prorrogar a bolsa de estudos: No último mês de vigência da bolsa de estudos, será consultada a lista de espera por bolsas do PPGCCS. A solicitação de prorrogação de bolsa só será analisada caso não existam candidatos na fila de espera e se a vigência da bolsa não tiver atingido 24 meses. Neste caso, se a prorrogação for atendida, o novo prazo de vigência da bolsa será estendido para coincidir com a data prevista da defesa ou até que se complete 24 meses de bolsa (o que ocorrer primeiro);
(III) aprovado o pedido de prorrogação de prazo para defesa de dissertação do aluno AAA, orientado pelo Prof. Dr. BBB por mais 06 meses com prazo final para setembro de 2015;
Comunicação da Presidência: a presidente do Conselho, Profa. Dra. AAA comunicou que: a Coordenação do curso assumirá a Secot em 2015 ; foi agendado para 25/11/2014 o pregão para a compra dos roteadores; foi feita a distribuição das salas do novo prédio do CCGT, que terá sala para os professores, salas de estudo, auditório, sala para as secretarias (coordenações de cursos de graduação, pós-graduação e departamentos); no dia 29/11/2014, os alunos inscritos no ENADE foram convocados para uma reunião com a profª. CCC, pró-reitora de graduação adjunta, que enfatizou sobre a importância da participação dos alunos na prova marcada para o dia 23/11/2014, e avisou que o aluno concluinte que não fizer a prova ficará em débito com o MEC e não terá seu diploma impresso; a Comissão Própria de Avaliação (CPA) da UFSCAR fará uma avaliação interna dos cursos que participam do ENADE este ano e conta com a participação dos docentes e discentes deste curso. 2. Comunicação dos Conselheiros - a professora AAA comunicou que em 07/11/2014 começa a final da XIX Maratona de Programação, que acontecerá em Fortaleza.
Divulgação da lista de candidatos: 25 de maio; Prazo para recursos da lista de candidatos, 26 de maio até às 12 horas; Divulgação da lista final de candidatos: 26/5;

Consulta 2 – Técnica 3

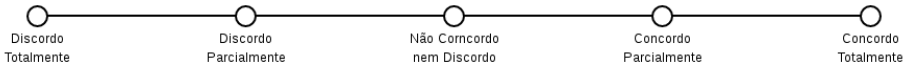
Questões:

Com base na análise dos trechos, responda as questões a seguir:

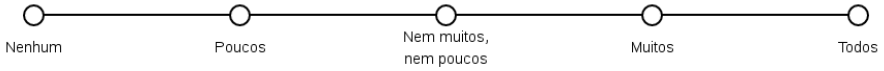
1. Todos os trechos apresentados compartilham um mesmo assunto.



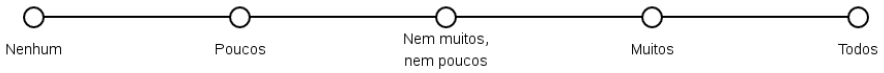
2. As palavas *orientada, prazo, bolsa, meses e defesa* resumem bem o assunto tratado nos trechos.



3. Existem trechos que não tratam de um único assunto?



4. Existem trechos incompletos e insuficientes para compreensão do assunto do trecho?



APÊNDICE C – Distribuição dos Tópicos obtidos pelos extratores

Nesse apêndice podem ser observadas as distribuições dos tópicos obtidos pelos extratores, conforme discutido no Capítulo 3.

