UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Investigação de Mapas de Relacionamento para Sumarização Multidocumento

Rafael Ribaldo

Investigação de Mapas de Relacionamento para Sumarização Multidocumento

Rafael Ribaldo

Orientador: Prof. Dr. Thiago A. S. Pardo

Monografia de conclusão de curso apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP - para obtenção do título de Bacharel em Ciências de Computação.

Área de Concentração: Processamento de Linguagem Natural

USP – São Carlos
Novembro de 2013

"My interest is in the future because I am going to spend the rest of my life there."

~ Charles Kettering

Dedicatória

À Deus que tudo me proporciona na vida.

À minha mãe Silvia e ao meu pai Joaquim, os quais amo muito, pelo exemplo de vida e família.

À minha irmã Vivian por estar sempre ao meu lado.

À minha namorada Isis, pelo carinho, compreensão e companheirismo.

E ao meu professor orientador Thiago, por todo o incentivo e ajuda.

Resumo

A necessidade de se ter uma ferramenta que lide com diversos textos de assuntos relacionados e consiga extrair conteúdos relevantes dos mesmos é, atualmente, cada vez maior. Neste contexto, tem-se a tarefa de sumarização automática multidocumento, a qual consiste em produzir automaticamente um único sumário a partir de um grupo de textos sobre um mesmo assunto. Por isso, neste projeto, são exploradas estratégias de construção de sumário com base nos chamados "mapas de relacionamento" propostos por Salton et al. (1997), em que se tenta representar no sumário os principais subtópicos presentes nos textos de origem ao mesmo tempo em que se faz a manutenção da informatividade do sumário produzido. Em especial, o trabalho foi desenvolvido e avaliado para textos em língua portuguesa, dando continuidade às pesquisas já realizadas na comunidade de pesquisa brasileira. Em adição, foi desenvolvida uma ferramenta web, de uso geral, que incorpora algumas técnicas estudadas.

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1: INTRODUÇÃO

1.1. Contextualização e Motivação

Nas últimas décadas, muitas tecnologias novas têm surgido, trazendo com isso um crescente aumento no volume de informação. Hoje em dia, muitos recursos como buscadores de notícias, blogs e redes sociais fazem accessível uma enorme quantidade de informação e, em consequência, o processamento desta se faz cada vez mais difícil. Para se ter uma ideia, o informe da *International Data Corporation* (IDC)[1] mostra que somente em 2012 a Web foi responsável pela disponibilização de aproximadamente 2.7 zettabytes[2], uma quantidade nove vezes maior do que a produzida cinco anos atrás. A maioria desta informação encontra-se em formato não estruturado, textual, com muitas informações similares e diferentes. Neste contexto, a tarefa de Sumarização Multidocumento mostra-se como um recurso importante.

A Sumarização Multidocumento consiste na produção de um único sumário a partir de um conjunto de documentos que versam sobre um mesmo assunto (Mani, 2001), sendo que este sumário deve conter as informações mais relevantes ao tópico em questão e lidar com os fenômenos multidocumento, como informações redundantes, complementares e contraditórias, estilos de escrita variados (já que os textos provêm de diferentes autores), ordenação temporal dos eventos/fatos (pois os textos são escritos em diferentes momentos) e perspectivas e focos diferentes, assim como a própria questão da coerência e coesão do sumário. Esta tarefa surgiu como uma extensão natural da tradicional Sumarização Monodocumento, que visa à construção de um sumário a partir de um único documento.

A Sumarização Monodocumento tem sido bastante explorada e discutida por vários autores (por exemplo, Luhn, 1958; Baxendale, 1958; Edmundson, 1969; O'Donnell, 1997; Marcu, 2000; Conroy e O'leary, 2001; Pardo e Rino, 2002; Pardo et al., 2003; Salton et al., 1997; Svore et al., 2007; Uzêda et al., 2009, 2010). Por outro lado, a sumarização

[1] www.idc.com/research/Predictions12/Main/index.jsp
[2] Unidade de medida de informação que corresponde a 2^{70} Bytes

multidocumento representa uma área mais nova que tem adquirido relevância nos últimos anos. As primeiras pesquisas datam dos anos 90 (McKeown e Radev, 1995; Carbonell e Goldstein, 1998) e as investigações continuam até os anos mais recentes (por exemplo, Radev et al., 2000; Zhang et al., 2002; Otterbacher et al., 2002; McKeown et al., 2005; Wan e Yang, 2006; Afantenos et al., 2004, 2008; Castro Jorge e Pardo, 2010, 2011), motivadas pela relevância da aplicação da sumarização em importantes sistemas de recuperação de informação, como buscadores de notícias (ex. Google News[3], Wiki News[4]) ou sistemas de bibliotecas digitais (por exemplo, CiteSeer[5], DBLP[6]). Por exemplo, poder-se-ia construir um sumário contendo as informações mais relevantes do que foi noticiado sobre as recentes manifestações no estado de São Paulo.

No Brasil, as pesquisas em sumarização multidocumento com a língua portuguesa são mais recentes, iniciando-se oficialmente em 2005 com o sistema simples chamado GistSumm (Pardo, 2005), mas somente recentemente produzindo recursos e sistemas de ponta, como o córpus de referência CSTNews (Cardoso et al. 2011) e sistemas de sumarização do estado da arte das linhas superficial (por exemplo, Ribaldo et al., 2011, 2012; Akabane et al., 2011) e profunda (Castro Jorge e Pardo, 2010, 2011; Castro Jorge et al., 2011; Cardoso et al., 2011). Sistemas ditos superficiais são aqueles que fazem pouco ou nenhum uso de conhecimento linguístico, sendo mais escaláveis e robustos, em geral. Sistemas da linha profunda, por outro lado, fazem uso massivo de conhecimento linguístico, como gramáticas, repositórios semânticos e modelos de discurso, sendo capazes de produzir melhores resultados, mas sendo mais caros e de aplicação mais restrita, normalmente.

Como ilustração dos avanços na área, a Figura 1 mostra um sumário automático multidocumento produzido a partir de um grupo de três textos (pertencentes ao córpus CSTNews, citado anteriormente) sobre um terremoto no Japão em 2007. O sumário foi gerado pelo sistema GistSumm. Como se pode notar, há vários problemas de redundância,

[3]http://news.google.com.br/
[4]http://pt.wikinews.org/
[5]http://citeseerx.ist.psu.edu/
[6]http://dblp.uni-trier.de/

coerência e coesão no sumário, já que o GistSumm é um sistema muito simples que baseia sua seleção de sentenças para o sumário na presença de palavras frequentes, não tratando apropriadamente os desafios multidocumento.

> O tremor atingiu a região às 10h13 (horário local, 22h13 de domingo, em Brasília) e seu epicentro foi localizado a 260 km da costa de Niigata, ao nordeste da capital, Tóquio, onde também foi sentido. Um terremoto de 6,8 graus na escala Richter atingiu a costa noroeste do Japão nesta segunda-feira, 16, matando pelo menos sete pessoas na cidade de Kashiwazaki e deixando outros 700 feridos. O terremoto de 7,4 graus, que pôde ser sentido em Tóquio, foi seguido por outro tremor de menor magnitude, de 4,2 graus na escala Ritcher, às 10h34m (22h34m de domingo em Brasília).

Figura 1 – Exemplo de sumário multidocumento produzido pelo sistema GistSumm

A Figura 2 mostra um sumário automático para os mesmos textos produzidos pelo sistema CSTSumm (Castro Jorge e Pardo, 2010). O CSTSumm é um dos melhores sistemas atuais para o português e faz uso de um modelo sofisticado de discurso que relaciona as partes dos vários textos a serem sumarizados, o que permite lidar apropriadamente com grande parte dos fenômenos multidocumento.

> Um terremoto de 6,8 graus na escala Richter atingiu a costa noroeste do Japão nesta segunda-feira, 16, matando pelo menos sete pessoas na cidade de Kashiwazaki e deixando outros 700 feridos. O terremoto, que pôde ser sentido em Tóquio, foi seguido por outro tremor de menor magnitude, de 4,2 graus na escala Ritcher, às 10h34m (22h34m de domingo em Brasília). Chamas e rolos de fumaça preta foram vistos na usina nuclear de Kashiwazaki, que foi automaticamente fechada durante o terremoto.

Figura 2 – Exemplo de sumário multidocumento produzido pelo sistema CSTSumm

Pode-se notar, no sumário acima, uma sensível melhora nos resultados, em relação ao sumário da Figura 1. Ainda assim, os sistemas atuais para o português não atacam todos os problemas encontrados na área da sumarização multidocumento, por exemplo, não há um tratamento adequado de textos que contenham passagens com variações de subtópicos, mesmo que estes estejam relacionados ao tópico principal dos textos. Como ilustração,

observe os dois textos da Figura 3. O tópico de ambos é o caso de saúde do ex-jogador argentino Maradona, entretanto, pode-se encontrar subtópicos sobre a doença em si e a partida de futebol entre o Boca Juniors e o River Plate.

Texto 1

O médico pessoal do argentino Diego Maradona, Alfredo Cahe, revelou nesta segunda-feira que uma recaída da hepatite aguda de que sofre foi o motivo da nova internação do ex-craque.

Maradona havia recebido alta no último dia 11, mas voltou a ser internado na sexta-feira e os boletins médicos não especificaram o que se passava com o ex-jogador -- Cahe descartou pancreatite ou úlcera.

"Maradona teve uma recaída na hepatite aguda. Agora está estável. Apesar de ter melhorado no domingo, deverá continuar internado", disse Cahe, em declarações ao jornal "La Nación".

Maradona, 46, desenvolveu um hepatite tóxica por excesso de consumo de álcool, o que já o manteve internado durante 13 dias antes da primeira alta.

Cahe disse ainda que Maradona não voltou a consumir bebidas alcoólicas e que as causas da recaída estão sendo investigadas.

Texto 2

Maradona voltou a ter problemas de saúde no fim de semana. Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe.

"Agora está estável. Mesmo com esta melhora, ele continuará internado", disse o médico, que descartou a possibilidade do ex-jogador ter uma pancreatite (inflamação do pâncreas, órgão situado atrás do estômago e que influencia na digestão). Cahe reforçou que Maradona ainda tem problemas. "Os valores hepáticos dele na avaliação não estão equilibrados e ele não está bem. Mas não é nada grave", afirma, em entrevista ao diário La Nación.

No domingo, Maradona assistiu ao empate por 1 a 1 no clássico Boca Juniors e River Plate pela televisão. Os torcedores do Boca, que compareceram em grande número ao Estádio La Bombonera, levaram muitas faixas e bandeiras com mensagens de apoio ao ídolo argentino. Sua filha, Dalma, foi ao estádio assistir ao jogo.

Figura 3 – Exemplo de textos com assuntos diferentes

Percebe-se nos textos acima, principalmente no Texto 2, que o primeiro subtópico "Maradona com problemas de saúde" (primeiro parágrafo) se mostra diferente de "Maradona assistiu ao jogo" (último parágrafo). Por esta razão, o sumário envolvendo estes dois textos pode não conter todas as informações importantes descritas nos textos da Figura 3, ou melhor, as passagens selecionadas podem estar desconexas (subtópicos sem transições).

1.2. Objetivo

Neste contexto, este projeto tem como objetivo adaptar e explorar uma técnica clássica para a sumarização multidocumento que ataca o problema mencionado acima. Em particular, dá-se continuidade ao trabalho de iniciação científica anterior (Ribaldo et al., 2012), no qual foram obtidos bons resultados com alguns dos métodos baseados em mapas de relacionamento de Salton et al. (1997). Neste trabalho, explora-se mais um desses métodos, o qual aborda a questão de subtópicos. Para isso, propõe-se, em primeiro lugar, identificar os subtópicos presentes nos diversos textos de um mesmo tópico, utilizando técnicas que têm se mostrado promissoras para tal tarefa, tanto superficiais (utilizando pouco conhecimento linguístico) quanto profundas (que utilizam mais conhecimento). O segundo passo deve considerar que subtópicos de diferentes textos (cenário multidocumento) podem, eventualmente, ser semelhantes, uma vez que textos que versam sobre um mesmo assunto (tópico) podem conter passagens muito parecidas. Para isso, necessita-se realizar um agrupamento/junção deste subtópicos. Por fim, usa-se o método de Salton et al. citado anteriormente para a construção de sumários a partir dos segmentos textuais relevantes detectados.

1.3. Organização da Monografia

A seguir, no Capítulo 2, a revisão literária sobre os conceitos estudados para a realização desde projeto é feita, seguido do Capítulo 3, onde são apresentados mais detalhes de cada etapa/atividade desenvolvida. Por fim, no quarto capítulo, algumas considerações finais são feitas.

CAPÍTULO 2: REVISÃO BIBLIOGRÁFICA

2.1. Considerações Iniciais

Neste capítulo, são introduzidos os conceitos e técnicas relevantes utilizados neste trabalho. São apresentados (1) conceitos mais formais relacionados a sumários, bem como modelos clássicos da sumarização automática, incluindo o método de Salton et al. (1997) utilizado neste projeto; (2) técnicas utilizadas para o descobrimento de subtópicos; (3) algoritmos para o eventual agrupamento de subtópicos semelhantes e (4) um modelo discursivo que pode auxiliar no tratamento de subtópicos nas próximas etapas deste trabalho.

