



UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

JOÃO HOLANDA FREIRES JUNIOR

**SUMOPINIONS: SUMARIZAÇÃO AUTOMÁTICA DE OPINIÕES SOBRE PONTOS
TURÍSTICOS**

FORTALEZA

2018

JOÃO HOLANDA FREIRES JUNIOR

SUMOPINIONS: SUMARIZAÇÃO AUTOMÁTICA DE OPINIÕES SOBRE PONTOS
TURÍSTICOS

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da computação. Área de Concentração: Banco de dados

Orientador: Prof. Dr. José Antônio Fernandes de Macêdo

Coorientador: Dr. Igo Ramalho Brilhante

FORTALEZA

2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

- F933s Freires Junior, João Holanda.
SumOpinions: Sumarização automática de opiniões sobre pontos turísticos / João Holanda Freires Junior. – 2018.
80 f. : il. color.
- Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2018.
Orientação: Prof. Dr. José Antônio Fernandes de Macêdo.
Coorientação: Prof. Dr. Igo Ramalho Brilhante.
1. Aprendizado de máquina. 2. Sumarização de textos. 3. Modelagem de tópicos. 4. Mineração de opinião.
I. Título.

CDD 005

JOÃO HOLANDA FREIRES JUNIOR

SUMOPINIONS: SUMARIZAÇÃO AUTOMÁTICA DE OPINIÕES SOBRE PONTOS
TURÍSTICOS

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da computação. Área de Concentração: Banco de dados

Aprovada em: 29 de Novembro de 2018

BANCA EXAMINADORA

Prof. Dr. José Antônio Fernandes de Macêdo (Orientador)
Universidade Federal do Ceará (UFC)

Dr. Igo Ramalho Brilhante (Coorientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. César Lincoln Cavalcante Mattos
Universidade Federal do Ceará (UFC)

Prof. Dr. João Paulo Pordeus Gomes
Universidade Federal do Ceará (UFC)

Profa. Dra. Vlândia Célia Monteiro Pinheiro
Universidade de Fortaleza (UNIFOR)

Dedico à Deus. Aos meus pais, Aila Maria e João Holanda. Meus irmãos, Karina Holanda e Germano Tavares. E à minha esposa, Luana Montenegro.

AGRADECIMENTOS

Primeiro agradeço à Deus por toda saúde, energia, amor, paciência e condições de poder realizar a jornada de trabalho e mestrado.

Segundo, agradeço à minha família, em especial aos meus pais, Aila Maria e João Holanda, que lutaram a vida toda para me oferecer as condições de estudo necessárias e me permitirem conquistar tudo que tenho desejado. Aos meus irmãos, Karina Holanda e Germano Tavares, por serem grandes influências na minha vida e me direcionarem para o caminho certo. E claro, minha companheira e recente esposa, Luana Montenegro, que esteve ao meu lado em todos os momentos, me incentivando tanto nos momentos felizes quanto nos difíceis, sendo paciente durante várias ausências minhas e que, sem dúvida, contribuiu muito nesta jornada.

Agradeço demais ao meu orientador, Prof. Dr. José Antônio Fernandes de Macêdo, e ao meu coorientador, Dr. Igo Ramalho Brilhante, por toda a parceria, o apoio, os incentivos e, com certeza, a oportunidade a mim dada para a realização do mestrado.

Sou grato aos meus amigos, seja com participação direta ou indireta na realização do mestrado. Agradeço a companhia, o compartilhamento de conhecimento, a energia em realizar pesquisa e todas as conversas do dia-a-dia. Em especial aos amigos de trabalho, Cleilton Rocha e Josué Machado, por todos os incentivos e grande amizade. Aos amigos de mestrado, Emanuel Oliveira, Guilherme Estevão e Tercio Jorge, pela parceria e ajuda para atingirmos o objetivo maior de realizar o mestrado. E claro, a todos os meus colegas em que tive o prazer de conhecer e passar um tempo do mestrado aprendendo com cada um.

Aos meus professores de graduação, que foram influências e me capacitaram para alcançar os meus desejos profissionais e acadêmicos.

E claro, agradeço ao meu grupo de pesquisa da UFC, que me ajudou em todo o desenvolvimento deste trabalho: o Insight Data Science Lab, coordenado pelo Prof. Dr. José Antônio Fernandes de Macêdo. Agradeço a infraestrutura fornecida, aos professores, amigos e colegas em que pude conviver e aprender.

Por fim, agradeço ao Doutorando em Engenharia Elétrica, Ednardo Moreira Rodrigues, e seu assistente, Alan Batista de Oliveira, aluno de graduação em Engenharia Elétrica, pela adequação do *template* utilizado neste trabalho para que o mesmo ficasse de acordo com as normas da biblioteca da Universidade Federal do Ceará (UFC).

“A persistência é o menor caminho do êxito”

(Charles Chaplin)

RESUMO

As plataformas de viagens on-line (e.g. TripAdvisor) tornaram-se muito populares nos últimos anos, pois permitiram o acesso fácil à uma grande quantidade de opiniões, a partir das experiências prévias de seus usuários sobre pontos turísticos, acomodações, restaurantes, serviços, etc. Dessa forma, um viajante pode usar tais informações para criar um planejamento de viagem mais assertivo e adequado às suas preferências. Porém, dado o grande volume de opiniões dos usuários, a leitura e seleção de opiniões relevantes é demorada e cansativa, tornando-se inviável de ser realizada manualmente. Neste contexto, este trabalho propõe um novo método para realizar a sumarização de opiniões, visando a detecção de tópicos relevantes sobre pontos turísticos, com o objetivo de reduzir a quantidade de avaliações a serem lidas e ampliar a cobertura e diversidade de assuntos relevantes representados no sumário. Para validar a abordagem proposta, coletamos dados do TripAdvisor e aplicamos técnicas de modelagem de tópicos, processamento de linguagem natural, aprendizado de máquina, similaridade de texto e análise de sentimento para construir o sumário referente às opiniões postadas pelos usuários. Experimentos foram realizados e comparados com um método que representa o estado da arte em sumarização de multi-documentos. Os resultados foram avaliados com base em três pontos de avaliação: cobertura de tópicos, redundância dos sumários e dificuldade de leitura. A diversidade de tópicos cobertos relacionados aos pontos turísticos apresentou um aumento considerável de assuntos abordados no sumário em relação ao algoritmo competidor. Em relação à análise de redundância, os resultados demonstraram que foram gerados sumários com baixa redundância. Para as avaliações de dificuldade de leitura, os resultados também foram satisfatórios, dado que os sumários não eram difíceis de serem lidos.

Palavras-chave: Aprendizado de Máquina. Sumarização de textos. Modelagem de tópicos. Mineração de opinião.

ABSTRACT

Online travel platforms (e.g. TripAdvisor) have become very popular in recent years as they have provided easy access to a wide range of opinions, from their users' previous experiences of tourist places, accommodations, restaurants, services, etc. In this way, travelers can use such information to create more assertive travel planning according to their preferences. However, given the large volume of user opinions, the reading and selection of relevant opinions are time-consuming and tiring, making it unfeasible to be performed manually. In this context, this work proposes a new method for summarizing opinions, aiming at the detection of relevant topics on tourist places, with the objective of reducing the number of opinions to be read and to expand the coverage of the relevant issues represented in the summary. To validate the proposed approach, we collected data from TripAdvisor and applied topic modeling algorithms, natural language processing techniques, machine learning, text similarity, and sentiment analysis to construct the summary about the opinions posted by users. Experiments were performed and compared with a state-of-the-art method in multi-document summarization. The results were evaluated based on three evaluation points: topic coverage, summary redundancy and reading difficulty. The diversity of covered topics related to the tourist places presented a considerable increase of subjects addressed in the summary in relation to the competing algorithm. Regarding the redundancy analysis, the results showed that summaries with low redundancy were generated. For assessments of reading difficulty, the results were also satisfactory, since the summaries were not difficult to be read.

Keywords: Machine Learning. Automatic summarization. Topic Modeling. Opinion mining.