K-Means		LDA		PLSA	
Descritores	*#Seg*	*Descritores*	*#Seg*	*Descritores*	*#Seg*
dia; realizada; chamada; estado; conselho;	116	disciplinas; cursadas; fichas; caracterização; aprovado;	107	docentes; presidente; dia; discente; téc;	76
informado; compra; ofício; pedido; processo;	106	colocar; deve; poderia; referentes; seriam;	94	disciplinas; álgebra; linear; geometria; analítica;	75
computação; conselho; aprovado; acordo; ficou;	102	docentes; presidente; técnica; dia; administrativo;	91	computação; acordo; levada; chefia; conselho;	62
docentes; técnica; administrativo; presidente; dia;	72	dia; aprovado; aprovação; ordem; anterior;	85	aprovado; aprovação; unanimidade; foram; ordinária;	57
representante; discente; presidente; secretária; turma;	55	representante; técnica; administrativo; secretária; discente;	79	representante; discente; piccoli; turma; pauta;	51
cursadas; conselho; coordenação; computação; presidente;	45	conselho; junto; assina; acordo; ficou;	69	técnica; administrativo; representante; secretária; discente;	45
aprovado; aprovação; atividades; relatórios; lido;	44	seguintes; chamada; conselho; data; ano;	67	comunicação; presidência; informado; verba; ocorrerá;	38
computação; cursadas; conselho; título; discente;	37	presença; realizada; cidade; leme; métricas;	55	bacharelado; coordenação; cursadas; encerrada; acordo;	37
disciplinas; cursadas; libras; conselho; aprovado;	36	presidência; comunicação; informado; iniciou; agradeceu;	54	afastamento; aprovado; aprovação; apresentar; referentes;	35
professores; colocar; regras; seriam; sugeriu;	30	atividades; extensão; relatórios; aprovado; aprovação;	52	dia; ordem; anterior; informado; pedido;	35
pedido; informado; substituí; laboratório; disciplinas;	30	havendo; lavra; iniciou; trabalho; trata;	46	discente; representante; lúcio; presidente; seki;	34
aprovado; trocar; pedido; analisada; fichas;	28	verba; compra; pagamento; valor; material;	46	informado; compra; ofício; material; verba;	34
dia; ordem; concurso; dezembro; foram;	27	afastamento; aprovado; aprovação; relatórios; lido;	44	presidente; presentes; lavra; havendo; comunicou;	30
representante; administrativo; técnica; discente; secretário;	26	coordenação; deliberar; restrito; junto; assina;	39	dia; seguintes; presidente; realizada; participante;	28
afastamento; aprovado; referentes; relatórios; final;	26	discente; representante; presidente; lúcio; seki;	34	relatórios; aprovado; lido; aprovação; atividades;	28
extensão; atividades; coordenadores; programa; relatórios;	25	computador; tópicos; disciplinas; software; dados;	30	dia; realizada; gestão; tecnologia; centrada;	27
compra; informado; verba; tem; valor;	23	disciplinas; calculada; diferentes; introdução; algoritmos;	27	dcomp; semestre; calendário; anexo; data;	26
aprovado; conselho; orientada; pedido; meses;	22	gestão; conhecimento; conselho; computação; sistema;	26	título; suplente; computação; iniciou; havendo;	25
dia; ordem; aprovação; anterior; aprovado;	22	processo; semestre; seletivo; resultado; ingresso;	26	fichas; caracterização; obrigatório; disciplinas; computação;	20
presidente; secretária; associado; chistine; dia;	20	laboratório; máquina; técnica; pedido; manutenção;	22	chamada; terceiro; dia; segunda; estado;	19
unanimidade; aprovado; conselho; ordinária; apreciação;	19	aprovado; defesa; pedido; dissertação; orientada;	17	dia; realizada; estado; presentes; cidade;	17
foram; aprovado; lido; relatórios; documentos;	19	conselho; cursadas; sala; computação; ano;	16	cursadas; disciplinas; coordenação; optativas; conselho;	17
comunicação; presidência; presidente; conselho; computação;	18	concurso; dados; bancos; vaga; comunicação;	15	explicou; enviar; aprovação; trazido; pauta;	17