2.2. Sumarização

Segundo Mani e Maybury (1999), sumários podem ser classificados de várias formas. Quanto à informação que contêm, sumários podem ser de 3 tipos: indicativos, informativos e críticos/avaliativos. Sumários indicativos contêm apenas os tópicos essenciais dos textos-fonte, não necessariamente contendo detalhes de resultados, argumentos e conclusões. Índices são bons exemplos de sumários indicativos. Sumários informativos, por sua vez, são considerados substitutos dos textos, devendo conter todos os seus aspectos principais. Estes são os mais tradicionais. *Abstracts* de artigos são ótimos representantes deste tipo de sumário. Sumários críticos, além de sumarizar o conteúdo dos textos, adicionam crítica em relação ao conteúdo. As resenhas de livros são exemplos de sumários críticos.

Em termos de formação, sumários podem ser classificados como extratos ou *abstracts*. Extratos são sumários compostos por trechos inalterados dos textos. Eles são construídos por operações de cópia e cola de trechos integrais dos textos, literalmente. *Abstracts*, por sua vez, apresentam partes (ou mesmo tudo) reescritas, ou seja, há algum nível de modificação na estrutura e/ou significado dos trechos extraídos dos textos.

Há ainda tipos de sumários que são formados, ou não, de acordo com uma determinada consulta, por exemplo, é muito comum alguém que deseja um sumário querer que este seja focado em um determinado tópico ou palavra-chave. Por isso, surge a

sumarização focada nos interesses do usuário, em oposição à genérica, em que se produz um sumário para uma ampla audiência. A sumarização genérica acontece de forma natural: modela-se o documento utilizando alguma estrutura de dados, dá-se importância aos segmentos textuais e, por fim, selecionam-se os mais salientes para compor o sumário final. Tal processo ocorre um pouco diferente quando se trata dos interesses do usuário, pois quando se tem alguma palavra-chave, por exemplo, a ponderação dos segmentos deve mudar de acordo com ela. Assim, todos os segmentos textuais que contiverem todo ou parte do requisito de procura ou consulta do usuário serão levados em conta para uma seleção mais cuidadosa.

Por fim, a criação de um sumário envolve outro fator relevante: seu tamanho. A quantidade de informação que se deseja ter em um sumário deve ser limitada para que este possa cumprir seu propósito. Logo, o usuário deve poder delimitar certo valor, chamando taxa de compressão, a qual determina o tamanho do sumário final em relação aos textos de origem, em número de palavras, normalmente. Por exemplo, um sumário com 70% de compressão informa que o sumário deve conter, no máximo, 30% do número de palavras de um dos textos.

Abaixo, na Figura 4, exibe-se uma arquitetura genérica de sumarização proposta por Mani e Maybury (1999).

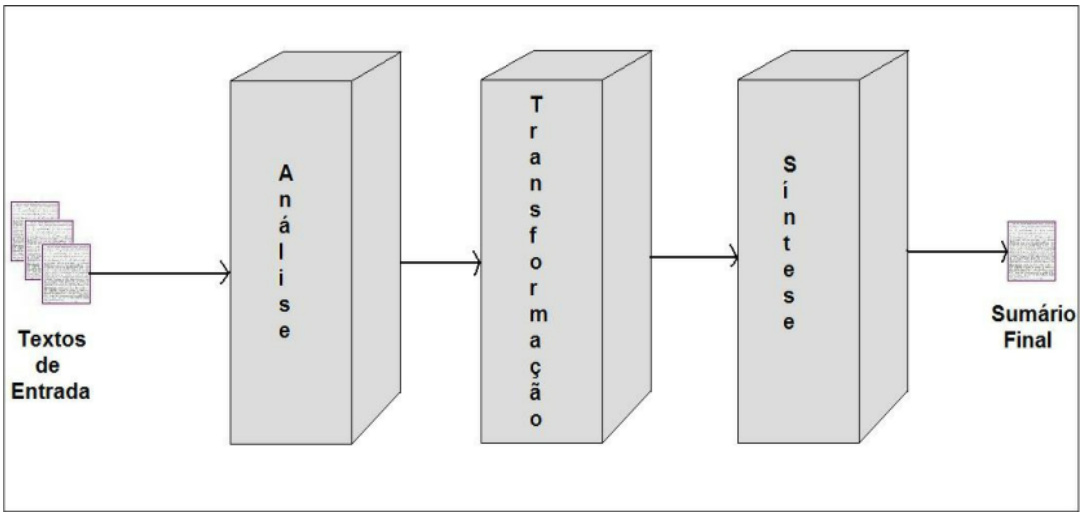

Figura 4 – Arquitetura genérica de sumarização

A primeira fase corresponde à análise dos textos de entrada, na qual estes são modelados de acordo com a necessidade e o método utilizado. Na segunda fase, ocorre principalmente a seleção do conteúdo relevante, onde os segmentos textuais são ranqueados de acordo com alguma métrica ou método. Por fim, a fase de síntese corresponde, em resumo, à expressão do conteúdo condensado em língua natural, podendo envolver etapas de junção, ordenação e representação deste conteúdo.

Na literatura, dispõem-se de diversos métodos que são associados às diferentes abordagens citadas anteriormente e que requerem maior ou menor grau de investimento em cada uma das etapas da arquitetura genérica de sumarização. Na abordagem superficial, há métodos como, por exemplo, método das Palavras-Chave e método da Localização. O primeiro (Palavras-Chave) parte do pressuposto de que as ideias principais de um texto podem ser expressas por algumas palavras-chave. Segundo Black e Johnson (1988), conforme as ideias vão sendo desenvolvidas no texto, os termos-chave aparecem com maior frequência. A ideia é, então, determinar a distribuição estatística das palavras-chave do texto e, a partir de sua frequência, extrair as sentenças que as contenham, agrupando-as de forma a constituir um sumário, na ordem em que aparecem originalmente. Já para o segundo método (Localização), Baxendale (1958) verificou que a posição de uma sentença em um texto poderia ser associada a sua importância no contexto textual. Por exemplo, a primeira e a última sentença de um texto jornalístico podem conter suas ideias principais e, portanto, estas seriam as sentenças consideradas para a produção de um sumário. Além dos métodos acima descritos, há também o que parte do descobrimento da ideia principal do texto, e é amplamente explorado pelo GistSumm (Pardo et al., 2003; Pardo, 2005). Além deste, outros métodos, dentro do cenário de sumarização, podem ser utilizados de forma a conseguir encontrar os segmentos textuais mais salientes de determinado conjunto de textos, como os métodos de Salton et al. (1997), Mani e Bloedorn (1997) e Mihalcea e Tarau (2005), dos quais o de Salton et al. é o que tem sido trabalhado, já que consiste em um método bem conhecido e difundido na literatura.

O descobrimento de segmentos textuais mais salientes também pode ser feito por meio da utilização de grafos. Grafos têm sido muito utilizados recentemente por serem estruturas bem entendidas e estudadas, com diversas métricas interessantes e, além de tudo,

resultarem em abordagens elegantes e claras para diversos problemas. Esse é o caso do trabalho de Salton et al. (1997). Esses autores modelam seu texto (originalmente no cenário monodocumento) em um grafo, sendo que os vértices representam segmentos textuais e a conexão entre eles (arestas) é feita em função da similaridade de palavras existentes (ligações intratextuais).

Segue abaixo, na Figura 5, um exemplo de modelagem de textos como grafos: cada vértice é uma sentença (determinada por S) de um documento (determinado por D) e as arestas são valores numéricos que indicam o quão próximas duas sentenças são lexicalmente. Tais valores podem ser calculados de diferentes formas, por exemplo, similaridade de cosseno (quanto menor o ângulo, mais similares são as sentenças) (Salton, 1988).

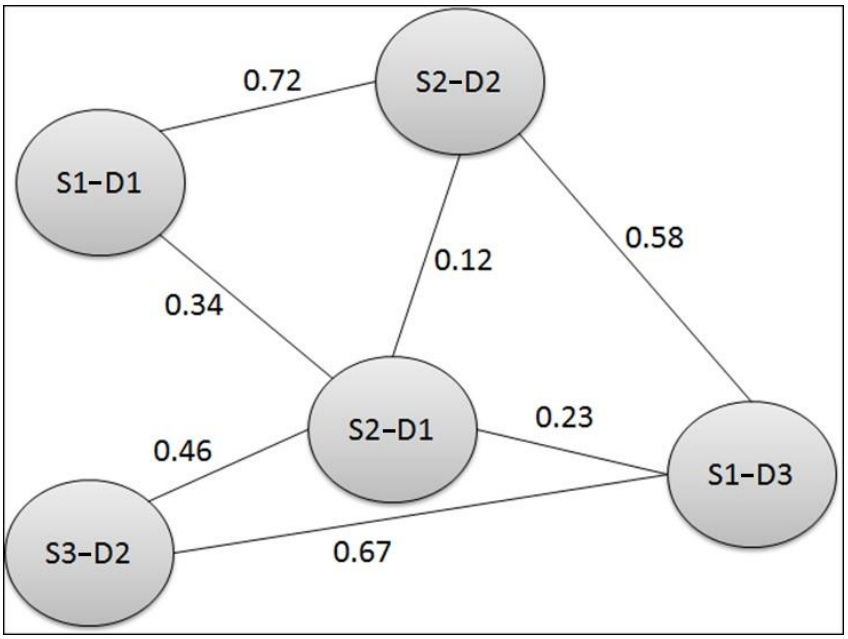

Figura 5 – Exemplo de grafo

Em relação aos métodos de Salton et al. (1997), os quais são a base para este trabalho, tem-se que, após então a transformação do texto em um grafo não direcionado (correspondente ao que Salton et al. chamam de mapa de relacionamentos), os algoritmos de seleção de conteúdo relevante podem então ser aplicados. São eles: Caminho Denso (no original, *bushy path*), Caminho Profundo (*depth-first path*) e Caminho Denso Segmentado (*segmented bushy path*). No Caminho Denso, a densidade de um nó (vértice) é definida

como o número de conexões que este tem com o resto do grafo; logo, um nó altamente conectado tem uma sobreposição grande de vocabulário com diversos parágrafos e, por este fato, tal nó é desejável em um sumário, pois se entende que o assunto principal de um texto será descrito ao longo dele. Portanto, um *bushy path* é construído com os nós mais densos, os quais são ordenados cronologicamente, isto é, na ordem em que originalmente aparecem no documento em questão, para então formar o sumário. Entretanto, os nós do Caminho Denso são conectados a outros parágrafos, mas não necessariamente entre si; por esta razão, o Caminho Profundo pode ser uma resolução para a má legibilidade possivelmente formada pelo caminho apresentado anteriormente. Em vez de selecionar os nós mais relacionados, começa-se por algum vértice importante (de preferência o de maior densidade) e, a partir dele, há a escolha de seu filho que tem mais ligações. Ainda que este algoritmo resolva a possível falta de coerência, ele causa outro problema que corresponde ao caminho não cobrir todos os subtópicos provenientes do texto de origem, sendo necessária a utilização do caminho denso segmentado. O Caminho Denso Segmentado ataca o problema mencionado anteriormente de forma que são construídos diversos caminhos densos e para cada subtópico e, logo depois, concatena-os em ordem textual. Garante-se não somente que pelo menos um parágrafo de cada subtópico é selecionado para compor o sumário, mas também algum parágrafo de transição entre subtópicos também seja incluído (respeitando a taxa de compressão), de forma que o sumário tenha uma melhor legibilidade.

Conforme citado na seção anterior, Ribaldo et al. (2012) exploraram e adaptaram os caminhos denso e profundo para o cenário multidocumento, produzindo bons resultados, mas não tratando a questão dos subtópicos. Assim, foi proposta aqui a continuação do trabalho iniciado previamente, considerando agora o terceiro caminho de Salton et al.: o Caminho Denso Segmentado. Este caminho, como antes dito, implica na construção de pequenos sumários para cada subtópico para, mais tarde, uni-los de forma coerente, os quais são devidamente demarcados por técnicas que realizam a segmentação topical de textos, como se descreve na próxima seção. É importante notar que, neste projeto, foram produzidos sumários genéricos e informativos, voltados para uma ampla audiência. Em particular, foram construídos extratos, visto que a proposta de Salton et al. (1997) é extrativa, em princípio.

2.3. Segmentação Topical

Vários sistemas de sumarização fazem uso de segmentadores topicais. Tem-se, como exemplos de segmentadores topicais, *C99* (Choi, 2000), o método implementado para o sumarizador *SUMMARIST* (Lin e Hovy, 2000) e *TextTiling* (Hearst, 1997).

O primeiro sistema (*C99*) descreve um método para a segmentação linear de textos, na qual é construída uma matriz de similaridade representando as conexões entre sentenças de forma a delimitar os subtópicos.

Em relação à Lin e Hovy, são definidas assinaturas correspondentes a cada subtópico, isto é, são gerados sumários a partir das palavras que mais identificam um determinado subtópico. Por último, tem-se o sistema *TextTiling* (utilizado neste trabalho), originalmente desenvolvido por Hearst (1997), adaptado ao português inicialmente por Leite (2010) e, finalmente, aprimorado por Cardoso et al. (2013), o qual é uma técnica para subdivisão de um texto em múltiplas unidades textuais (parágrafos ou sentenças) que representam passagens ou subtópicos. Neste contexto, passagem se refere a qualquer segmento de texto isolado do restante do mesmo, enquanto que subtópico, previamente introduzido, significa uma parte do texto sobre algum assunto específico, em contraste com o termo tópico, o qual indica o tema de um determinado documento por completo. Inicialmente, Hearst (1997) utiliza esta abordagem sobre textos expositivos (cenário monodocumento), os quais são caracterizados, na sua maioria, como longas sequências de parágrafos explicativos com pouca delimitação estrutural. Logo, o objetivo foca no particionamento desses textos em segmentos subtopicais contínuos e não redundantes. Abaixo, explica-se mais detalhadamente este método, pois ele é o utilizado neste projeto.