LISTA DE FIGURAS

Figura 1 – "Travelers talk about": o que os usuários mais comentam sobre o ponto turístico.	15
Figura 2 – "Show reviews that mention": uma busca por opiniões que citam termos específicos.	16
Figura 3 – Sumarização de opiniões sobre o Coliseu de Roma, Itália.	17
Figura 4 – Texto representado como uma distribuição de tópicos.	18
Figura 5 – Representação de modelos de tópicos	28
Figura 6 – Metodologia de Hu <i>et al.</i> (2017)	39
Figura 7 – Exemplo de sumário	41
Figura 8 – Exemplo de sentenças relacionadas aos respectivos tópicos e polaridade de sentimento	43
Figura 9 – Metodologia desenvolvida	46
Figura 10 – Exemplo de informações coletadas sobre um autor: número de opiniões (163), agradecimentos (85) e o nível do autor (6) atribuído pela plataforma.	48
Figura 11 – Informações coletadas sobre a opinião de um autor para um ponto turístico.	49
Figura 12 – Representação da etapa de pré-processamento dos dados.	50
Figura 13 – Palavras/frases indicadoras utilizadas no artigo.	56
Figura 14 – Processo de sumarização com seleção das sentenças representativas de K medóides.	59
Figura 15 – Comportamento da diversidade de tópicos sobre os 100 sumários gerados com 20 sentenças sobre a Torre Eiffel.	65
Figura 16 – Comportamento da diversidade de tópicos sobre os 100 sumários gerados com 30 sentenças sobre a Torre Eiffel.	65
Figura 17 – Comportamento da diversidade de tópicos sobre os 100 sumários gerados com 20 sentenças sobre o Parque Central.	65
Figura 18 – Comportamento da diversidade de tópicos sobre os 100 sumários gerados com 30 sentenças sobre o Parque Central.	66
Figura 19 – Distribuição de tópicos sobre os 100 sumários gerados com 20 sentenças para a Torre Eiffel.	68
Figura 20 – Distribuição de tópicos sobre os 100 sumários gerados com 30 sentenças para a Torre Eiffel.	68

Figura 21 – Distribuição de tópicos sobre os 100 sumários gerados com 20 sentenças para o Parque Central.	68
Figura 22 – Distribuição de tópicos sobre os 100 sumários gerados com 30 sentenças para o Parque Central.	69
Figura 23 – Comparativo de redundância de acordo com a quantidade de sentenças para a Torre Eiffel.	69
Figura 24 – Comparativo de redundância de acordo com a quantidade de sentenças para o Parque Central.	70
Figura 25 – Comparativo da similaridade de sentenças em relação aos tópicos para a Torre Eiffel.	70
Figura 26 – Comparativo da similaridade de sentenças em relação aos tópicos para o Parque Central.	70
Figura 27 – Dificuldade de leitura dos sumários com 20 sentenças sobre a Torre Eiffel. . .	71
Figura 28 – Dificuldade de leitura dos sumários com 30 sentenças sobre a Torre Eiffel. . .	72
Figura 29 – Dificuldade de leitura dos sumários com 20 sentenças sobre o Parque Central. .	73
Figura 30 – Dificuldade de leitura dos sumários com 30 sentenças sobre o Parque Central. .	73

LISTA DE TABELAS

Tabela 1 – As cinco palavras mais frequentes em cada tópico descoberto.	28
Tabela 2 – Conjunto de valores para a métrica Flesch.	35
Tabela 3 – Conjunto de valores para a métrica ARI (Automated Readability Index). . .	35
Tabela 4 – Conjunto de valores para a métrica ARI (Automated Readability Index). . .	36
Tabela 5 – Comparativo de trabalhos relacionados.	45
Tabela 6 – Informações das opiniões coletadas da plataforma TripAdvisor.	60
Tabela 7 – Tópicos extraídos sobre a Torre Eiffel	63
Tabela 8 – Tópicos extraídos sobre o Parque Central	63
Tabela 9 – Cobertura de tópicos para os 10 primeiros sumários sobre a Torre Eiffel. . .	64
Tabela 10 – Cobertura de tópicos para os 10 primeiros sumários sobre o Parque Central.	64
Tabela 11 – Resultados sobre a cobertura de tópicos nos sumários	66

SUMÁRIO

1	INTRODUÇÃO	15
1.1	Objetivo	17
1.2	Organização do trabalho	19
2	FUNDAMENTAÇÃO TEÓRICA	20
2.1	Sumarização automática	20
2.1.1	<i>Processo de composição do sumário</i>	20
2.1.1.1	<i>Extrativo</i>	20
2.1.1.2	<i>Abstrativo</i>	21
2.1.2	<i>Tipo de Sumarização</i>	21
2.1.2.1	<i>Genérica</i>	21
2.1.2.2	<i>Orientada à consulta</i>	22
2.1.3	<i>Cardinalidade da sumarização</i>	22
2.1.3.1	<i>Individual</i>	22
2.1.3.2	<i>Múltiplos documentos</i>	22
2.1.4	<i>Sumarização em fases</i>	23
2.1.4.1	<i>Sumarização em fase única</i>	23
2.1.4.2	<i>Sumarização em duas fases</i>	23
2.2	Abordagens para sumarização automática	23
2.2.1	<i>Abordagens não supervisionadas</i>	24
2.2.2	<i>Abordagens supervisionadas</i>	24
2.2.3	<i>Abordagens baseadas em otimização</i>	25
2.3	Requisitos para Sistemas de Sumarização	25
2.3.1	<i>Cobertura</i>	25
2.3.2	<i>Diversidade</i>	26
2.3.3	<i>Coerência</i>	26
2.3.4	<i>Equilíbrio</i>	26
2.4	Sumarização de opiniões	26
2.5	Modelagem de tópicos	27
2.5.1	<i>Métodos para modelagem de tópicos</i>	29
2.5.1.1	<i>Latent Semantic Analysis</i>	29

2.5.1.2	<i>Probabilistic Latent Semantic Analysis</i>	29
2.5.1.3	<i>Latent Dirichlet Allocation</i>	30
2.5.2	Sumarização baseada em tópicos	30
2.5.2.1	<i>Identificação de tópicos</i>	31
2.5.2.2	<i>Análise de sentimento</i>	32
2.5.2.3	<i>Sumarização</i>	32
2.5.3	Avaliação da Sumarização	32
2.5.3.1	<i>Medida ROUGE</i>	32
2.5.3.2	<i>Avaliação da dificuldade de leitura</i>	34
3	TRABALHOS RELACIONADOS	37
3.1	Sumarização automática	37
3.1.1	<i>SumView: A Web-based engine for summarizing product reviews and customer opinions</i>	37
3.1.2	<i>Multi-document summarization using closed patterns</i>	38
3.1.3	<i>Opinion mining from online hotel reviews - A text summarization approach</i>	39
3.2	Modelagem e extração de tópicos	40
3.2.1	<i>Mining and summarizing customer reviews</i>	40
3.2.2	<i>A logic programming approach to aspect extraction in opinion mining</i>	41
3.2.3	<i>Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs</i>	42
3.3	Mineração de opinião	43
4	SUMOPINIONS: SUMARIZAÇÃO AUTOMÁTICA DE OPINIÕES SOBRE PONTOS TURÍSTICOS	46
4.1	Opiniões do TripAdvisor	47
4.1.1	<i>Informações sobre o autor</i>	48
4.1.2	<i>Informações sobre a opinião</i>	48
4.2	Pré-processamento	49
4.3	Modelagem e descoberta de tópicos	51
4.4	Cálculo de importância da sentença	54
4.5	Similaridade de sentenças	56
4.6	Sumarização	58
5	EXPERIMENTOS	60
5.1	<i>Dataset</i>	60

5.2	Configuração dos experimentos	60
5.3	Resultados	61
5.3.1	<i>Diversidade de tópicos</i>	61
5.3.2	<i>Análise de redundância</i>	69
5.3.3	<i>Resultados da qualidade de leitura</i>	71
6	CONCLUSÕES E TRABALHOS FUTUROS	74
	REFERÊNCIAS	76

1 INTRODUÇÃO

Para aqueles que gostam de viajar, obter informações sobre prévias experiências em plataformas de viagens é um processo importante (CHUNG; KOO, 2015; ADY *et al.*, 2015). Informações como comidas típicas da região, trajetos que otimizem custo e, principalmente, pontos turísticos, são fatores cruciais para uma viagem agradável. A plataforma TripAdvisor, por exemplo, proporciona relatos de usuários sobre hotéis, restaurantes e pontos turísticos. Porém, sabemos que as informações relevantes estão dispersas no meio do grande conjunto de opiniões dos usuários, o que torna o processo de leitura demorado e cansativo. Portanto, coletar informações significativas de maneira rápida e efetiva pode proporcionar um planejamento de viagem objetivo e decisivo.