processo; seletivo; semestre; resultado; aprovado;	18	pauta; inclusão; pedido; aceita; suplente;	14	extensão; atividades; coordenadores; aprovado; proposta;	17
secretária; representante; presentes; técnica; administrativo;	17	condicionado; informado; compra; faltando; aparelhos;	13	computação; cursadas; conselho; representante; discente;	16
fichas; caracterização; disciplinas; aprovação; aprovado;	16	discussão; decidido; regras; colocar; seriam;	12	computador; sistema; software; disciplinas; engenharia;	16
aprovação; aprovado; política; atas; anterior;	16	próxima; trazido; tomadas; informado; trouxe;	12	atividades; extensão; processo; edital; coordenadores;	16
computação; teoria; paralela; tópicos; tutoria;	13	aprovado; aprovação; referentes; vista; política;	10	pedido; deve; informado; realizada; aulas;	16
candidatos; concurso; lista; divulgação; bolsa;	12	pedido; atendida; compra; professores; informado;	10	cursadas; recurso; dcomp; laboratório; contar;	15
semestre; conceito; fronteiras; manutenção; esquema;	12	implantação; serviços; horária; prestados; unidades;	10	orientada; prazo; meses; defesa; final;	15
aprovação; realizada; laboratório; correções; trazido;	11	votação; votaram; equipe; verificaram; opção;	9	dia; conceito; laboratório; chamada; primeiro;	14
redigida; lavra; presidente; trabalho; coube;	10	foram; material; estado; detalhes; conseguiu;	8	extensão; programa; coordenadores; tecnologia; semana;	14
deve; normalizado; assunto; estudos; desejamos;	10	foram; conselho; aprovado; realizada; referendum;	7	projeto; comissão; esclarecido; faltando; bolsa;	13
havendo; legal; número; iniciou; introdução;	10	site; informado; dcomp; laboratório; ficou;	7	ficou; novo; colocar; regimento; votação;	13
realizada; pagamento; apresentação; the; learning;	9	deve; laboratório; aprovado; controlados; manter;	6	planos; ensino; foram; disciplinas; adequação;	13
aprovação; anterior; máquina; aprendizado; atas;	9	learning; the; and; artigo; apresentação;	6	provas; candidatos; presidente; solicitada; concurso;	12
pauta; inclusão; pedido; aceita; conselho;	8	projeto; extensão; mudanças; atividades; desenvolvimento;	6	professores; cursadas; justificativa; computação; ausência;	12
presentes; lavra; junto; assina; castilho;	8	aprovado; dcomp; proposta; foram; processo;	3	área; concurso; problemas; enviar; cursadas;	12
presidente; docentes; dia; cancelamento; governo;	7	sugeriu; mail; poderia; enviar; contato;	2	valor; compra; empenho; auxílio; verba;	12
ausência; justificativa; solicitação; períodos; afastamento;	7	informado; aprovado; comissão; eleição; deve;	1	graduação; pós; min; dia; realizada;	11
técnica; administrativo; docentes; dia; eihara;	7	informado; aprovado; dia; deve; substituí;	1	vaga; transferência; foram; informática; cursadas;	11
presidência; comunicações; iniciou; comunicação; presidente;	6	aprovado; informática; informado; dia; dcomp;	1	bancos; dados; aprovado; estrutura; proposta;	11
comunicou; comunicação; conselho; presidente; presentes;	6	deve; lista; informado; conselho; aprovado;	1	demanda; compra; pedido; verba; professores;	10
dados; bancos; ccs; software; engenharia;	6	informado; deve; aprovado; dia; conselho;	1	verba; cursadas; pagamento; disciplinas; calculada;	10
informática; sociedade; docentes; ética; casadei;	6	deve; informática; conselho; informado; aprovado;	1	laboratório; manutenção; suplente; trocar; cabo;	10
estudos; liberados; instalação; apuração; eleição;	6	conselho; informado; dia; aprovado; ordem;	1	área; criação; conselho; projeto; suplente;	9