O algoritmo *TextTiling*, um dos mais utilizados na área, pode ser dividido em 3 partes principais: 1) divisão dos parágrafos em unidades sentenciais e radicalização de suas palavras; 2) associação de uma nota lexical para cada par de unidade sentencial e 3) detecção de limites subtopicais. Na primeira parte, a radicalização se refere à redução de entradas textuais (palavras) em unidades lexicais menores (radicais). É importante notar que, neste projeto, é utilizado o radicalizador *Snowball* (Porter, 2001) por mostrar grande eficiência e facilidade de uso. Completada essa fase de uniformização de palavras, o texto

é subdividido em pseudo-sentenças (*token-sequences*) de um tamanho pré-definido *w* (parâmetro do algoritmo). Isso é feito, em contraste com o uso de sentenças reais, por permitir que comparações possam ser realizadas entre unidades de tamanhos iguais, uma vez que comparar duas sentenças, uma muito grande e outra muito pequena, produziriam notas incomparáveis. Outro ponto relevante a ser considerado são as *stopwords*; estas contribuem para a computação do tamanho das *token-sequences*, mas são retiradas a partir deste passo, uma vez que são palavras muito comuns e não carregam significado importante para serem consideradas no processamento.

Em relação à segunda parte do algoritmo (associação de uma nota lexical para cada par de unidade sentencial), têm-se três estratégias que podem ser utilizadas. São elas: comparação de blocos adjacentes de textos (*comparing adjacent blocks of text*), mudanças de vocabulário (*vocabulary introductions*) e cadeias lexicais (*lexical chains*). No caso de comparação de blocos adjacente de textos, o qual é considerado neste trabalho, a nota é dada pela soma dos produtos das frequências com que as palavras aparecem nos blocos sendo comparados, os quais, neste contexto, são formados por conjuntos de pseudo-sentenças. Abaixo, tem-se a Figura 6, com um exemplo hipotético, para uma melhor exemplificação. Neste exemplo, assumimos que cada letra corresponde a uma palavra, sendo que o bloco formado por um conjunto de letras representa uma *token-sequence*. Desta forma, para os dois primeiros blocos da figura, a nota é calculada como sendo 2*1 (para A) + 1*1 (para B) + 2*1 (para C) + 1*1 (para D) + 1*2 (para E).

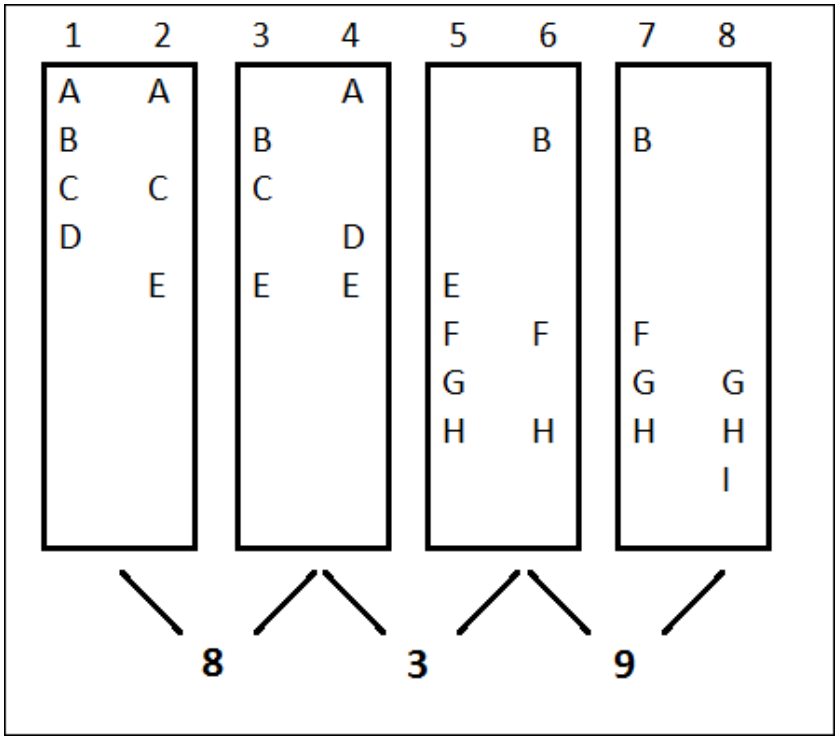

Figura 6 – Comparação de blocos

Desta forma, é atribuída uma nota de profundidade (*depth score*) para a identificação de vales para cada lacuna entre pseudo-sentenças. Assim, quanto maior a profundidade existente entre as pseudo-sentenças, maior é a indicação de que há uma mudança de subtópico. No exemplo acima, tem-se a nota 3 (considerada baixa) entre duas notas relativamente altas. Logo, é definida então uma mudança de subtópico entre a *token-sequence* 4 e 5. Um problema em potencial com tal método pode acontecer se, por um acaso, houver pequenos vales que "interrompem" a medição. Nestes casos, aplica-se uma técnica chamada de suavização, a qual ajuda a eliminar tais perturbações levando em conta o tamanho de ambos os lados do vale.

A escolha da técnica *TextTiling* se deve ao fato de, além de ser uma técnica bastante utilizada na área, já existir ferramentas para o português (Leite, 2010; Cardoso et al., 2013), como antes mencionado, que atendem os propósitos deste projeto. É importante notar que foi utilizada a ferramenta de Cardoso et al. neste projeto pelo simples fato da mesma ser uma versão aprimorada e mais atual da adaptação original.

Por fim, a utilização do *TextTiling* para o cenário multidocumento tem de ser feita considerando um passo adicional de correlação das passagens de subtópicos semelhantes, uma vez que tal técnica pertence ao cenário monodocumento. Desta forma, a identificação de subtópicos semelhantes entre documentos de mesmo assunto é feita considerando o agrupamento entre os primeiros, melhor descrito na Seção 2.4. Além disso, o uso desta técnica, juntamente com o terceiro caminho de Salton et al. (Caminho Denso Segmentado), deve ser incorporada após a fase de pré-processamento das sentenças, para que haja o descobrimento de subtópicos e, por fim, a seleção de conteúdo relevante pelo método de Salton et al. (1997).

2.4. Agrupamento

Agrupamento (*Clustering*) é uma noção que surge naturalmente em muitos campos, sempre que se tem um conjunto heterogêneo de objetos. É natural a busca por métodos para agrupar tais objetos com base em alguma medida de semelhança. Por exemplo, para definir a distância entre objetos, pode-se considerar que quanto mais perto eles estão entre si, mais parecidos eles são. Assim, o agrupamento é centrado em torno de um intuitivo, mas vago, objetivo: dado um conjunto de objetos, particioná-los em uma coleção de aglomerados em que os objetos no mesmo grupo estão próximos, enquanto os objetos em diferentes grupos estão distantes.

Há diversos trabalhos na área, como a ferramenta para a língua portuguesa SiSPI (Seno, 2008), a qual aplica o agrupamento com base do algoritmo *Single-pass* (Van Rijsbergen, 1979) (considerado neste projeto) para o agrupamento de sentenças similares. Tal como o nome sugere, *Single-pass*, no contexto do descobrimento de subtópicos correlacionados, requer uma única passagem sequencial ao longo do conjunto de subtópicos a serem agrupados. É um algoritmo de agrupamento incremental (grupos são criados de forma incremental analisando todos os outros previamente criados). O algoritmo do método *Single-pass* já adaptado ao caso de subtópicos é mostrado na Figura 7. É importante notar que a medida de similaridade utilizada, neste caso, é a similaridade de cosseno, a qual é utilizada neste projeto.

Entrada: Conjunto D = $<d_1,...,d_n>$ com *n* documentos, onde cada $d_i = <s_1,...,s_m>$ com *m* subtópicos para cada *n* e *m* >= 1.
Saída: Conjunto C = $<c_1,...,c_x>$ com *x clusters*, onde cada $c_i = <s_1,...,s_y>$ com *y* subtópicos para *y* >= 1.

Passo 1: Defina um conjunto inicial de *clusters* C vazio
Passo 2: Selecione um subtópico s_i de um documento d_i seguindo uma determinada ordem
Se C está vazio
Então adicione o primeiro *cluster* a C inserindo o elemento s_i
Senão compare s_i (tratado como um novo *cluster* com um único elemento) com todos os *clusters* em C
Se a similaridade entre s_i e quaisquer outros *clusters* em C está acima de um limiar pré-definido
Então agrupe s_i com o *cluster* em C mais similar
Senão adicione um novo *cluster* a C

Passo 3: Repita o passo acima até que todos os subtópicos de todos os documentos sejam processados.

Figura 7 – Algoritmo *Single-pass*

Inicialmente, o algoritmo cria o primeiro grupo, selecionando o primeiro subtópico de uma coleção de documentos a ser agrupado. Então, este primeiro grupo inicia o trabalho de agrupamento com todos os subtópicos remanescentes. Em cada análise dos grupos, o algoritmo decide se um subtópico recém-selecionado deve ser colocado em um grupo já criado ou em um novo. Esta decisão é feita de acordo com a condição especificada pela função de similaridade empregada, isto é, um limite de similaridade previamente determinado. Neste trabalho, a função de similaridade é a medida de similaridade lexical do cosseno, explicada anteriormente – quanto maior o valor de similaridade entre dois subtópicos, mais semelhantes eles são. A decisão do limiar utilizado baseou-se no cálculo da média de similaridade entre todos os grupos. É importante salientar que há diversos meios a serem considerados para a realização do agrupamento, os quais são abordados mais abaixo.

Por fim, a tarefa de agrupamento é necessária, uma vez que a segmentação topical aplicada pela técnica *TextTiling* é feita de forma linear, isto é, sobre somente um documento, acarretando no seguinte problema: se dois subtópicos de diferentes textos forem encontrados, não há a garantia de que, entre eles, tais subtópicos correspondam à informações centrais distintas. Desta forma, com a aplicação de um agrupamento, a identificação de tais subtópicos pode ser descoberta.

2.5. *Cross-document Structure Theory*

A *Cross-document Structure Theory* (CST) (Radev, 2000) é uma teoria usada para descrever conexões semânticas entre unidades topicamente relacionadas, como a relação de equivalência (correspondente à paráfrase) exemplificada abaixo.

> (1) A polícia também vai abrir nova investigação sobre a participação de desembargadores e conselheiros do Tribunal de Contas no suposto esquema.
>
> (2) A PF abriu uma nova frente de atuação para apurar o caso com o objetivo de apurar a participação de desembargadores e conselheiros do Tribunal de Contas na quadrilha.

Além da relação demonstrada acima, a CST propõe um conjunto de mais 23 (vinte e três) relações discursivas para relacionamento multidocumento. Na Tabela 1, listam-se todas estas.

Tabela 1. Conjunto original de relações da CST

Identity	*Modality*	*Judgment*
Equivalence	*Attribution*	*Fulfillment*
Translation	*Summary*	*Description*
Subsumption	*Follow-up*	*Reader profile*
Contradiction	*Elaboration*	*Contrast*
Historical background	*Indirect speech*	*Parallel*
Cross-reference	*Refinement*	*Generalization*
Citation	*Agreement*	*Change of perspective*

Ainda que existam 24 relações CST, neste trabalho foram utilizadas somente 14, pelo fato do número limitado de relações encontradas e anotadas no conjunto de textos aqui considerado (mencionado na Seção 3.3). Tais relações, definidas por Maziero et al. (2010), podem ser encontradas na Figura 8.

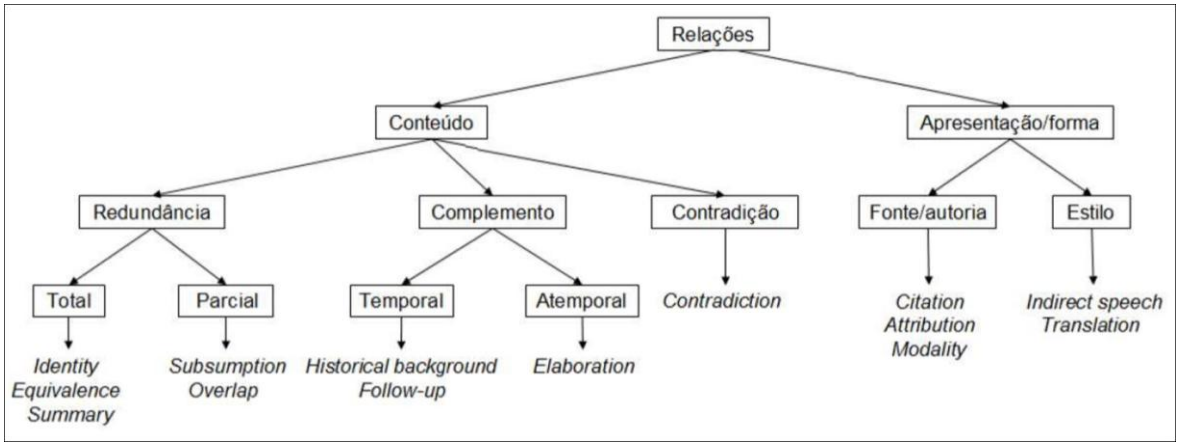

Figura 8 – Árvore das relações CST

A tipologia acima classifica as relações CST em dois grandes grupos: conteúdo e apresentação/forma. Estes ainda são subdivididos, como pode ser observado. O primeiro indica as relações de conteúdo entre segmentos. Já o segundo grupo se preocupa com a apresentação e forma com a qual o conteúdo foi expresso. Foi constatado que quando há a ocorrência de fonte/autoria ou estilo, alguma categoria do grupo conteúdo também ocorre.