No TripAdvisor, por exemplo, busca-se resolver esse tipo de problema apresentando os termos mais discutidos nos textos de opiniões publicadas pelos usuários, como podemos ver na Figura 1.

Figura 1 – "Travelers talk about": o que os usuários mais comentam sobre o ponto turístico.

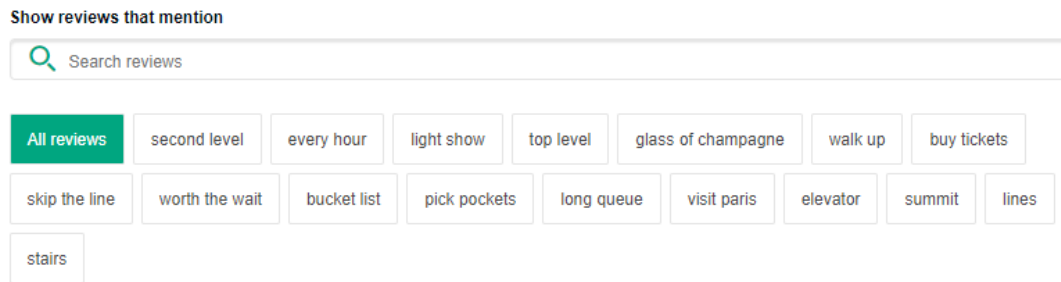


Fonte: Imagem retirada da plataforma TripAdvisor.

Na parte de "*Travelers talk about*", os termos mais comentados são listados juntos com a quantidade de opiniões que os citam. Assim, é possível que o leitor possa identificar o que mais é destacado no conjunto de opiniões publicadas. Utilizando-se da listagem de tópicos mais falados, é possível usá-los no momento da pesquisa, segundo a Figura 2.

Perceba que o TripAdvisor otimiza a busca de informações a partir do sistema de busca utilizando os termos mais citados. No entanto, mesmo selecionando os termos em destaque,

Figura 2 – "Show reviews that mention": uma busca por opiniões que citam termos específicos.



Fonte: Imagem retirada da plataforma TripAdvisor.

ainda é preciso realizar a leitura de diversas opiniões, com diferentes contextos e experiências para que se possa obter as informações importantes. Por isso, uma sumarização de opiniões sobre os pontos turísticos tornaria o consumo dessas informações um processo rápido e fácil.

Trabalhos recentes da literatura buscam otimizar o processo de obtenção de informações em plataformas de compartilhamento de experiências produzindo sumários que condensem as informações mais relevantes no conjunto de opiniões (HU *et al.*, 2017; WANG *et al.*, 2013; CATALDI *et al.*, 2013; SHIMADA *et al.*, 2011). No trabalho de Hu *et al.* (2017), por exemplo, é proposto um sumarizador de opiniões sobre hotéis na plataforma TripAdvisor. Utilizando-se de técnicas de processamento de linguagem natural e aprendizado de máquina aplicados em áreas de sumarização automática de textos e mineração de opinião, foi possível extrair as informações salientes das diferentes opiniões sobre hotéis em um conjunto de sentenças que sumarizasse tais informações.

Contudo, com a metodologia proposta por Hu *et al.* (2017), durante seu processo de sumarização é utilizado um algoritmo de máquina não supervisionado para gerar diferentes grupos conforme a similaridade entre as sentenças e, implicitamente, pode separá-las em diferentes grupos de acordo com o assunto abordado. No entanto, esse processo não identifica, explicitamente, os tópicos mais falados sobre os hotéis, assuntos estes que são constantemente compartilhados pelos usuários, seja de maneira positiva ou negativa, e relevantes de serem contemplados no sumário final para que qualquer outro leitor possa identificar rapidamente. Dessa forma, seus sumários podem apresentar conteúdos irrelevantes no ponto de vista do leitor, dado que informações pessoais ou mesmo isoladas são compartilhadas e possivelmente abordadas no sumário, dependendo da sua escrita e relevância identificada pelo sistema sumarizador.

Por outro lado, em Wang *et al.* (2013) aplica-se técnicas de extração de aspectos baseadas no trabalho de Hu e Liu (2004) onde os tópicos relevantes são extraídos sobre opiniões de usuários em relação aos produtos eletrônicos. No entanto, em sua abordagem de sumarização,

não há um texto unificado apresentando as informações de maneira construtiva e linear, são apresentadas apenas as sentenças relevantes separadas por cada tópico escolhido pelo usuário.

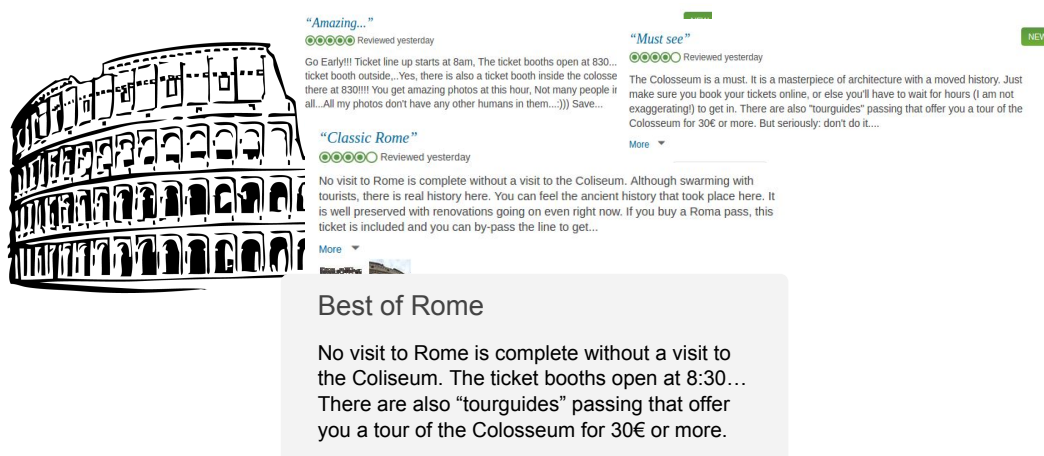
Portanto, esta dissertação aborda o problema de sumarização de opiniões sobre pontos turísticos compartilhadas pelos usuários na plataforma TripAdvisor, onde as sentenças mais relevantes do texto são extraídas com base em sua estrutura, significado semântico conforme o sentimento expresso e a presença dos tópicos mais discutidos em seu conteúdo.

Dada a natureza dinâmica de informações compartilhadas, devido ao contexto de experiência única vivida por cada usuário e a diversidade de características próprias em que cada ponto turístico possui, a descoberta de tópicos sobre as opiniões e a maximização de sua presença representa o grande problema pelo qual este trabalho se propõe a resolver.

1.1 Objetivo

Nesse cenário, pretende-se desenvolver um método de sumarização de opiniões sobre pontos turísticos buscando contemplar os tópicos mais relevantes e maximizar a diversidade de informações, no intuito de produzir sumários mais representativos e de maior importância para a leitura do usuário, otimizando o planejamento de uma viagem.

Figura 3 – Sumarização de opiniões sobre o Coliseu de Roma, Itália.