processo; participação; ficou; imagens; sinais;	6	conselho; dia; aprovado; informado; próxima;	1	laboratório; disciplinas; votaram; proposta; créditos;	9
extensão; atividades; coordenadores; mitsuru; lúcio;	6	informática; conselho; informado; dcomp; comissão;	1	disciplinas; cursadas; oferta; horário; oferecida;	9
presidência; informado; comunicação; iniciou; presença;	5	aprovado; deve; informado; conselho; comissão;	1	manutenção; informado; laboratório; conseguiu; evento;	9
votaram; conselho; favor; eleitos; relação;	5	informática; informado; conselho; aprovado; deve;	1	colocar; lecionar; disciplinas; fechaduras; realizada;	9
iniciou; poderia; gestante; encontrados; compareceram;	5	conselho; informado; aprovado; final; deve;	1	laboratório; conselho; colocar; disciplinas; nus;	9
unidades; positivo; leitura; proposta; alterações;	5	aprovado; conselho; informado; recurso; eleição;	1	estágio; cursadas; atividades; coordenação; complementares;	8
discussão; dcomp; normalizado; material; item;	5	deve; dia; bolsa; lista; próxima;	1	colocar; diz; esclarecido; professores; seriam;	8
havendo; trata; deus; encerrada; presentes;	5	informado; conselho; deve; eleição; aprovado;	1	conselho; computação; cursadas; reuniu; ordinária;	7
votação; entrou; solicitada; transferência; responsável;	5	aprovado; unanimidade; conselho; informado; dia;	1	candidatos; dados; concurso; palestra; área;	6
recurso; foram; adequação; encaminhada; levantadas;	5	próxima; informado; aprovado; conselho; proposta;	1	sala; informado; orçamento; valor; site;	6
avaliação; extensão; discussão; projeto; ficou;	5	aprovado; deve; dcomp; informado; unanimidade;	1	proposta; colocar; copq; pesquisa; terá;	6
associado; presidente; chistine; existam; discente;	4	conselho; informado; dia; pedido; deve;	1	cursadas; existam; pedido; ofício; laboratório;	6
persianas; pedido; foram; cota; procurando;	4	palestra; apoio; sustentabilidade; informática; documentos;	1	bolsa; cursadas; distribuição; realizada; disciplinas;	6
estágio; apresentação; secretário; lúcio; eduardo;	4	próxima; conselho; discussão; informado; proposta;	1	disciplinas; tem; colocar; acadêmico; seriam;	6
gestão; ambiental; noções; vaga; distribuídos;	4	informado; conselho; dia; deve; aprovado;	1	presidente; discussão; cursadas; acrescentou; professores;	5
piccoli; ausência; justificativa; cancelamento; governo;	4	implantação; horária; serviços; dia; prestados;	1	junto; assina; participante; lavra; estado;	5
calendário; apresentar; dcomp; dia; ordinária;	4	informado; deve; transferência; informática; dcomp;	1	colocar; verba; poderia; informado; terá;	5
sistema; técnica; relatórios; publicação; proposta;	3	aprovado; dia; conselho; deve; informado;	1	chefia; seriam; criar; conselho; conflito;	5
gasto; verba; custeio; veio; lista;	3	participação; evento; palestra; sustentabilidade; partir;	1	capacitação; afastamento; regras; docentes; discussão;	5
sistema; operacionais; ccs; gestão; planos;	3	comissão; eleição; aprovado; conselho; dcomp;	1	seriam; edital; colocar; item; terá;	5
semana; estudos; perfil; ficou; apresentar;	2	deve; dia; conselho; aprovado; informado;	1	cursadas; proposta; atividades; esclarecido; colocar;	4
professores; férias; resposta; unanimidade; questões;	2	conselho; informado; substituição; aprovado; deve;	1	ano; dia; interessados; realizada; mês;	3
moraes; assunto; trata; breve; discussão;	2	dia; deve; conselho; docentes; aprovado;	1	projeto; foram; pesos; pedido; texto;	3
fapesp; rti; coccs; ordinária; extraordinária;	2	conselho; pedido; dia; informática; informado;	1	disciplinas; requisitos; pré; professores; colocar;	1

Tabela 22 – Distribuições dos tópicos obtidos pelos extratores.