Com base no significado das relações CST definidas por Maziero et al., estas podem auxiliar na fase de agrupamento, de forma a obter uma melhor seleção de subtópicos semelhantes. Isto se deve ao fato de que foi constatado no trabalho anterior que as relações semântico-discursivas estão intimamente relacionadas com a similaridade lexical de segmentos textuais, isto é, quanto mais próximos tais segmentos forem, mais relações CST os mesmos terão entre si e, portanto, maior a chance dos segmentos pertencerem a um mesmo subtópico. Logo, esse modelo foi utilizado para investigar a questão do relacionamento entre subtópicos.

CAPÍTULO 3: Atividades no período

3.1. Considerações Iniciais

Estudos foram realizados para um melhor entendimento da área de sumarização e processamento de linguagem natural, na qual este projeto se encaixa. A seleção de trabalhos a serem estudados foi feita de acordo com a proposta de pesquisa, seguindo cautelosamente o foco de cada tarefa. O Capítulo 2, já apresentado, sintetizou os principais trabalhos estudados, a saber, trabalhos de sumarização em si, de delimitação topical e de agrupamento. Tal revisão fundamenta a modelagem proposta na Seção 3.4.

3.2. Familiarização com recursos e ferramentas necessárias ao projeto

Para o português, há alguns recursos e ferramentas de base para a sumarização. Em termos de córpus de textos e sumários, há dois recursos importantes e amplamente utilizados na área: o TeMário (Pardo e Rino, 2003), para sumarização monodocumento, e o CSTNews (Cardoso et al., 2011), para sumarização mono e multidocumento. O CSTNews foi e tem sido a base para os trabalhos recentes em sumarização multidocumento no Brasil, incluindo o trabalho de Ribaldo et al. (2012), que corresponde à primeira iniciação científica do aluno. Logo, neste projeto, foi utilizado tal córpus, que conta com 50 grupos de textos jornalísticos, sendo que cada grupo tem 2 ou 3 textos sobre um mesmo assunto e é acompanhado por sumários humanos, tanto mono quanto multidocumento. Tais textos jornalísticos, com as devidas anotações manuais de subtópicos e de acordo com a CST, forneceram os dados de referência necessários para o desenvolvimento e avaliação deste trabalho.

Os grupos de textos contêm, cada um, um sumário feito manualmente, como mencionado acima e, ainda, sumários extrativos manuais produzidos no trabalho anterior. Tais extrações seguiram um mesmo esquema de formação iniciando-se com a leitura minuciosa dos textos contidos em cada grupo do córpus e dos resumos manuais para que o entendimento do assunto fosse claro. Após isso, cada sentença do sumário foi analisada e

comparada com as dos textos-fonte para que houvesse a certificação de que a sentença extraída (texto-fonte) continha todo o conteúdo da sentença pertencente ao sumário manual. Eventualmente, uma sentença extraída poderia não abranger todas as informações que eram requisitadas, para isso duas ou mais sentenças eram selecionadas para compor o sumário extrativo.

Além disso, sentenças de textos diferentes no conjunto corrente, que continham todo o conteúdo necessário para o seu envio ao sumário, foram tratadas de forma que a menor delas era escolhida para que o tamanho do sumário permanecesse, com alguma pequena variação, igual ao manual. É notável que alguns desses extratos acabaram ultrapassando a quantidade de sentenças esperadas, mas o conteúdo foi severamente verificado e constatado que a diminuição destas acarretariam em uma significante perda de informação.

Abaixo, nas figuras 9, 10 e 11, temos as primeiras sentenças dos textos pertencentes ao primeiro grupo do córpus e, em seguida, na Figura 12, o sumário manual correspondente.

Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.

Figura 9. Sentença 1 – Texto 1

Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira, informou nesta sexta-feira um porta-voz das Nações Unidas.

Figura 10. Sentença 1 – Texto 2

Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo, matou 17 pessoas na quinta-feira à tarde, informou hoje um porta-voz das Nações Unidas.

Figura 11. Sentença 1 – Texto 3

17 pessoas morreram após a queda de um avião na República Democrática do Congo.

Figura 12. Sumário manual

Percebe-se que a sentença do Texto 1 (Figura 9) contém todas as informações para o claro entendimento do Sumário Manual, logo será selecionada para a composição do sumário extrativo. Outro ponto importante é o descobrimento futuro da origem da sentença escolhida, que foi tratada de forma que foi adicionada uma *tag* que contém o necessário para sua identificação, como número do grupo, número do texto do conjunto, parágrafo e posição da sentença dentro deste.

Para uma melhor análise da variação de tamanho entre os sumários extrativos e os manuais, seguem, na Tabela 2, seus dados representantes.

Tabela 2. Dados referentes aos córpus CSTNews

Nº de Textos	**Nº de palavras**	**Nº de Sentenças**	**Tamanho sumário manual (nº de palavras)**	**Tamanho sumário extrativo (nº de palavras)**	**Aumento (%)**
3	940	41	137	190	41.6

Assim, temos dois tipos de sumários que puderam ser, então, analisados automaticamente após seus textos-fonte serem sumarizados. Nota-se que há um considerável aumento no tamanho do sumário extrativo em relação ao manual, mas este crescimento deve-se ao fato de que, algumas vezes, a informação de uma sentença do sumário manual estava contida em duas ou mais sentenças dos textos-fonte, fazendo com que o sumário extrativo aumentasse de tamanho. Em adição, o córpus serve tanto para desenvolvimento de métodos de sumarização, pois permite aprendizado de estratégias de sumarização e a customização delas, quanto para avaliação dos sistemas desenvolvidos. Em geral, os sumários automáticos produzidos pelos sistemas são comparados com os sumários humanos correspondentes via métricas tradicionais, como a ROUGE (Lin e Hovy, 2003; Lin, 2004), introduzida posteriormente na Seção 3.5.

3.3. Descrição das Atividades Realizadas

Nas seções seguintes, cada uma das atividades realizadas para o desenvolvimento deste trabalho de conclusão de curso é descrita, de forma que primeiramente é apresentado o passo a passo do projeto, seguido de exemplos e explicações mais detalhadas. Ao final, discute-se e avalia-se os resultados obtidos.

3.4. Modelagem de textos como grafos e de algoritmos para seleção da informação relevante

Como antes mencionado, foi proposto o percurso no grafo a partir do modelo de Salton et al. (1997). É importante salientar que (1) foi considerado somente o caminho denso segmentado neste trabalho, devido a conclusão dos dois outros no projeto anterior; (2) o método escolhido, originalmente, foi desenvolvido para o cenário monodocumento; logo, adaptações foram realizadas para que este conseguisse identificar informações relevantes em, não somente um, mas diversos documentos. Segue abaixo, na Figura 13, um esquema a ser seguido para a realização deste projeto.

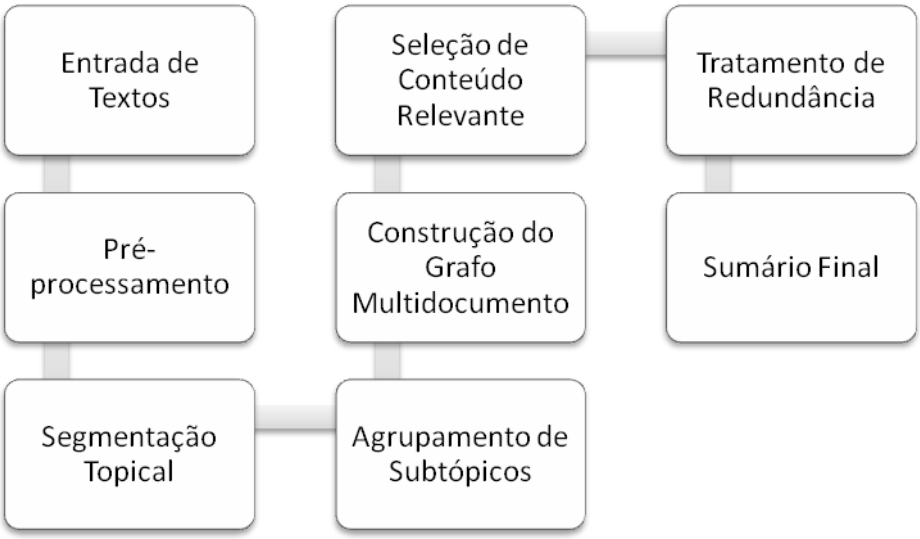

Figura 13 – Passo a passo do projeto

Primeiramente, o método não foi mais modelado sobre um texto somente, mas sim sobre um conjunto finito de textos, fazendo com que as sentenças pudessem se relacionar com

outras que eram fornecidas por um documento diferente do primeiro. Assim que os textos são escolhidos pelo usuário, estes são modelados como um grafo, onde seus vértices são sentenças e as ligações são valores numéricos que indicam o quão próximas elas são lexicalmente (como mencionado anteriormente, utilizando a Similaridade de Cosseno). É válida a explicação de que muitas vezes nos deparamos com palavras similares, mas que estão em gêneros diferentes, por exemplo: casa e casas. Logo, é necessário um tratamento para tais tipos de palavras, o qual pode ser feito pela lematização ou radicalização (Porter, 2001). Este tratamento tem por finalidade a uniformização das palavras das sentenças.

É importante notar que o tratamento indicado acima elimina as chamadas *stopwords*, que são palavras muito comuns. Por este fato, sua remoção é essencial para que somente palavras que carregam significado possam ser consideradas no processamento. O descobrimento das *stopwords* foi apoiado em uma lista pré-definida para a língua portuguesa.

Abaixo, tem-se um exemplo de todo o processo (tratamento) realizado em cada sentença de todos os textos-fonte envolvidos. Para uma melhor compreensão, são apresentadas sentenças hipotéticas, na Figura 14, que contêm as informações necessárias para esta exemplificação.

Considerando duas sentenças de diferentes textos que versam sobre um mesmo assunto, temos:

Sentença 1 (S1): Todas as casas foram destruídas.
Sentença 2 (S2): A destruição não poupou nenhuma casa.

1º Passo ➔ Remoção de *stopwords*

S1: casas foram destruídas
S2: destruição não poupou casa

2º Passo ➔ Radicalização das palavras (*Stemmer* de Porter, 2001)

S1: cas ser destríd
S2: destruiç não poup cas

3º Passo ➔ Aplicação da medida do cosseno

3.1º Passo ➔ Junção de todas as palavras distintas

cas ser destruíd destruiç não poup

3.2º Passo ➔ Atribuição de valores a cada palavra de acordo com o número de aparições da mesma em cada sentença

S1: 1 1 1 0 0 0
S2: 1 0 0 1 1 1

3.3º Passo ➔ Finalmente, aplicação direta da fórmula da medida do cosseno

$$\text{cosseno}(S1,S2) = \frac{S1.S2}{|S1|*|S2|} = 0.29$$

Figura 14 – Exemplo de pré-processamento

Completada a fase de pré-processamento, a técnica *TextTiling* começa a ser aplicada. Nota-se que esta técnica é aplicada para cada documento de entrada, como segue na Figura 15.

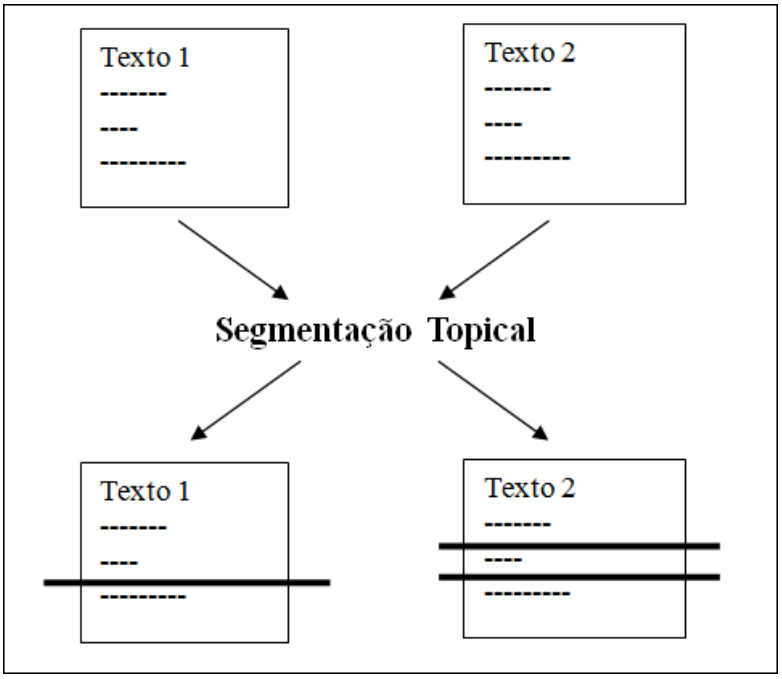

Figura 15 – Exemplo de segmentação Topical

Como pode ser visto na figura acima, após a Segmentação Topical realizada pelo *TextTiling*, os subtópicos dos textos são descobertos (divisão caracterizada pela linhas mais escuras). Como uma melhor exemplificação, temos as figuras abaixo que compreendem dois documentos sobre as Olimpíadas de Pequim.