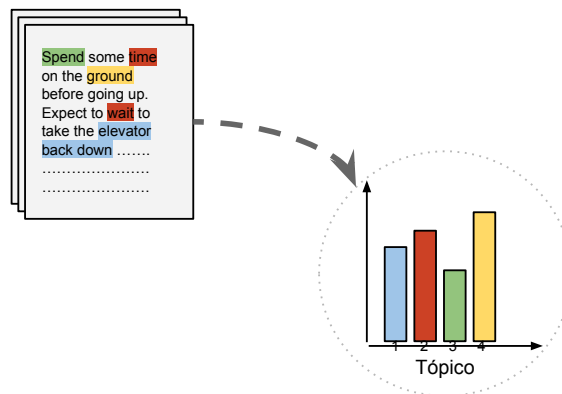


Na Figura 3, onde são apresentadas opiniões de usuários na plataforma TripAdvisor sobre o ponto turístico Coliseu localizado em Roma, representamos o objetivo deste trabalho, o qual é extrair as informações mais relevantes dos textos e gerar uma sumarização apresentando os principais tópicos, conforme apresentado no texto "*Best of Rome*", onde podemos identificar informações sobre horários de funcionamento e preços de possíveis serviços no local.

Para a identificação dos assuntos em destaque, utilizaremos algoritmos de modelagem de tópicos para representar as opiniões como um conjunto de tópicos, permitindo então extrair os termos mais importantes do conjunto de opiniões, conforme Figura 4. Algoritmos de modelagem de tópicos, diferentemente de outras abordagens de extração de tópicos (identificado também como "aspectos") baseados na frequência de substantivos (HU; LIU, 2004), a partir das relações sintáticas utilizando uma gramática de dependência (QIU *et al.*, 2011) ou mesmo regras lógicas (LIU *et al.*, 2013), são modelos probabilísticos não supervisionados que constroem uma representação semântica implícita sobre o texto a partir da coocorrências das palavras

Assim, baseado no contexto de sumarização, poderemos extrair as sentenças mais relevantes em termos de cobertura de tópicos para, em seguida, realizar o processo de sumarização. Com esse tipo de processo, podemos atingir sumários mais relevantes em termos de assuntos, os tópicos, além de caracterizar uma maior diversidade de informações em destaque. Trabalhando com modelagem de tópicos no processo de sumarização é possível também flexibilizar o conteúdo compreendido no sumário, onde poderemos gerar textos compreendendo todos os tópicos descobertos ou parte deles, caso o sistema permita o usuário submeter uma consulta informando o tipo de conteúdo que deseja ler.

Figura 4 – Texto representado como uma distribuição de tópicos.



Conforme a contextualização do objetivo deste trabalho, podemos apresentar os seguintes objetivos específicos:

- Estender o trabalho de Hu *et al.* (2017) para desenvolver uma sumarização extrativa baseada em tópicos;
- Gerar um único sumário que contemple um conjunto de tópicos relevantes a partir da extração de opiniões sobre pontos turísticos;
- Propor e avaliar os resultados da abordagem utilizando métricas de leitura para atestar a

qualidade dos sumários gerados de acordo com o grau de dificuldade de leitura, dada a ausência de sumários referências para comparação;

- Disponibilizar os dados coletados de opiniões sobre pontos turísticos, bem como a disponibilização do aplicativo coletor como um projeto *open source*;
- Facilitar o consumo de informação, especificamente sobre pontos turísticos, disponíveis em plataformas online;

1.2 Organização do trabalho

No restante deste trabalho, apresentamos a seguinte organização: no Capítulo 2 abordaremos os conceitos e definições sobre as áreas de Sumarização automática e Modelagem de tópicos, compreendidas neste trabalho. No Capítulo 3 apresentaremos os trabalhos relacionados, utilizando-os como referências para o desenvolvimento deste. Posteriormente, o Capítulo 4 apresentará a metodologia utilizada, descrevendo também os aspectos propostos em relação aos trabalhos anteriores. Após a explanação do processo, no Capítulo 5 serão apresentados os resultados de experimentos em relação a diversidade de tópicos, análise de redundância e a dificuldade de leitura dos sumários gerados em comparação com o trabalho de Hu *et al.* (2017). Por fim, o Capítulo 6 expõe a visão final do projeto, esclarecendo os resultados alcançados e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Para que os objetivos sejam alcançados, duas grandes áreas de pesquisa foram abordadas: sumarização automática e modelagem de tópicos. Logo, os conceitos e características serão apresentados nas próximas seções.

2.1 Sumarização automática

A sumarização automática consiste em criar uma representação textual envolvendo os principais pontos de informação de uma coleção de textos (MANI; MAYBURY, 2001). Segundo a literatura, possui como características: o processo de composição do sumário, o tipo de sumarização, a cardinalidade da coleção textual utilizada e a quantidade de fases no processo de sumarização (NENKOVA *et al.*, 2011).

2.1.1 Processo de composição do sumário

Tendo como objetivo criar um sumário ao extrair, representar e condensar informações relevantes sobre os documentos originais, representados como um conjunto de n documentos $D = \{d_1, d_2, d_3, \dots, d_n\}$, temos dois processos principais de composição: extrativa e abstrativa. Abaixo, os processos serão discutidos, respectivamente, com mais detalhes.

2.1.1.1 Extrativo

O sumário é construído a partir de fragmentos semânticos (sentenças ou parágrafos) não modificados dos documentos originais, selecionando os fragmentos com a melhor cobertura em relação aos aspectos da coleção original. Contudo, os métodos mais comuns de sumarização ocorrem quando o sumário é construído a partir das sentenças extraídas dos documentos originais contendo as informações relevantes (ALIGULIYEV, 2009). Assim, suponha o conjunto de todas as sentenças do corpo textual formado por k sentenças de cada documento d como $U_{s,d} = \{s_{1,1}, s_{2,1}, s_{1,2}, \dots, s_{k,d}\}$, a sumarização extrativa pode ser representada como $Sum_{ext} : U_{s,d} \rightarrow S_e$, onde $S_e \subseteq U_{s,d}$.

De modo geral, métodos extrativos para sumarização podem ser subdivididos em três atividades, não necessariamente independentes: (1) pré-processamento dos documentos, que envolve tarefas de Processamento de Linguagem Natural (PLN), por exemplo a segmentação de

texto, análise morfossintática e reconhecimento de entidades nomeadas, bem como a Mineração de Textos, como remoção de *stopwords*, redução de palavras ao radical, normalização e representação vetorial de fragmentos textuais; (2) geração de escores, que estabelecem a saliência das sentenças e indicam a preferência de inclusão no sumário; (3) seleção das sentenças, visando produzir sumários com diversidade de tópicos (NENKOVA; MCKEOWN, 2012).

Em geral, esse processo de composição é considerado mais simples que o abstrativo, mas apresenta como fragilidade a tendência em selecionar de sentenças longas e que apresentam termos irrelevantes (BING *et al.*, 2015).

2.1.1.2 Abstrativo

O sumário é obtido através da geração ou reformulação de linguagem natural, buscando reproduzir a maneira como as pessoas sintetizam textos. Os itens do sumário são produzidos mediante técnicas de reescrita de texto ou abordagens generativas. No primeiro tipo, reformula-se as sentenças originais de sumários extrativos por meio de operações como fusão, substituição e compressão. No segundo tipo, o texto é gerado diretamente com base na sentença original e pode conter termos não presentes na coleção de documentos (DAS; MARTINS, 2007; NENKOVA *et al.*, 2011). Pode ser formalmente representado como $Sum_{abs}: U_{s,d} \rightarrow S_e \times L$, onde L é o conjunto léxico de uma determinada linguagem natural e $S_e \subseteq U_{s,d}$, sendo $U_{s,d}$ o conjunto de sentenças do *corpus*, conforme foi previamente definido.