<t1>
A ginasta Jade Barbosa, que obteve três medalhas nos Jogos Pan-Americanos do Rio, em julho, venceu votação na internet e será a representante brasileira no revezamento da tocha olímpica para Pequim-2008.
</t1>
<t2>
A tocha passará por vinte países, mas o Brasil não estará no percurso olímpico.
Por isso, Jade participará do evento em Buenos Aires, na Argentina, única cidade da América do Sul a receber o símbolo dos Jogos.
</t2>
<t3>
O revezamento terminará em 8 de agosto, primeiro dia das Olimpíadas de Pequim.
</t3>

Figura 16 – Documento 1

<t4>
Um dos destaques desta temporada do esporte brasileiro, a ginasta Jade Barbosa foi escolhida, na noite desta terça-feira, para ser a representante do Brasil no revezamento da tocha dos Jogos Olímpicos de Pequim.
Em votação pela internet, a ginasta recebeu mais de 100 mil votos e superou o nadador Thiago Pereira, que ganhou seis ouros nos Jogos Pan-Americanos.
</t4>
<t5>
O Brasil não faz parte do trajeto da tocha olímpica.
Na América do Sul, a chama passará por Buenos Aires, onde Jade participará do revezamento, no dia 11 de abril.
</t5>
<t6>
Aos 16 anos, Jade conquistou três medalhas no Pan: ouro na disputa dos saltos, prata na apresentação por equipes e bronze no solo.
</t6>
<t7>
Ao todo, a chama olímpica percorrerá 20 países antes de chegar a Pequim para a abertura da competição, no dia 8 de agosto.
</t7>

Figura 17 – Documento 2

As figuras 16 e 17 mostram claramente, através das *tags* <t>, a segmentação topical aplicada pela técnica *TextTiling*. Como essa técnica foi desenvolvida para o cenário monodocumento, como mencionado anteriormente, não há correlação dos subtópicos. Desta forma, precisa-se então aplicar a técnica de agrupamento de subtópicos, como ilustra a Figura 18.

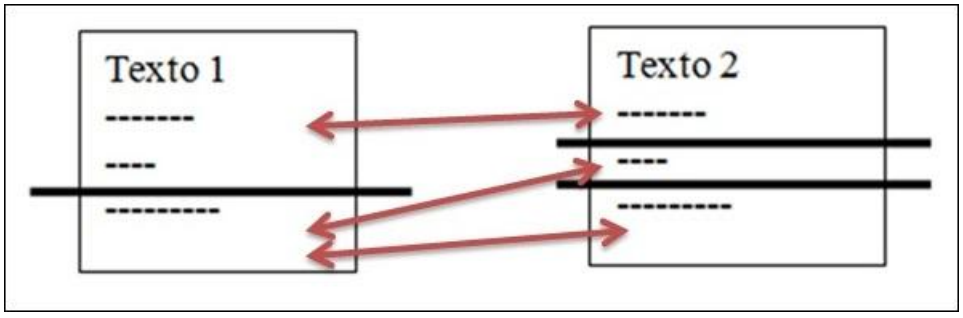

Figura 18 – Correlação de subtópicos

Nesta fase de correlação de subtópicos, ou melhor, de agrupamento, foram considerados 4 modos para encontrar subtópicos semelhantes: 1) Palavras-Chave – as palavras mais frequentes são descobertas; desta forma, a similaridade aplicada entre os grupos é feita analisando-se somente tais palavras; 2) Similaridade Subtopical – todas as palavras dos subtópicos são consideradas para a análise de proximidade lexical, ou melhor, são analisados pares de subtópicos para determinar se os mesmos compõem o mesmo grupo; 3) CST sem Peso - utilização do número de relações CST entre subtópicos para investigar sua correlação, isto é, quanto maior o número de conexões entre dois subtópicos, maior a chance de serem correlacionados; e 4) CST com Peso – utilização do valor numérico de cada relação CST entre subtópicos. Tais valores numéricos, no intervalo de 0 a 1, correspondem ao nível de similaridade entre cada par de subtópico, sendo que uma relação de Identidade, por exemplo, corresponde ao valor 1 de similaridade. Assim, quanto maior for a soma de cada relação CST entre um par de subtópicos, maior será a chance de agrupá-los. Abaixo, encontra-se, na Figura 19, a correlação dos subtópicos definida após aplicada a técnica de agrupamento (2) Similaridade Subtopical, como um exemplo.

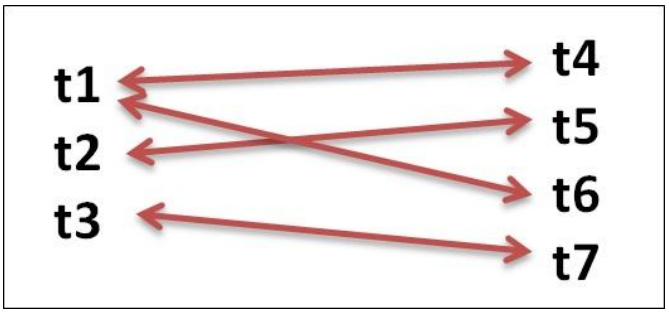

Figura 19 – Correlação de subtópicos

É possível notar que o subtópico 1 do primeiro documento é similar ao subtópicos 4 e 6 do documento 2. Portanto, estes três subtópicos serão agrupados em um único grupo. O mesmo acontece, de forma análoga, para os subtópicos remanescentes.

Feita tal modelagem, as sentenças começam a ser selecionadas de acordo com o que foi descrito na Seção 2.2, onde todo o trabalho de Salton et al. foi explicado. Ainda assim, para uma melhor visualização do método, temos a Figura 20 abaixo que retrata a modelagem dos documentos com a segmentação topical e o agrupamento já realizados.

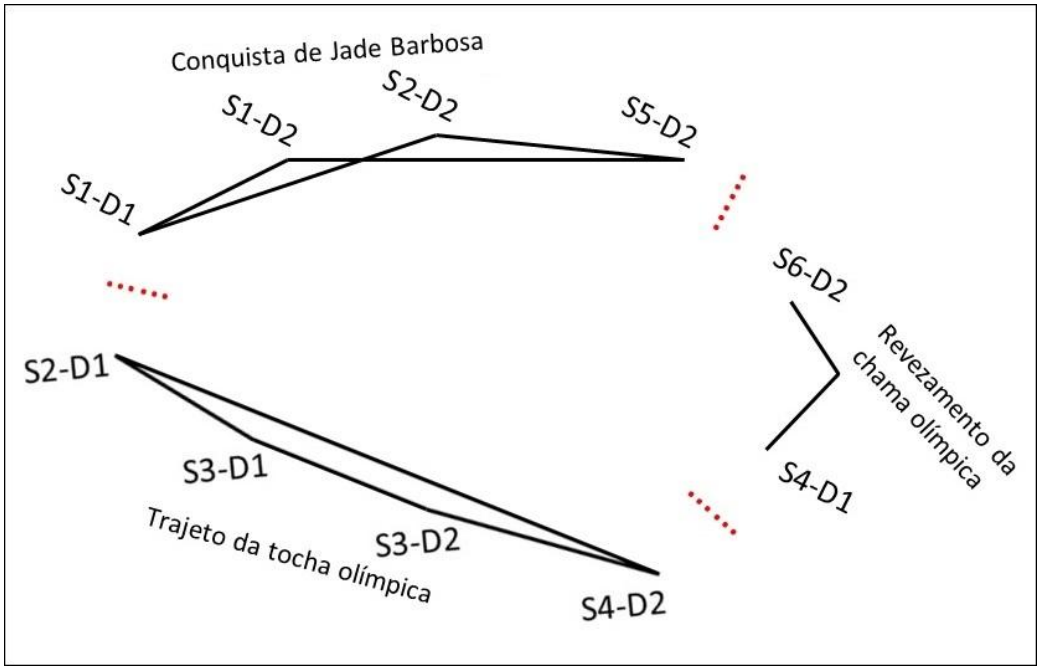

Figura 20 – Mapa de Relacionamentos

A figura acima mostra o mapa de relacionamento modelado para a aplicação do terceiro caminho proposto por Salton et al. (1997). Como antes mencionado, este caminho constrói o sumário final selecionando as sentenças mais importantes de cada subtópico; para isso, escolhe-se a sentença mais conectada dentro de um determinado subtópico. Feita a primeira escolha da sentença do primeiro conjunto (subtópico), é necessário selecionar uma sentença de transição antes mesmo da escolha da segunda sentença mais relevante de outro subtópico. É importante notar que esta sentença não é escolhida aleatoriamente, mas sim de forma que a mesma seja cronologicamente anterior à sentença principal subsequente para que a passagem de um subtópico para outro se dê de forma coerente. Por fim, após construir o sumário com as primeiras sentenças mais importantes (conectadas) de cada subtópico e ainda considerar a seleção de sentenças de transição quando há a mudança de subtópicos, tem-se a seleção das sentenças subsequentes através da mesma consideração: escolhe-se a sentença mais conectada que não tenha ainda sido levada ao sumário.

É notável que, utilizando tal método, muitas arestas eventualmente podem indicar um grau de similaridade altíssimo (devido à redundância entre textos), por isso deve-se calcular o limite de redundância que duas sentenças podem ter entre si de acordo com uma

média dos valores de cosseno no sumário para descartar sentenças redundantes, ou seja, caso um nó (sentença) tenha um valor de similaridade maior (em relação aos nós que já compõem o sumário) do que o calculado como limitante, este não é levado ao sumário, pois é considerado redundante. O cálculo desse limitante se dá através da média dos valores de similaridade entre cada uma das sentença de todos os documentos.

Por fim, a construção do sumário final deve levar em conta o seu tamanho e, para isso, utiliza-se uma determinada taxa de compressão (mencionada anteriormente), a qual limita a quantidade de informações que o sumário irá conter. Nota-se que, justamente pela aplicação da taxa de compressão, sentenças relevantes podem ficar de fora do sumário em prol da seleção de sentenças de transição. Optou-se por manter tal abordagem para que, mais uma vez, a coerência fosse mantida.

3.5. Desenvolvimento de um protótipo computacional com base na modelagem da proposta

À medida que a necessidade de uma ferramenta que lide com o grande cenário multidocumento atual cresce, neste trabalho, foi desenvolvida uma extensão para o navegador *Google Chrome*, a qual sumariza os documentos retornados de uma pesquisa através do site *Google News.* Foi cogitada tal implementação para uma melhor aplicação das técnicas propostas, bem como a utilidade da ferramenta. A técnica de sumarização aplicada utiliza as seguintes ferramentas: 1) Interface de Programação de Aplicativos do *Google News* (*Google News API*) para a recuperação dos documentos; 2) *NCleaner* (Evert, 2008) – remoção de conteúdo não relevante de uma página da *web* (propagandas e *links* para outras páginas, por exemplo); 3) os métodos descritos por Ribaldo et al. (2012) – para a sumarização dos documentos aplicando os métodos de Salton et al. É importante notar que a incorporação do método estudado atualmente está em andamento, pois o mesmo produziu, como apresentado na seção 3.5, resultados ainda melhores que o anterior.

A Figura 21 mostra a extensão ativa com a pesquisa pelo termo "Manifestações São Paulo" no site *Google News*, seguida da Figura 22, onde o retorno da pesquisa foi sumarizado (neste caso, foram considerados os oito primeiros textos mais relevantes).

Figura 21 – Extensão ativa no navegador *Google Chrome*

Pode-se notar na figura acima que, a partir da pesquisa por um determinado termo no site *Google News*, o botão "Sumarizar" (canto superior direito) aparece, dando ao usuário a opção de ter os textos referentes à sua pesquisa sumarizados.

É importante notar que o sistema está atualmente customizado para a língua portuguesa, uma vez que o radicalizador e a *stoplist* utilizados são para tal língua. Entretanto, tal customização pode ser facilmente feita para a língua inglesa.

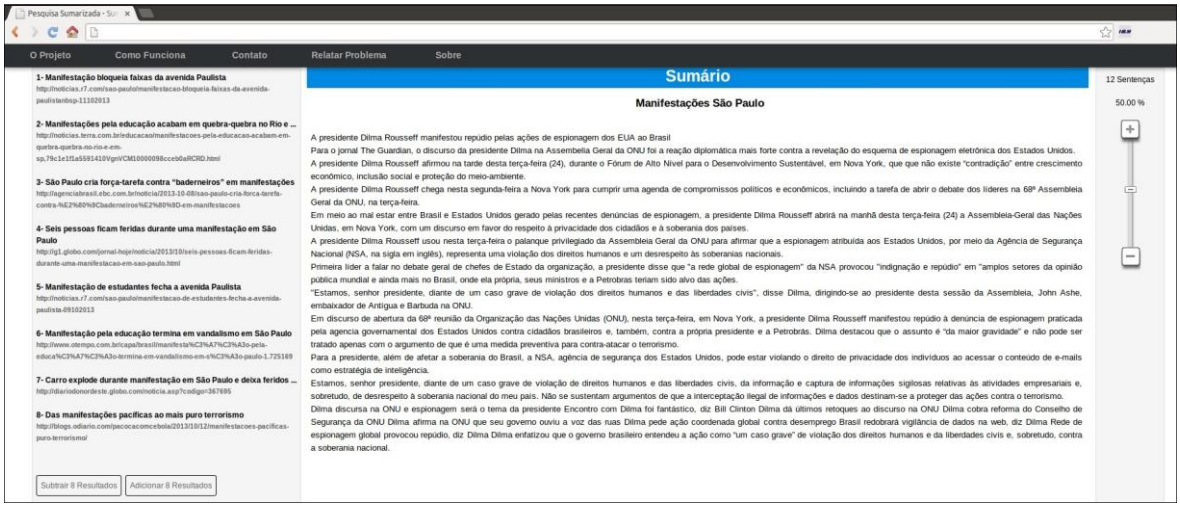

Figura 22 – Sumário gerado pela extensão

A figura acima apresenta um sumário com uma taxa de compressão de 70% sobre a pesquisa efetuada anteriormente. Pode-se notar que no lado esquerdo têm-se os *links* referentes à busca, além dos *links* em cada sentença do sumário que as conecta a seu texto de origem.

3.6. Avaliação

Para avaliar a qualidade dos sumários gerados, deve-se levar em conta não somente o sumário final, mas também os passos anteriores: a segmentação topical e o agrupamento.

A avaliação da técnica utilizada (*TextTiling*) realizada por Cardoso et al. (2013) levou-se em conta três medidas tradicionais: Precisão (*Precision*), Cobertura (*Recall*) e Medida-F (*F-Measure*). Segue abaixo a explicação sucinta de cada uma das medidas apresentadas acima no contexto da segmentação topical seguida Tabela 3, a qual contém os valores obtidos da avaliação:

- Precisão: representa a quantidade de subtópicos descobertos automaticamente que se encontram na segmentação topical manual;

- Cobertura: representa a quantidade de subtópicos segmentados manualmente que se encontram na segmentação automática;

- Medida-F: média harmônica da precisão e da cobertura. Essa medida é utilizada para medir a exatidão do classificador, ou seja, quanto mais próximo de 1, melhor a segmentação, e os resultados mais próximos de 0 demonstram que as segmentações topicais são ruins. A medida-F é útil por combinar em uma única medida tanto a Precisão quanto a Cobertura.

Tabela 3. Avaliação do *TextTiling*

Precisão	**Cobertura**	**Medida-F**
0.7732	0.4056	0.4978

A tabela acima mostra os valores obtidos pelo *TextTiling* quando avaliado sob o córpus CSTNews. Apesar desta técnica ser bastante conservativa e capaz de detectar somente alguns segmentos, obteve-se valores consideráveis quando comparados com outros métodos também avaliados por Cardoso et al. (2013).

Realizada a avaliação da técnica de segmentação topical, tem-se agora a avaliação dos resultados providos pelo agrupamento de subtópicos. A qualidade do método de agrupamento utilizado pode ser avaliada por medidas de qualidade externa. Essas medidas avaliam o quão próximos os grupos produzidos automaticamente estão em relação aos grupos de referência. Para tal avaliação, foram utilizadas as medidas de Precisão, Cobertura e Medida-F (Steinbah et al., 2000; Fung et al., 2003) descritas abaixo (Fórmula 1, Fórmula 2 e Fórmula 3).

(1)

$$P(k_i, c_j) = \frac{n_{ij}}{|c_j|}$$

(2)

$$C(k_i, c_j) = \frac{n_{ij}}{|k_i|}$$

(3)

$$F(k_i, c_j) = \frac{2 * C(k_i, c_j) * P(k_i, c_j)}{C(k_i, c_j) + P(k_i, c_j)}$$

Nas fórmulas acima, tem-se 1) k_i: indica cada grupo de referência (manual); 2) c_j: indica os grupos formados automaticamente; e 4) n_{ij}: número de segmentos da classe k_i que estão presentes no grupo c_j.

A Medida-F de cada classe de todo o conjunto de dados se baseia no grupo que melhor descreve cada classe k_i . Assim, o valor da Medida-F global, F(S), pode ser denotado pela fórmula abaixo.

(4)

$$F(S) = \sum_{k_i \in K} \frac{|k_i| \max c_j \in C\{F(k_i, c_j)\}}{N}$$

Na Fórmula 4, temos 1) S: solução de agrupamento; 2) N: número total de segmentos a serem agrupados; 3) K: conjunto de classes (grupos de referência); e 4) C: conjunto de grupos (grupos automáticos) .

Após utilizar os quatro métodos de agrupamentos descritos anteriormente (Palavras-Chave, Similaridade Subtopical, CST sem Peso e CST com Peso) e compará-los com os agrupamentos de referência realizados por humanos, seguindo-se as fórmulas 1, 2, 3 e 4, temos a seguinte tabela de resultados.

Tabela 4. Resultados de agrupamento (segmentação automática) – Medida-F Global

Palavras-Chave	**Similaridade Subtopical**	**CST sem Peso**	**CST com Peso**
0.6790	0.4856	0.4475	0.4519

Os resultados apresentados acima mostram valores bastante consideráveis quando os mesmos são analisados da forma: quanto mais próximo de 1, melhor. Conclui-se ainda que a melhor técnica utilizada, levando em consideração o método automático de segmentação topical (*TextTiling*), foi a consideração das palavras mais frequentes em cada subtópico para o agrupamento: Palavras-Chave.

Porém, um passo adicional teve de ser realizado para que as conclusões pudessem ser ainda mais concisas: a avaliação dos agrupamentos de referência/manual (Tabela 5).

Tabela 5. Resultados de agrupamento (segmentação manual) – Medida-F Global

Palavras-Chave	Similaridade Subtopical	CST sem Peso	CST com Peso
0.7227	0.5072	0.4850	0.4789

Pode-ser notar os valores acima superam os obtidos pelos agrupamentos automáticos. Esta superação já era esperada, pois, como mencionado anteriormente, os agrupamentos foram realizados por humanos. Em adição, percebe-se que os valores obtidos pelo método Palavras-Chave (novamente os mais altos entre todos) diferem consideravelmente dos demais métodos; isto se dá pelo fato da técnica considerar somente as palavras mais relevantes de um subtópico, logo, agrupando todos eles de forma mais precisa.

Segue abaixo, mais detalhadamente na Tabela 6, a comparação entre os resultados dos agrupamentos obtidos pela seguinte fórmula de desvio (Fórmula 5).

(5)

$$D = \frac{r - a}{a}$$

Tem-se na fórmula acima: *r* - resultado de referência; *a* - resultado automático; *D* - desvio obtido. Com isso, temos, na Tabela 6, os valores correspondentes aos desvios de cada método de agrupamento.

Tabela 6. Comparação entre os agrupamentos automático e manual

Palavras-Chave	Similaridade Subtopical	CST sem Peso	CST com Peso
0.0643	0.0444	0.0837	0.0597

Através da análise dos desvios encontrados a partir da comparação entre os agrupamentos automático e manual, pode-se concluir que os métodos automáticos utilizados não estão muito distantes dos utilizados por humanos. Isto demonstra a boa qualidade do algoritmo e técnicas utilizados.

O último passo da etapa de avaliação corresponde à análise dos sumários automáticos em relação aos manuais, o qual, para os devidos fins apoiou-se no uso da *ROUGE* (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin e Hovy, 2003), que é um pacote de métricas para tal finalidade. Basicamente, essa ferramenta foi criada para possibilitar a comparação direta entre um resumo gerado automaticamente e o respectivo sumário humano. Segundo as definições de Lin e Hovy, a ROUGE faz o cálculo baseando-se nos conjuntos de palavras em sequência (ou seja, os n-gramas que podem variar de 1 a 4) em comum entre os sumários de referência humanos e os sumários gerados automaticamente. Tem-se também a ROUGE-L, utilizada neste trabalho, em que se faz uma comparação da maior subsequência de palavras em comum entre os sumários humanos e automáticos. Sabe-se que a ROUGE avalia a informatividade de um sumário, ou seja, qual é o nível de informação que o mesmo conseguiu capturar dos textos-fonte correspondentes.

Uma forma de avaliar os sumários extrativos, como antes mencionados que fazem parte do córpus CSTNews, é por meio das medidas tradicionais de Precisão (*Precision*), Cobertura (*Recall*) e Medida-F (*F-Measure*). Tal avaliação é necessária, pois, ao contrário dos sumários manuais, os sumários extrativos fornecem uma comparação entre sentenças inteiras, sem perda de conteúdo. Segue então abaixo a explicação sucinta de cada uma das medidas apresentadas acima:

- Precisão: representa a quantidade de sentenças relevantes selecionadas dentre as sentenças do sumário automático;

- Cobertura: representa a quantidade de sentenças selecionadas dentre as sentenças do sumário humano;

- Medida-F: média harmônica da precisão e da cobertura. Essa medida é utilizada para medir a exatidão do classificador, ou seja, quanto mais próximo de 1, melhor o sumário, e os resultados mais próximos de 0 demonstram que os sumários são ruins. A medida-F é útil por combinar em uma única medida tanto a Precisão quanto a Cobertura.

Abaixo, temos, na Tabela 7, os resultados provenientes deste trabalho. É válido notar que: a) as colunas ROUGE (P), ROUGE (C) e ROUGE (F) correspondem à Precisão, Cobertura e Medida-F da medida ROUGE-L, respectivamente; b) o método de seleção de conteúdo '*Baseline*' corresponde aos sumários gerados com sentenças aleatórias para fins de comparação; c) como os valores de Precisão, Cobertura e Medida-F para o *Baseline* estariam muito próximos de zero (pelo seu caráter aleatório), decidiu-se que não seria relevante analisar tais resultados; d) o método utilizado neste trabalho, como antes mencionado, condiz com o Denso Segmentado, logo as diferentes medidas correspondem aos diferentes métodos de agrupamento aplicados (1: agrupamento entre subtópicos completos; 2: agrupamento entre palavras-chave; 3: agrupamento utilizando somente a quantidade de conexões CST; 4: agrupamento utilizando as ponderações das conexões CST).

Tabela 7. Resultados

Método de Seleção de Conteúdo	**ROUGE (P)**	**ROUGE (C)**	**ROUGE (F)**	**Precisão**	**Cobertura**	**Medida F**
Caminho Denso Segmentado[1]	0.5472	0.3517	0.4190	0.2850	0.1714	0.2141
Caminho Denso Segmentado[2]	0.5507	0.3297	0.4023	0.2722	0.1600	0.2015
Caminho Denso Segmentado[3]	0.6079	0.2802	0.3571	0.3206	0.1470	0.2016
Caminho Denso Segmentado[4]	0.6033	0.2879	0.3637	0.3186	0.1498	0.2038
Baseline	0.3015	0.2900	0.2948	—	—	—

Os resultados apresentados acima, obtidos pela ROUGE, ainda que um pouco distantes do ideal (valor 1) foram muito satisfatórios superando não somente os sumarizadores atuais do estado da arte, como também os dois primeiros caminhos propostos por Salton et al. (1997) implementados no trabalho anterior. A ocorrência dos valores acima se deu pelo fato de que a prioridade era de se selecionar ao menos uma sentença de cada subtópico, além de uma sentença de transição entre cada um deles. Desta forma, muitas vezes, atingiu-se a taxa de

compressão logo após a seleção de uma sentença de transição, deixando de fora outras informações importantes (sentenças de outros subtópicos) do sumário (isso pode ser verificado, na tabela acima, com a medida de Cobertura da ROUGE).

Em relação à análise de Cobertura, Precisão e Medida-F, estas ficaram abaixo do esperado pelo fato de que as comparações (sumários automáticos – sumários manuais) não suportavam qualquer divergência, neste caso, palavras diferentes. Porém, como o conteúdo das sentenças presentes no sumário automático e no extrativo eram muito similares, e pelo motivo apresentado acima (informações importantes eliminadas pela taxa de compressão), o resultado não foi tão alarmante.

A seguir, na Tabela 8, temos uma comparação com outros sumarizadores do Português. São apresentados os resultados obtidos para os sumarizadores RSumm (Ribaldo et al., 2012), CSTSumm (Castro Jorge e Pardo, 2010), MEAD (Radev et al., 2000), GistSumm e RCSumm (Akabane et al., 2011), possibilitando-se saber a qualidade do sumarizador desenvolvido neste projeto em relação a outros sumarizadores de referência para o português. O sumarizador RSumm condiz com o trabalho anterior onde os dois primeiros caminhos de Salton et al. (1997) foram implementados, como antes mencionado. Em relação ao sumarizador CSTSumm, este investiga métodos de seleção de conteúdo através da CST, produzindo sumários tanto genéricos quando baseados na preferência do usuário. A ferramenta MEAD incorpora múltiplas estratégias para seleção de sentenças para sumarização, a saber: 1) a posição das sentenças; 2) distância lexical das sentenças em relação ao centróide (o peso de cada aresta é demarcado com a similaridade de cosseno), ou melhor, à sentença central do texto (que contém mais relacionamentos); 3) maior subsequência em comum e 4) palavras-chave. Em relação ao GistSumm, este utiliza um método extrativo que usa uma única sentença (ideia principal) para então selecionar os melhores segmentos textuais que farão parte do sumário. Por fim, o sumarizador RCSumm explora medidas de redes complexas para encontrar as sentenças mais relevantes em um determinado conjunto de textos que versam sobre um mesmo assunto. Em adição, foram também adicionados resultados dos sumarizadores MEAD e GistSumm com a incorporação de conhecimento discursivo da CST para uma melhor comparação.

É importante notar que a ferramenta de análise (ROUGE), a fonte (CSTNews) e a taxa de compressão de 70% usadas para que os sumários fossem analisados e gerados, tanto pelo sistema desenvolvido neste projeto, quanto pelos outros, foram as mesmas. Em adição, não foram utilizadas as medidas tradicionais de Precisão, Cobertura e Medida-F por estas não considerarem n-gramas (necessários para as comparações).