2.1.2 Tipo de Sumarização

O sumário reflete uma coleção de aspectos relevantes condensados sobre os documentos originais. Contudo, podemos ter sumários que contemple genericamente um conjunto diverso de aspectos ou podemos especificar quais tipos de aspectos gostaríamos de identificar. Portanto, temos os tipos de sumarização genérica e orientada à consulta, detalhados a seguir:

2.1.2.1 Genérica

Apresenta um sumário geral que contém todos os aspectos da coleção ou documento, sem fazer suposições sobre o tipo de conteúdo de entrada. Assim, o objetivo é encontrar um subconjunto de sentenças relevantes para todos os documentos de forma a representá-los por completo, até mesmo substituindo a necessidade de leitura do documento original.

2.1.2.2 *Orientada à consulta*

O processo de sumarização se inicia após a submissão de um conjunto de termos (consulta) que definem o tópico a ser sintetizado. Durante a construção do sumário, fragmentos de texto mais relevantes para a consulta são favorecidos. Processo de sumarização semelhante a um motor de Busca Web: a coleção de documentos é recuperada, processada e gera como saída o sumário. Requer eficiência, já que o processo é disparado por meio de uma interação do usuário, cuja expectativa é obter a resposta no menor tempo possível.

2.1.3 *Cardinalidade da sumarização*

Para a produção do sumário podemos considerar um ou mais documentos, apresentando cardinalidade individual ou para múltiplos documentos, respectivamente.

2.1.3.1 *Individual*

O objetivo é produzir sumário para apenas um documento, mesmo que outros sejam avaliados durante o processo de sumarização. Abordagens abstrativas sumarizam um documento através da geração de títulos (LOPYREV, 2015). Métodos extrativos usam informações externas ao documento para expandir o vocabulário, através de outros documentos (WAN *et al.*, 2007) ou resolução de anáforas (DURRETT *et al.*, 2016). Comumente, a sumarização individual não é tão estudada quanto à multi-documentos, possivelmente por conta da dificuldade de sumarizar um único documento.

2.1.3.2 *Múltiplos documentos*

Há, no mínimo, dois documentos envolvidos na sumarização. Em certas abordagens, o sumário de múltiplos documentos é montado a partir de sumários individuais. Pode apresentar sentenças sobre um mesmo tópico geral ao longo dos documentos da coleção ou conter uma diversidade representativa sobre os diferentes aspectos em relação à coleção de textos. Na construção do sumário, para evitar ambiguidade, deve-se excluir sentenças similares a aquelas já escolhidas para compor o sumário (RADEV *et al.*, 2004a).

2.1.4 Sumarização em fases

Há dois processos principais na construção de um sumário: seleção de sentenças de um *corpus* e o processo de minimização da redundância de aspectos contemplados referentes aos documentos originais. Por isso, dividimos o processo de sumarização em fases, como explicado abaixo.

2.1.4.1 Sumarização em fase única

Nesse tipo de sumarização são escolhidas as sentenças mais relevantes do corpo textual para integrar ao sumário, ao mesmo tempo em que busca minimizar a redundância de assuntos compreendidos. Certos métodos de aprendizado não supervisionados não passam pela fase posterior de tratamento de redundância, pois se supõe que o cálculo dos escores leva em conta a sua saliência e o agrupamento das sentenças reduz a duplicidade de informação.

2.1.4.2 Sumarização em duas fases

Nos métodos de sumarização em duas fases o cálculo da saliência das sentenças e tratamento da redundância são realizados em etapas separadas:

Geração dos escores: para cada sentença no conjunto $U_{s,d}$, que representa todas as sentenças do corpo textual, atribui-se o escore E_s , calculados por meio da combinação de atributos e heurísticas que visam capturar a relevância das mesmas.

Seleção de sentenças: refere-se à construção do sumário $U \subset U_{s,d}$ para o usuário. Após o cálculo dos escores, o sistema seleciona K sentenças mediante alguma abordagem que visa maximizar a diversidade de tópicos contemplados em U .

2.2 Abordagens para sumarização automática

Há várias abordagens de soluções presentes na literatura para o processo de sumarização automática de textos e elas podem ser agrupadas em três tipos: não supervisionadas, supervisionadas e híbridas, detalhadas abaixo.

2.2.1 *Abordagens não supervisionadas*

Métodos extrativos clássicos, baseados em centroides, são embasados na atribuição de escores às sentenças a partir de funções que combinam pesos dos termos que a constituem. O peso atribuído aos termos é calculado com base em estatísticas extraídas da coleção de documentos (RADEV *et al.*, 2004b; LIN; HOVY, 2002; NENKOVA; MCKEOWN, 2012). Em trabalhos mais recentes, baseados em grafos (ERKAN; RADEV, 2004; MIHALCEA; TARAU, 2005), o conjunto de sentenças é transformado em um gráfico, de modo que sentenças semelhantes são conectadas. O cálculo de relevância é dado pela centralidade da sentença, obtida por meio de variações do algoritmo PageRank (PAGE *et al.*, 1999).

Em métodos abstrativos, encontram-se abordagens de reformulação de texto (compressão) não supervisionadas que utilizam recursos de linguagem natural para determinar a estrutura sintática das sentenças. Baseiam-se na construção de gramáticas que fazem a tradução de estruturas sintáticas complexas para outras mais simples (COHN; LAPATA, 2008). Em Ganesan *et al.* (2010), a informação redundante de múltiplos documentos foi associada com elementos sintáticos simples para reformular as sentenças.

2.2.2 *Abordagens supervisionadas*

Para gerar sumários extrativos, aprendem uma função capaz de atribuir escores às sentenças a partir de um conjunto de treinamento, dado por sumários de referência. Em geral, são modeladas como um problema de aprendizagem de ranking (TOUTANOVA *et al.*, 2007; METZLER; KANUNGO, 2008). Assim, para extrair o conjunto de exemplos para treinar o modelo, atribuem-se escores às sentenças do conjunto de treinamento com base na sobreposição entre seus termos e os termos dos sumários de referência (CAO *et al.*, 2015). Sentenças que apresentam maior sobreposição devem aparecer no início do ranking e, portanto, recebem escores maiores. A função a ser aprendida deve minimizar o erro entre os rankings gerados para as sentenças do conjunto de treinamento e ser capaz de generalizar a construção do ranking para coleções fora desse conjunto.

Métodos abstrativos supervisionados são treinados com pares formados por uma sentença maior e a sentença simplificada correspondente. Em geral, são métodos generativos, que fazem uso de técnicas de Aprendizagem Profunda (GOODFELLOW *et al.*, 2016) para aprender a gerar linguagem natural. Costumam fazer uso de pouca informação sintática, de modo que o

modelo de linguagem capaz de detectar sentenças válidas de um idioma é inferido do conjunto de treinamento (RUSH *et al.*, 2015).

2.2.3 Abordagens baseadas em otimização

Tiram proveito do escore gerado para sentenças, mediante técnicas supervisionadas ou não, para construir o sumário que maximiza uma função objetivo. Durante a construção do sumário, priorizam, além do escore, sentenças que diversificam o conjunto de termos do sumário (MCDONALD, 2007; GILLICK; FAVRE, 2009). Ao diversificar o conjunto de termos, implicitamente, trata-se a redundância. Enquanto certas abordagens supervisionadas e não supervisionadas precisam de uma etapa posterior para promover a diversidade, aqui a maximização do escore e a diversificação são feitas através de um arcabouço integrado (LIN; BILMES, 2011).

2.3 Requisitos para Sistemas de Sumarização

Abordagens para o problema de sumarização automática recebem como entrada um conjunto de documentos e produzem um sumário com os aspectos mais representativos da coleção. Além disso, há a necessidade de eliminar a redundância da informação oriunda de múltiplos documentos relacionados. Para alcançar essas características na sumarização, é possível enumerar um conjunto de requisitos tidos como efetivos para a geração de bons sumários (LI *et al.*, 2009).

2.3.1 Cobertura

O sumário deve conter aspectos importantes da coleção. Ao levar em consideração a cobertura, a perda de informação na sumarização é minimizada (ALGULIEV *et al.*, 2011). Em métodos extrativos, alcança-se esse objetivo quando são selecionadas sentenças que contêm termos importantes para o conjunto de documentos. Essa característica enfatiza a necessidade de construir sumários com sentenças que apresentam palavras-chave da coleção. A relevância de uma sentença, com base nesse requisito, é chamada de “saliência”. Nos métodos orientados à consulta, o sumário deve conter, adicionalmente, termos semanticamente similares à consulta.

2.3.2 *Diversidade*

Prioriza sumários concisos e com pouca redundância. Isto é, se duas sentenças fornecem informação similar, então ambas não devem estar presentes no sumário ao mesmo tempo. Quando há esforço para aumentar a diversidade, a redundância é minimizada como consequência.

2.3.3 *Coerência*

A ordem das sentenças no documento deve seguir uma sequência lógica. Com esse objetivo, sentenças similares ou que tratam da mesma entidade devem aparecer fisicamente próximas (LAPATA; BARZILAY, 2005). Quando se conhece o período de tempo que o documento foi publicado, um modo simples de lidar com essa situação é ordenar as sentenças pela data de publicação do documento de origem.

2.3.4 *Equilíbrio*

O sumário deve enfatizar os diversos aspectos mencionados acima de modo equilibrado. Sumários desbalanceados tendem a não transmitir ao usuário a informação sobre os vários aspectos presentes nos documentos.

Os requisitos descritos acima permitem construir sumários informativos. Além deles, há também outros aspectos relacionados a apresentação do sumário. Os sumários devem ser interessantes para o usuário, de modo que apresentem conteúdo significativo sobre os aspectos dos documentos de maneira legível. O requisito “coerência” contribui para que esse objetivo seja atingido. Adicionalmente, espera-se que a informação seja transmitida de maneira direta, concisa e com termos que facilitam a compreensão dos aspectos da coleção.

2.4 *Sumarização de opiniões*

A sumarização de opinião é o processo de construção de um sumário baseado na orientação semântica do autor quanto à sua opinião. Entende-se uma opinião como um texto construído onde representa um pensamento a respeito de uma entidade, objeto ou tópico, seja ele positivo, negativo ou neutro (PANG *et al.*, 2008). Dessa forma, a sumarização pode ser representada como a combinação da Mineração de Opiniões e o processo de Sumarização

automática, onde a mineração identifica o sentimento implícito envolvido nas sentenças e a sumarização, por sua vez, identifica a sentenças mais relevantes e as escolhe maximizando a diversidade (CONDORI, 2014).

É possível dividir o processo de sumarização de opiniões em três principais abordagens: sumarização tradicional, contrastiva e baseada em tópicos. Como já explanado, a sumarização tradicional remete ao processo de sumarizar o *corpus* sem levar em consideração o sentimento ou conjunto de tópicos envolvidos no(s) documento(s). Portanto, sumarizar uma opinião é construir um resumo apenas com a sentenças mais relevantes encontradas. Por outro lado, preocupando-se com o sentimento envolvido do autor, bem como os tópicos abordados na opinião, tem-se a sumarização contrastiva, a qual seleciona sempre pares de sentenças que remetem à uma mesma entidade ou tópico, mas que refletem sentimentos contrários. E, por fim, a sumarização baseada em tópicos constrói um sumário para cada tópico extraído (LIU, 2012).

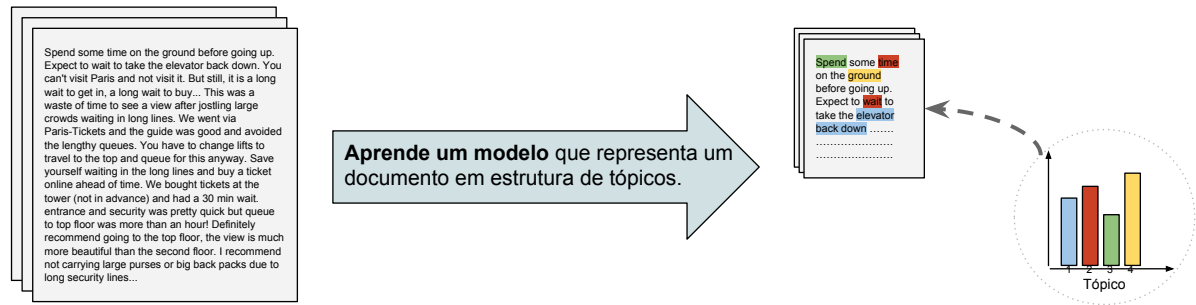
Diferentemente de uma abordagem exclusiva, este trabalho se propõe a unir as características dos três grupos explanados ao sumarizar os documentos originais com suas respectivas sentenças mais relevantes, considerando o sentimento implícito envolvido nas sentenças que, por sua vez, são selecionadas de acordo com sua importância em relação aos tópicos descobertos.

2.5 Modelagem de tópicos

Modelos de tópicos são baseados na ideia de que documentos são representados como uma distribuição de tópicos (BLEI *et al.*, 2003; STEYVERS; GRIFFITHS, 2007). Segundo Hu e Liu (2004), um aspecto ou tópico representa uma característica ou função de uma entidade. Por exemplo, "A Torre Eiffel tem uma bela paisagem durante a noite!", temos "Torre Eiffel" como uma entidade e "paisagem" como um tópico discutido. Contudo, no contexto geral da modelagem e descoberta de tópicos, um tópico representa uma distribuição de palavras correlacionadas. Portanto, são modelos probabilísticos com o objetivo de descobrir uma estrutura semântica subjacente de uma coleção de documentos, como podemos observar na Figura 5.

A modelagem de tópicos tem sido aplicada em diferentes tipos de documentos e pode ser expandido para diferentes domínios de aplicação que envolvam uma coleção de dados, como em áreas de recuperação de dados e filtragem colaborativa (BLEI *et al.*, 2003). Ao descobrir padrões de uso de palavras e conectar documentos que exibem semelhança nos padrões, tem se tornado uma técnica poderosa para a descoberta de estruturas implícitas em coleções de documentos não estruturados (BLEI; LAFFERTY, 2009). É importante destacar que esse tipo de

Figura 5 – Representação de modelos de tópicos



algoritmo não tem conhecimento prévio sobre a estrutura encontrada, que no caso de documentos para sumarização, nenhuma conhecimento a priori sobre os temas e tópicos abordados. Como exemplo na Tabela 1, temos um conjunto de tópicos descobertos a partir de opiniões sobre um ponto turístico:

Tabela 1 – As cinco palavras mais frequentes em cada tópico descoberto.

Tópico	1	2	3	4	5
#0	queue	tickets	long	queues	booked
#1	paris	tower	night	eiffel	day
#2	tower	eiffel	tickets	paris	views
#3	visit	tower	loved	place	queue
#4	view	long	people	lines	enjoy
#5	tower	eiffel	minute	good	trocadero
#6	line	amazing	structure	security	fantastic
#7	tour	early	wait	times	guide
#8	tour	booked	advance	eiffel	time
#9	visit	paris	place	view	beautiful

Nomeados de 0 a 9, os tópicos não possuem nome, mas são definidos pela distribuição de palavras correlacionados que representam um tema de destaque conforme sua estrutura. Em relação à Torre Eiffel, ponto turístico utilizado para a extração das informações, percebemos que um tópico fala sobre a paisagem, outro remete à compra de ingressos antecipados, entre outros assuntos. Portanto, em geral esse é o tipo de estrutura que pretendemos descobrir com a utilização de modelos de tópicos.

2.5.1 Métodos para modelagem de tópicos

2.5.1.1 Latent Semantic Analysis

O método Latent Semantic Analysis (LSA) é uma técnica na área de Processamento de Linguagem Natural para analisar documentos de textos de forma a representá-los vetorialmente com o objetivo de encontrar relações semânticas entre documentos e termos de maneira eficiente (DUMAIS, 2004; ALGHAMDI; ALFALQI, 2015).

Como principais etapas de funcionamento temos que, a partir de um conjunto de documentos, podemos criar uma matriz M de $n \times d$ dimensões com n termos e d documentos onde cada célula da matriz representa a frequência de um termo em um respectivo documento. Em seguida, o algoritmo Singular Value Decomposition (SVD) (KLEMA; LAUB, 1980) é utilizado para reduzir a dimensionalidade da matriz de ocorrência ao mesmo tempo que preserva as estruturas semânticas quando coloca os termos e o documentos fortemente relacionados em posições próximas na matriz.