Tabela 8. Comparação

Método de Seleção de Conteúdo	ROUGE (P)	ROUGE (C)	ROUGE (F)
Caminho Denso Segmentado[1]	0.5472	0.3517	**0.4190**
RCSumm	0.4218	0.4036	0.4102
Caminho Denso Segmentado[2]	0.5507	0.3297	**0.4023**
MEAD com CST	0.4257	0.3876	0.4018
RSumm (Denso)	0.4089	0.3704	0.3871
CSTSumm	0.4472	0.3557	0.3864
RSumm (Profundo)	0.3977	0.3630	0.3795
Caminho Denso Segmentado[4]	0.6033	0.2879	**0.3637**
MEAD sem CST	0.3691	0.3574	0.3616
GistSumm com CST	0.2800	0.5229	0.3583
GistSumm sem CST	0.3923	0.3343	0.3581
Caminho Denso Segmentado[3]	0.6079	0.2802	**0.3571**
Baseline	0.3015	0.2900	0.2948

Sobre os resultados coletados, a partir da ROUGE, estes foram muito satisfatórios. Tem-se que o melhor resultado, Caminho Denso Segmentado[1], o qual corresponde ao agrupamento comparando-se subtópicos inteiros para a então aplicado do terceiro caminho de Salton et al., ficou acima do sumarizador atual do estado da arte: RCSumm e ainda dos demais métodos de Salton et al.: RSumm (Denso) e RSumm (Profundo). Como mencionado anteriormente, o RCSumm utiliza redes complexas para o descobrimento de sentenças relevantes a serem levadas ao sumário.

É importante notar que o segundo método de agrupamento (Palavras-Chave) obteve resultados secundários em relação ao Caminho Denso Segmentado[1], ainda que os valores obtidos pela avaliação do agrupamento do primeiro tenham sido maiores que o do segundo. A hipótese da ocorrência desse fato implica que quando se utiliza as palavras que melhor definem um subtópico (palavras-chave) a comparação é mais rígida e concentrada. Levando em conta a comparação entre subtópicos inteiros, esta é mais relaxada e tolerante

à palavras que apareçam somente uma vez em cada subtópico; aumentando assim a similaridade lexical e, consequentemente, a chance de agrupamento (mesmo que os subtópicos não tenham uma similaridade significante).

Os resultados obtidos pelos terceiro e quarto métodos de agrupamento: utilização da *Cross-document Theory* não foram tão altos quanto o esperado, mas ainda assim a diferença entre os principais sumarizadores atuais não foi grande. Em adição, os valores de ROUGE do *Baseline* estão de acordo com o esperado, pois, como antes mencionado, esta abordagem seleciona sentenças de forma aleatória. Logo, esperava-se que o sumário gerado fosse de baixa qualidade.

Por fim, nas Tabelas 9, 10 e 11, verifica-se a porção de desvio encontrada nos valores coletados para cada um dos casos (caminhos Denso Segmentado – de 1 à 4 – e *Baseline*). A fórmula abaixo, análoga à Fórmula 5 previamente apresentada, representa o cálculo do desvio entre os caminhos Denso Segmentado e os sumarizadores de referência.

(6)

$$D = \frac{ds - sr}{sr}$$

Tem-se na fórmula acima: *ds* - resultado caminho Denso Segmentado; *sr* - resultado sumarizadores de referência; *D* - desvio obtido. Tal análise do desvio[7] é importante, pois demonstra o quão distante os resultados obtidos por este trabalho estão em relação aos sumarizadores analisados.

[7] Foi considerado o cálculo do Teste *t-Student* para uma melhor comparação. Porém, os dados necessários (sumários automáticos gerados por cada uma das ferramentas) não puderam ser adquiridos.

Tabela 9. Comparação Denso Segmentado[1]

Método de Seleção de Conteúdo	ROUGE (P)	ROUGE (C)	ROUGE (F)
RCSumm	0.2973	-0.1286	0.0215
MEAD com CST	0.2854	-0.0926	0.0428
RSumm (Denso)	0.3382	-0.0505	0.0824
CSTSumm	0.2236	-0.0112	0.0844
RSumm (Profundo)	0.3759	-0.0311	0.1041
MEAD sem CST	0.4825	-0.0159	0.1587
GistSumm com CST	0.9543	-0.3274	0.1694
GistSumm sem CST	0.3949	0.0520	0.1701

Neste caso, o melhor caminho comprovado, Denso Segmentado[1], obteve seu menor desvio da medida ROUGE (F) com o sumarizador RCSumm e o maior com o GistSumm sem a adição de conhecimento discursivo da CST. Esse resultado confirma a boa qualidade dos sumários gerados pelo caminho Denso Segmentado[1].

Tabela 10. Comparação Denso Segmentado[2]

Método de seleção de conteúdo	ROUGE (P)	ROUGE (C)	ROUGE (F)
RCSumm	0.3056	-0.1831	-0.0193
MEAD com CST	0.2936	-0.1494	0.0012
RSumm (Denso)	0.3468	-0.1099	0.0393
CSTSumm	0.2314	-0.0731	0.0411
RSumm (Profundo)	0.3847	-0.0917	0.0601
MEAD sem CST	0.4920	-0.0775	0.1126
GistSumm com CST	0.9668	-0.3695	0.1228
GistSumm sem CST	0.4038	-0.0138	0.1234

A tabela acima apresenta os desvios encontrados quando comparados o caminho Denso Segmentado[2] e os demais sumarizadores. Nota-se que os desvios foram um pouco menores dos apresentados na Tabela 9. Em adição, em comparação com o RCSumm, a diferença foi negativa; sendo assim, comprova-se que, neste caso, o sumarizador RCSumm foi melhor em relação a utilização de palavras-chave para o agrupamento de subtópicos.

Tabela 11. Comparação Denso Segmentado[3]

Método de seleção de conteúdo	ROUGE (P)	ROUGE (C)	ROUGE (F)
RCSumm	0.4412	-0.3057	-0.1294
MEAD com CST	0.4280	-0.2771	-0.1112
RSumm (Denso)	0.4867	-0.2435	-0.0775
CSTSumm	0.3593	-0.2123	-0.0758
RSumm (Profundo)	0.5285	-0.2281	-0.0590
MEAD sem CST	0.6470	-0.2160	-0.0124
GistSumm com CST	1.1711	-0.4641	-0.0033
GistSumm sem CST	0.5496	-0.1618	-0.0028

Na tabela acima é possível notar que as divergências foram bem maiores, ou melhor, o sumários gerados pelo método Denso Segmentado[3] são de menor qualidade quando comparados com outros sumarizadores. Esse resultado demonstra que somente a utilização do número de relações CST entre subtópicos não é suficiente para agrupá-los de forma correta.

Tabela 12. Comparação Denso Segmentado[4]

Método de seleção de conteúdo	ROUGE (P)	ROUGE (C)	ROUGE (F)
RCSumm	0.4303	-0.2867	-0.1134
MEAD com CST	0.4172	-0.2572	-0.0948
RSumm (Denso)	0.4754	-0.2227	-0.0604
CSTSumm	0.3491	-0.1906	-0.0587
RSumm (Profundo)	0.5170	-0.2069	-0.0416
MEAD sem CST	0.6345	-0.1945	0.0058
GistSumm com CST	1.1546	-0.4494	0.0151
GistSumm sem CST	0.5379	-0.1388	0.0156

Mais uma vez tem-se uma alta discrepância dos valores de desvio da comparação entre o caminho Denso Segmentado[4] (CST ponderada) e os sumarizadores da atualidade. Pode-se notar, ainda que há disparidade entre as tabelas 11 e 12; porém, estas não são tão grandes. Logo, conclui-se os sumarizadores avaliados aqui geram resultados melhores em comparação com o método Denso Segmentado[3].

Tabela 13. Comparação *Baseline*

Método de seleção de conteúdo	ROUGE (P)	ROUGE (C)	ROUGE (F)
RCSumm	-0.2852	-0.2815	-0.2813
MEAD com CST	-0.2918	-0.2518	-0.2663
RSumm (Denso)	-0.2627	-0.2171	-0.2384
CSTSumm	-0.3258	-0.1847	-0.2371
RSumm (Profundo)	-0.2419	-0.2011	-0.2232
MEAD sem CST	-0.1831	-0.1886	-0.1847
GistSumm com CST	0.0768	-0.4454	-0.1772
GistSumm sem CST	-0.2315	-0.1325	-0.1768

Os resultados obtidos pelo *Baseline* foram baixos, como mostram os desvios na Tabela 13, em relação aos caminhos anteriormente utilizados. Tais valores eram esperados, pelo fato do *Baseline* conter sentenças geradas aleatoriamente.

Finalmente, tem-se abaixo a última comparação feita em relação às variações do método Denso Segmentado. Essa comparação foi realizada para avaliar somente o caminho de Salton et al. (1997), isto é, quando são considerados a segmentação topical e o agrupamento de forma manual.

Tabela 14. Comparação Segmentação Topical e Agrupamento Manuais

Método de seleção de conteúdo	ROUGE (P)	ROUGE (C)	ROUGE (F)
Caminho Denso Segmentado (Manual)	**0.5803**	**0.3918**	**0.4407**
Caminho Denso Segmentado[1]	0.5472	0.3517	0.4190
Caminho Denso Segmentado[2]	0.5507	0.3297	0.4023
Caminho Denso Segmentado[4]	0.6033	0.2879	0.3637
Caminho Denso Segmentado[3]	0.6079	0.2802	0.3571

É possível notar que os resultados são melhores, quando se realiza a segmentação e o agrupamento manuais. Esse aumento, de 5.17% em relação à Medida-F, demonstra a qualidade do método utilizado neste trabalho quando tarefas anteriores são desempenhadas de forma quase não errônea.

3.5. Dificuldades, Limitações e Trabalhos Futuros

As dificuldades encontradas no desenvolvimento deste projeto se resumiram no entendimento dos algoritmos (segmentação topical, agrupamento e caminho Denso Segmentado), os quais requisitaram leituras minuciosas dos artigos que os contêm, além da variedade de exemplos desenvolvidos especificamente para auxiliar a compreensão de tais algoritmos. Uma dificuldade secundária esteve relacionada ao desenvolvimento da extensão da ferramenta para o navegador *Google Chrome*. Esta ferramenta, em particular, exigiu conhecimentos detalhados de linguagens de programação *web*, as quais ainda não eram dominadas. Em adição, não foram encontrados impasses no desenvolvimento da ferramenta *stand-alone* (sumarizador), pois as técnicas e linguagens de programação utilizados já eram de alto conhecimento.

Em relação às limitações encontradas durante a condução deste trabalho, somente uma se sobressaiu: o cálculo do Teste *t-Student* para uma melhor e mais precisa comparação entre os valores de ROUGE de todos os sumarizadores analisados. Como explicado anteriormente, tal limitação ocorreu pela falta de disponibilidade dos sumários de cada ferramenta para que o cálculo do teste fosse realizado.

Por fim, neste projeto, em particular, foram utilizadas medidas, como a Similaridade do Cosseno, que forneceram parâmetros muito importantes que influenciaram na maioria das tomadas de decisão (por exemplo, a sentença não deve ser levada ao sumário por este já conter uma sentença similar). Uma possível abordagem aprimorada seria a consideração de medidas mais robustas para decisões relevantes; logo, espera-se que em trabalhos futuros, melhores considerações, como tratamento de redundância aprimorado e outras, possam ser utilizadas para uma geração de sumários de melhor qualidade (sem a alteração do método de seleção de informação relevante, pois este já se provou eficiente).

CAPÍTULO 4: CONCLUSÃO

É evidente a grande utilidade da sumarização automática multidocumento nos dias atuais, pois é claro que a enorme quantidade *online* de informações disponíveis cresce de forma tão rápida que lidar com tal contingente torna-se uma tarefa quase impossível de ser realizada.

O desenvolvimento deste projeto forneceu uma visão e um entendimento mais críticos quando se trata da análise e síntese de informações relevantes que necessitam permanecer concisas e, ao mesmo tempo, sucintas. Mais ainda, não somente conhecimentos linguísticos puderem ser aprimorados, mas também conhecimentos que envolvem computação. Em particular, disciplinas como Algoritmos e Estrutura de Dados, Programação Orientada a Objetos e Engenharia de Software foram de suma importância para que este trabalho fosse concluído com êxito. Em primeiro lugar, a disciplina de Algoritmos e Estrutura de Dados forneceu toda a lógica e a habilidade de sintetizar procedimentos que deveriam ser implementados. Em Programação Orientada a Objetos, conceitos relacionados a abstração e comunicação de dados permitiram uma melhor visão de comportamento dos mesmos. Em terceiro lugar (Engenharia de Software), o planejamento e a correta execução de cada etapa deste projeto foram imprescindíveis.

Conclui-se que o conhecimento e experiência adquiridos ao decorrer deste trabalho foram de extrema importância. Pôde-se aplicar aqui assuntos discutidos em sala de aula e também desenvolver conceitos de pesquisa que eram previamente desconhecidos.

4.1. Contribuições

A conclusão deste projeto é de enorme utilidade não só individualmente, mas também para a academia, pois permite uma grande utilização dos métodos e ferramentas aqui desenvolvidos para futuros trabalhos e referências.

Por fim, tal projeto é de grande conveniência e lucratividade não somente para a academia, como mencionado, mas para o crescimento, entendimento e progresso do aluno relacionado a este e à área como um todo.

4.2. Considerações sobre o Curso de Graduação

Cursar Ciências de Computação no ICMC (USP-São Carlos) foi de grande valia tanto para minha vida pessoal, quanto profissional. Sempre fui exposto a novas perspectivas, ferramentas, tecnologias e abordagens que me fizeram sair da minha zona de conforto e melhorar como pessoa e aluno.

Diversas disciplinas proporcionaram conteúdos inovadores e desafiadores, em minha opinião, como Sistemas Operacionais (SO). Em SO somos desafiados a desenvolver nossa própria plataforma de comandos. Tudo isso me permitiu enxergar além da teoria em sala de aula e obter uma visão mais detalhada e precisa de como a ferramenta desenvolvida neste trabalho impacta no sistema operacional no qual ela está executando.

Disciplinas de outras áreas, como Matemática e Estatística também tiveram um impacto positivo no desenvolvimento deste projeto. Atualmente, não é mais possível desvincular tais disciplinas da área de Computação; pelo contrário, estas são vitais para que operações complexas e análises de dados possam ser desenvolvidas.

Um dos pontos fracos no curso de Ciências de Computação, na minha opinião, é a falta de integração com outras áreas da própria computação. Em particular, no decorrer deste trabalho, nenhuma disciplina da área de Processamento de Linguagem Natural foi oferecida. Dessa forma, perde-se a visibilidade de tal área e, consequentemente, impacta negativamente alunos envolvidos em projetos relacionados, pois é necessário focar no aprendizado da área, em vez do desenvolvimento e aprimoramento de novas técnicas que poderiam influenciar diretamente os resultados finais e, indiretamente, a comunidade em geral que poderia se apoiar em algum resultado específico obtido.

REFERÊNCIAS

AFANTENOS, S.D.; DOURA, I.; KAPELLOU, E.; KARKALETSIS, V. Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In the *Proceedings of SETN*, 2004, p. 410-419, 2004.

AFANTENOS, S.D.; KARKALETSIS, V.; STAMATOPOULOS, P.; HALATSIS, C. Using synchronic and diachronic relations for summarizing multiple documents describing evolving events. *Journal of Intelligent Information Systems*, 2008, v. 30, n. 3, p. 183-226, 2008.

AKABANE, A.T.; PARDO, T.A.S.; RINO, L.H.M. Sumarização Multidocumento com Base em Métricas de Redes Complexas. In Anais do *19o Simpósio Internacional de Iniciação Científica da Universidade de São Paulo - SIICUSP*, 2011, p. 1-1. November 21-25, São Carlos/SP, Brazil.

BAXENDALE, P.B. Machine-made index for technical literature – an experiment. IBM *Journal of Research and Development*, 1958, v. 2, p. 354-361, 1958.

BLACK, W.J. and JOHNSON, F.C. *Expert Systems for Information Management. A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques*. Department of Computation. University of Manchester, Institute of Science and Technology, 1988, v. 1, n. 3, 1988.

CARBONELL, J.; GOLDSTEIN, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In the *Proceedings of ACM-SIGIR*, 1998. Melbourne/Australia.

CARDOSO, P.C.F.; PARDO, T.A.S.; NUNES, M.G.V. Métodos para Sumarização Automática Multidocumento Usando Modelos Semântico-Discursivos. In the *Proceedings of the 3rd RST Brazilian Meeting*, 2011, p. 59-74. October 26, Cuiabá/MT, Brazil.

CARDOSO, P.C.F.; MAZIERO, E.G.; CASTRO JORGE, M.L.R.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the *Proceedings of the 3rd RST Brazilian Meeting*, 2011, p. 88-105. October 26, Cuiabá/MT, Brazil.

CARDOSO, P.C.F.; PARDO, T.A.S.; TABOADA, M. On the contribution of discourse to topic segmentation. In the *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue* – SIGdial, 2013, p. 92-96. August 22-23, Metz/France.

CASTRO JORGE, M.L.R. and PARDO, T.A.S. Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, 2010, p. 74-82. July 16, Uppsala/Sweden.

CASTRO JORGE, M.L.R. and PARDO, T.A.S. A Generative Approach for Multi-Document Summarization using the Noisy Channel Model. In the *Proceedings of the 3rd RST Brazilian Meeting*, 2011, p. 75-87. October 26, Cuiabá/MT, Brazil.

CASTRO JORGE, M.L.R.; AGOSTINI, V.; PARDO, T.A.S. Multi-document Summarization Using Complex and Rich Features. In *Anais do VIII Encontro Nacional de Inteligência Artificial*, 2011, p. 1-12. July 19-22, Natal/RN, Brazil

CHOI, F. Y. Y. Advances in domain independent linear text segmentation. In the *Proceedings of ANLP*, 2000, p. 26-33, 2000.

CONROY, J. M. and O'LEARY, D. P. Text summarization via Hidden Markov models. In the *Proceedings of SIGIR '01*, 2001, p. 406-407, New York/ USA.

EDMUNDSON, H.P. New Methods in Automatic Extracting. *Journal of the ACM*, 1969, v. 16, n. 2, p. 264-285, 1969.

EVERT, Stefan. A lightweight and efficient tool for cleaning Web pages. In the Proceedings of the *6th International Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Morocco.

FUNG, B.C.M.; WANG, K.; ESTER, M. Hierarchical Document Clustering using Frequent Itemsets. In: Barbará, D., Kamath, C. (eds.) *3rd SIAM International Conference on Data Mining*, 2003, p. 59–70, 2003.

HEARST, M. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages, *Computational Linguistics*, 1997, v. 23, n. 1, p. 33-64, 1997.

LEITE, D. S. *Um estudo comparativo de modelos baseados em estatísticas textuais, grafos e aprendizado de máquina para sumarização automática de textos em português.* Dissertação apresentada ao Curso de Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, 2010, SP, Brasil.

LIN, C.Y. ROUGE: a Package for Automatic Evaluation of Summaries. In the *Proceedings of the Workshop on Text Summarization Branches Out* (WAS 2004). Barcelona, Spain.

LIN, C.Y. and HOVY, E. H. The automated acquisition of topic signatures for text summarization. In the *Proceedings of the 18th Conference on Computational Linguistics*, 2000, p. 495-501.

LIN, C.Y. and HOVY, E. H. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In the *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, p. 71-78. Edmonton, Canada.

LUHN, H. P. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 1958, n. 2, p. 157-165, 1958.

MANI, I. *Automatic Summarization*. John Benjamins Publishing Co., 2001, Amsterdam.

MANI, I. and BLOEDORN, E. Multi-document summarization by graph search and matching. In the *Proceedings of the 14th National Conference on Artificial Intelligence* (AAAI 1997), p. 622-628, 1997. American Association for Artificial Intelligence.

MANI, I. and MAYBURY, M. T. *Advances in automatic text summarization*. MIT Press, 1999, Cambridge, MA.

MARCU, D. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, 2000. Cambridge, Massachusetts.

MAZIERO, E.G.; JORGE, M.L.C.; PARDO, T.A.S. Identifying Multidocument Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science* - NLPCS 2010, p. 60-69, 2010. June 8-12, Funchal/Madeira, Portugal.

MCKEOWN, K. and RADEV, D. R. Generating summaries of multiple news articles. In the Proceedings of the *18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1995, p. 74-82, 1995, Seattle, WA.

MCKEOWN, K.; PASSONNEAU, R.; ELSON, D.; NENKOVA, A.; HIRSCHBERG J. Do Summaries Help? A Task-Based Evaluation of Multi-Document Summarization. In the *Proceedings of the 28th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 2005 Salvador, Brazil.

MIHALCEA, R. and TARAU, P. An Algorithm for Language Independent Single and Multiple Document Summarization. In the *Proceedings of the International Joint Conference on Natural Language Processing* (IJCNLP 2005). Korea.

O'DONNELL, M. Variable-Length On-Line Document Generation. In the *Proceedings of the 6th European Workshop on Natural Language Generation*, 1997. Gerhard-Mercator University, Duisburg, Germany.

OTTERBACHER, J.C.; RADEV, D.R.; LUO, A. Revisions that improve cohesion in multi-document summaries: a preliminary study. In the *Proceedings of the Workshop on Automatic Summarization*, 2002, p. 27-36, 2002. Philadelphia.

PARDO, T.A.S. and RINO, L.H.M. DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), Advances in *Natural Language Processing*, 2002, p. 263-273, 2002 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.

PARDO, T.A.S. e RINO, L.H.M. *TeMário: Um Corpus para Sumarização Automática de Textos*. Série de Relatórios do NILC, 2003. NILC-TR-03-09. São Carlos-SP, Outubro, 13p.

PARDO, T.A.S.; RINO, L.H.M.; NUNES, M.G.V. GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken – PROPOR (Lecture Notes in Artificial Intelligence 2721)*, 2003, p. 210-218, 2003. Faro, Portugal. June 26-27.

PARDO, T.A.S. *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades*. Série de Relatórios do NILC, 2005. NILC-TR-05-05. São Carlos-SP/Brasil.

PORTER, M.P. Snowball: A language for stemming algorithms. 2001.

RADEV, D.R. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, 2000. Hong Kong.

RADEV, D.R.; JING, H.; BUDZIKOWSKA, M. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In the *Proceedings of the ANLP/NAACL Workshop*, 2000, p. 21-29, 2000.

RADEV, D.R. and MCKEOWN, K. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 1998, v. 24, n. 3, p. 469-500, 1998.

RIBALDO, R.; AKABANE, A.T.; PARDO, T.A.S. Aplicando Grafos e CST (*Cross-document Structure Theory*) na Sumarização Automática Multidocumento. In Anais do *18o Simpósio Internacional de Iniciação Científica da Universidade de São Paulo - SIICUSP*, 2010. November 16-19, São Paulo/SP, Brazil.

RIBALDO, R.; PARDO, T.A.S.; RINO, L.H.M. Sumarização Automática Multidocumento com Mapas de Relacionamento. In the *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*, 2011. October 24-26, Cuiabá/MT, Brazil.

RIBALDO, R.; PARDO, T.A.S.; RINO, L.H.M. Um Método de Sumarização Multidocumento Baseado em Grafos e Informação Semântico-Discursiva. In *Anais do 19o Simpósio Internacional de Iniciação Científica da Universidade de São Paulo - SIICUSP*, 2011, p. 1-1, 2011. November 21-25, São Carlos/SP, Brazil.

RIBALDO, R.; AKABANE, A.T.; RINO, L.H.M.; PARDO, T.A.S. Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In the *Proceedings of the 10th International Conference on Computational Processing of Portuguese (LNAI 7243)*, 2012, p. 260-271, 2012. April 17-20, Coimbra, Portugal.

RIBALDO, R.; AKABANE, A.T.; PARDO, T.A.S. Multi-document Summarization with Graph Metrics. In the (on-line) *Proceedings of the PROPOR 2012 Demonstrations*, 2012, p. 1-3, 2012. April 17-20, Coimbra, Portugal.

RIBALDO, R.; CARDOSO, P.C.F.; PARDO, T.A.S. Investigação de Métodos de Segmentação e Agrupamento de Subtópicos para Sumarização Multidocumento. In *Anais do 3o Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana* - TILic, 2013, p. 1-3, 2013. October 21-23, Fortaleza/Brazil.

RIBALDO, R.; PARDO. T.A.S. Investigação de Mapas de Relacionamento para Sumarização Multidocumento. In *Anais do 21o Simpósio Internacional de Iniciação Científica da Universidade de São Paulo* - SIICUSP, 2013, p. 1-1, 2013. October 21-25, São Paulo/SP, Brazil.

SALTON, G. *Automatic text processing*. Addison-Wesley Longman Publishing Co., Inc., 1988, Boston, MA, USA.

SALTON, G.; SINGHAL A.; MITRA, M; BUCKLEY C. Automatic Text Structuring And Summarization. *Information Processing & Management*, 1997, v. 33, n. 2, p. 193-207, 1997.

SENO, E.R.M. and NUNES, M.G.V. Some Experiments on Clustering Similar Sentences of Texts in Brazilian Portuguese. In: *Proceedings of the International Conference on Computational Processing of Portuguese Language - PROPOR (Lecture Notes in Artificial Intelligence, 5190)*, 2008, p. 133-144, 2008.

STEINBACH, M.; KARYPIS, G.; KUMAR, V. A comparison of document clustering techniques. In *International Conference on Knowledge Discovery & Data Mining* - KDD 2000.

SVORE, K.; VANDERWENDE, L.; BURGES, C. Enhancing single-document summarization by combining RankNet and third-party sources. In *Proceedings of the EMNLP-CoNLL*, 2007, p. 448-457, 2007.

UZÊDA, V.R.; PARDO, T.A.S.; NUNES, M.G.V. A comprehensive summary informativeness evaluation for RST-based summarization methods. *International Journal of Computer Information Systems and Industrial Management Applications* - IJCISIM 2009, v. 1, p. 188-196, 2009.

UZÊDA, V.R.; PARDO, T.A.S.; NUNES, M.G.V. A Comprehensive Comparative Evaluation of RST-Based Summarization Methods. *ACM Transactions on Speech and Language Processing*, 2010, v. 6, n. 4, p. 1-20, 2010.

VAN RIJSBERGEN, C.J. *Information Retrieval*. 2nd edition, 1979. Butterworths, Massachusetts.

ZHANG, Z.; BLAIR-GOLDENSOHN, S.; RADEV, D.R. Towards CST-enhanced summarization. In the *Proceedings of the AAAI 2002 Conference*. Edmonton, Alberta.

WAN, X. and YANG, J. Improved affinity graph based multi-document summarization. In the *Proceedings of HLT-NAACL*, 2006, p.181-184, 2006. New York/USA.