2.5.1.2 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) é uma melhoria em relação ao LSA ao utilizar modelo generativo. É baseado em um modelo estatístico chamado modelo de aspectos. Um modelo de aspecto é um modelo de variável latente para dados co-ocorrentes que associa uma classe de dados não observados com cada observação (HOFMANN, 1999).

No PLSA, a probabilidade condicional entre documentos d e palavra w é modelada através de uma variável latente z , que pode ser identificada como uma classe ou tópico. Portanto, o modelo PLSA é parametrizado por $P(w|z)$ e $P(z|d)$, sendo que uma palavra pode pertencer a mais de um tópico e, por sua vez, um documento pode ser representado por vários tópicos (BRANTS *et al.*, 2002). Contudo, assume-se que, dada a distribuição das palavras em um tópico, $P(w|z)$ é condicionalmente independente do documento. Portanto, a probabilidade conjunta de um documento d e uma palavra w é representada como:

$$P(d, w) = P(d) \sum_z P(w|z)P(z|d) \quad (2.1)$$

Os parâmetros do modelo, $P(w|z)$ e $P(z|d)$, são estimados utilizando o algoritmo Expectation-Maximization (EM) (MOON, 1996) com o objetivo de maximizar a função de

log-verossimilhança L sobre um conjunto de documentos D :

$$L = \sum_{d \in D} \sum_{w \in d} f(d, w) \log P(d, w) \quad (2.2)$$

Onde $f(d, w)$ representa a frequência da palavra w no documento d .

2.5.1.3 Latent Dirichlet Allocation

O aparecimento do modelo Latent Dirichlet Allocation (LDA) se deu com o objetivo de melhorar o processo de captura de permutabilidade entre as palavras e os documentos, comparado aos modelos mais antigos, como LSA e PLSA. LDA é um modelo probabilístico generativo de documentos que, de maneira simples, podemos descrever como o processo de modelagem de documentos d em uma distribuição de tópicos, onde cada tópico é uma distribuição de probabilidade discreta de palavras conforme a probabilidade dessa palavra estar presente neste tópico (BLEI *et al.*, 2003).

No processo generativo do modelo, a distribuição Dirichlet é a utilizada para amostrar a distribuição de tópicos, sendo as palavras designadas para os diferentes tópicos de acordo com o resultado da amostragem. Portanto, dado um conjunto T de tópicos, podemos definir a função densidade da distribuição de Dirichlet como:

$$Dir(z, \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^K z_k^{\alpha_k - 1} \quad (2.3)$$

Onde $z = (z_1, \dots, z_k)$ é uma variável k -dimensional, $0 \leq z \leq 1$ e $\sum_{i=1}^k z_i = 1$. Temos que $\alpha = (\alpha_1, \dots, \alpha_k)$ representa os hiper-parâmetros da distribuição. A função $B(\alpha)$ acima representa a função *Beta* expressa com base na função gama Γ (ARTIN, 2015) :

$$B(\alpha) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)} \quad (2.4)$$

2.5.2 Sumarização baseada em tópicos

A sumarização baseada em tópicos representa a junção das áreas de Sumarização automática e Modelagem de tópicos, com o papel de identificar o conjunto de tópicos presentes no corpo textual de entrada e produzir um sumário para cada tópico encontrado. O objetivo

desse tipo de sumarização é otimizar o processo de leitura e identificar diferentes pontos de vista quanto aos tópicos e assuntos abordados, desde que é possível condensar diferentes opiniões e perspectivas para cada tópico. Logo, diversas abordagens têm sido exploradas no últimos anos (GANESAN *et al.*, 2010; SHIMADA *et al.*, 2011; WANG *et al.*, 2013).

Para a geração de sumários baseados em tópicos, há três etapas principais: (i) identificação dos tópicos, (ii) análise de sentimento e (iii) processo de sumarização (KIM *et al.*, 2011).

2.5.2.1 Identificação de tópicos

Esta primeira etapa se propõe a identificar os tópicos presentes no conjunto de opiniões analisadas utilizando diferentes abordagens possíveis, como:

Extração baseada em substantivos: Identifica os substantivos mais frequentes que podem representar um tópico utilizando-se de um analisador gramatical (HU; LIU, 2004). Tal processo é o mais simples e tem servido como referência para avanços na área. No entanto, esse tipo de abordagem pode sofrer de problemas no reconhecimento dos tópicos caso os substantivos, de fato, não os representem apropriadamente.

Extração baseada em aprendizado supervisionado: Utiliza-se de modelos em aprendizado de máquina para identificar uma entidade ou tópico abordado em uma opinião (PONTIKI *et al.*, 2016). Dessa forma, esse tipo de técnica exige uma grande quantidade de dados rotulados para treinamento (LIU, 2012). Como principais técnicas, temos HMM (Modelo Oculto de Markov) (RABINER, 1989) e CRF (Campos Aleatórios Condicionais) (LAFFERTY *et al.*, 2001), por exemplo.

Extração utilizando modelos de tópicos: É baseada em métodos de agrupamentos de palavras onde os tópicos são representados de acordo com as distribuições das palavras relacionadas presentes no texto. Na literatura, temos a utilização do pLSA (Probabilistic Latent Semantic Analysis) (HOFMANN, 1999) e LDA (Latent Dirichlet Allocation) (BLEI *et al.*, 2003), por exemplo.

2.5.2.2 *Análise de sentimento*

Na segunda etapa, é possível classificar o sentimento de uma opinião com as seguintes abordagens (LIU, 2012):

Classificação baseada em aprendizado supervisionado: Utiliza-se de um modelo de aprendizado supervisionado a nível de sentença para classificá-las no correto sentimento. Importante perceber que, caso seja necessário determinar a orientação de sentimento presente na opinião, heurísticas para generalizar o sentimento devem ser definidas, pois cada sentença presente pode representar sentimentos distintos.

Classificação baseada em informações léxicas: Faz uso de dicionários e outros recursos de processamento como analisadores gramaticais e árvores de dependências para determinar a polaridade de uma opinião.

2.5.2.3 *Sumarização*

Finalizando o processo de três etapas, a sumarização é executada seguindo suas características e etapas de acordo com a Seção 2.1.

2.5.3 *Avaliação da Sumarização*

A sumarização automática tem como objetivo gerar um sumário que represente múltiplos documentos a partir do processo de entendimento e extração dos assuntos relevantes. Contudo, para que se possa identificar a qualidade dos sumários resultantes, é necessário definir métricas para avaliar a qualidade das sínteses produzidas.

2.5.3.1 *Medida ROUGE*

Técnicas automáticas para avaliar a sumarização são divididas em dois tipos: “intrínsecas” ou “extrínsecas” (MANI, 1999). Na avaliação intrínseca, a qualidade do sumários gerados é avaliada diretamente mediante a comparação com sumários de referência elaborados por humanos. Já na avaliação extrínseca, verifica-se o quanto o sumário produzido pode auxiliar em outras tarefas, como classificação, análise de sentimento, etc. Pode-se dizer que a avaliação intrínseca mais largamente utilizada corresponde à família de medidas ROUGE (LIN, 2004). Essas métricas contam o número de coocorrências de unidades (n-gramas, skip-n-gramas ou

sequência de palavras) entre o sumário avaliado e a referência. As medidas que compõem o ROUGE são definidas da seguinte maneira:

ROUGE-N: é uma medida que avalia a cobertura dos n-gramas do sumário gerado sobre um conjunto de sumários de referência. Valores comuns para N são 1 (ROUGE-1) e 2 (ROUGE-2), nos quais as unidades verificadas são unigramas e bigramas, respectivamente. ROUGE-N é computado da seguinte forma:

$$ROUGE - N = \frac{\sum_{S \in S_{ref}} \sum_{gram_n \in S} count_{match}(gram_n)}{\sum_{S_i \in S_{ref}} \sum_{gram_n \in S} count(gram_n)} \quad (2.5)$$

Onde n é o tamanho da n-grama $gram_n$; S_{ref} é um conjunto de sumários S de referência; $count_{match}(gram_n)$ é o número máximo de coocorrências entre os n-gramas do sumário candidato e um sumário de referência; $count(gram_n)$ conta o número de ocorrências de $gram_n$ em S .

ROUGE-L: usa o tamanho da maior subsequência comum (LCS) (PATERSON; DANČÍK, 1994) entre as sentenças dos sumários sob comparação como unidade a ser contabilizada.

ROUGE-W: usa uma versão ponderada do LCS que favorece subsequências maiores. Por exemplo, na comparação com a referência S , se no sumário S_1 há quatro LCS de tamanho 1 e no sumário S_2 há uma LCS de tamanho 4, S_2 apresenta-se como uma escolha melhor de acordo com essa métrica.

ROUGE-S: usa skip-bigramas em vez de n-gramas. Skip-bigramas são pares de palavras ordenadas conforme a sequência que aparecem na sentença, podendo ser sequenciais (reduz-se a n-grama) ou separadas por uma distância arbitrária.

ROUGE-SU usa estatísticas sobre skip-bigramas e unigramas para medir as coocorrências.

As medidas ROUGE são avaliações de cobertura, ou seja, medem a fração de termos do sumário de referência cobertos pelo sumário gerado. Tende a favorecer a escolha de sentenças maiores para compor a síntese e não garante que o sumário gerado tenha boa legibilidade, pois este pode conter termos irrelevantes. A medida de precisão avalia o percentual de termos do sumário gerado relevantes para a referência. O cálculo da precisão é feito de modo semelhante, dividindo-se o número de sobreposição de termos pelo tamanho da sentença. Essa medida favorece a escolha de sentenças menores. Para balancear as duas métricas, costuma-se calcular a

medida F1, dada pela média harmônica entre a cobertura e a precisão (ACHANANUPARP *et al.*, 2008).

Contudo, há problemas em que não é possível obter sumários de referência para atestar a qualidade dos sumários conforme a medida ROUGE. Para tal cenário, adotamos métricas que identificam a dificuldade de leitura para avaliar a qualidade do sumário sintetizado. No tópico seguinte, detalharemos as métricas utilizadas.

2.5.3.2 Avaliação da dificuldade de leitura

Com base na sumarização automática e a falta de sumários de referência, inerente ao contexto de sumarização de opiniões em plataformas online, faz da avaliação ROUGE (LIN, 2004) inadequada para julgar a qualidade dos resultados. Diante desse cenário, a abordagem adotada é avaliar a qualidade do sumário com base na dificuldade de leitura. Pois uma exata compreensão e fácil visualização textual indicam versatilidade do conhecimento, melhores relações interdisciplinares e, no segundo caso, melhor disposição para a leitura. Em termos práticos, pretendemos avaliar os sumários obtendo um escore representando a idade e o grau de escolaridade americano (desde que a língua inglesa é utilizada) adequado para a leitura e entendimento do texto. Portanto, consideramos as seguintes métricas:

Flesch Reading Ease: A Facilidade de Leitura Flesch (FLESCH, 1948) é uma métrica de legibilidade adequada para todos os tipos de texto. Esta métrica é calculada usando o número médio de sílabas por palavra e comprimento médio da frase. O resultado da fórmula cai no intervalo de 0 a 100, o valor de 0 indica uma baixa legibilidade, enquanto que 100 indica que o texto tem uma alta legibilidade. Define-se como:

$$FLF = 206,835 - (1,015 \times CMF) - (84,6 \times MSP) \quad (2.6)$$

Onde *CMF* é comprimento médio da frase (número de palavras dividido pelo número de frases) e *MSP* o número médio de sílabas por palavra (número de sílabas dividido pelo número de palavras). Na Tabela 2 detalhamos os escores possíveis.

Tabela 2 – Conjunto de valores para a métrica Flesch.

Escore	Escolaridade
90-100	5° série
80-90	6° série
70-80	7° série
60-70	8° série
50-60	9° série
30-50	10° ano
0-30	11° ano

Fonte: Lyra e Amaral (2012).

Coleman-Liau Index: A de Coleman Liau - CL (COLEMAN; LIAU, 1975) calcula o nível de ensino necessário baseado nas médias dos comprimentos das sentenças e a média do número de caracteres por palavra. A equação abaixo define o cálculo CL:

$$CL = (5.89 \times ACW) - 0.3 \times ASL - 15.8 \quad (2.7)$$

Onde *ACW* é o número médio de caracteres por palavra e *ASL* é comprimento médio da frase (número de palavras dividido pelo número de frases). Abaixo temos a Tabela 3 de equivalência dos anos de escolaridade exigidos, valor esse resultante da métrica de Coleman Liau, e a equivalência escolar.

Tabela 3 – Conjunto de valores para a métrica ARI (Automated Readability Index).

Anos de escolaridade	Equivalência Escolar
1 a 5 anos	1°, 2° ou 3° série do ensino fundamental
6 a 8 anos	4° série do ensino fundamental
9 a 12 anos	1° e 2° grau do ensino médio
13 a 16 anos	3° grau do ensino médio ou 1°, 2°, 3° semestre do ensino superior
17 ou mais	A partir do 4° semestre do ensino superior ou mestrado e doutorado

Fonte: Lyra e Amaral (2012).

ARI (Automated Readability Index): Semelhante à métrica de Coleman-Liau, ARI também obtém a quantidade de anos na educação formal americana necessária para a compreensão de um texto. Leva-se em consideração o comprimento das sentenças e a média de caracteres por palavra para calcular seus resultados. Na Tabela 4 listamos os escores.

$$ARI = 0.50 \times ASL + 4.71 \times ACW - 21.43 \quad (2.8)$$

Semelhante a métrica anterior, onde *ACW* é o número médio de caracteres por palavra e *ASL* é comprimento médio da frase (número de palavras dividido pelo número de frases). Para os resultados possíveis, temos a Tabela 4 abaixo.

Tabela 4 – Conjunto de valores para a métrica ARI (Automated Readability Index).

Escore	Idade	Escolaridade
1	5-6	Jardim de infância
2	6-7	1° e 2° série
3	7-9	3° série
4	9-10	4° série
5	10-11	5° série
6	11-12	6° série
7	12-13	7° série
8	13-14	8° série
9	14-15	9° série
10	15-16	10° série
11	16-17	11° ano
12	17-18	12° ano
13	18-24	Universitário
14	24+	Professor

Fonte: Tabela retirada da Wikipédia (ARI, 2018)

3 TRABALHOS RELACIONADOS

Muitos estudos têm sido realizados para analisar opiniões online com técnicas de processamento de linguagem natural, aprendizado de máquina e mineração de opinião para produzir sumários automáticos (Pang and Lee, 2002; Hu e Liu, 2004; Qui et al., 2011, Zhang et al., 2011; Hu et al., 2017). Neste capítulo serão apresentados trabalhos na área de sumarização de textos, divididos por assuntos de interesse, os quais serviram de base para o desenvolvimento deste trabalho.

3.1 Sumarização automática

Com o objetivo de condensar as informações mais relevantes sobre os documentos, podemos aplicar diferentes abordagens e técnicas de aprendizado de máquina para construir um sumário relevante. Abaixo apresentamos alguns trabalhos relacionados na área de sumarização automática.

3.1.1 *SumView: A Web-based engine for summarizing product reviews and customer opinions*

Com foco na sumarização automática de texto, Wang *et al.* (2013) propuseram um sistema de sumarização baseado na *web*, extraindo automaticamente as opiniões e expressões de maior relevância de um conjunto de opiniões sobre produtos para produzir um sumário extrativo.

Primeiramente, um *crawler* foi desenvolvido para coletar os comentários dos usuários sobre produtos na plataforma da Amazon. Com o conjunto de dados em mãos, um conjunto de etapas em processamento de linguagem natural foram aplicadas: tokenização em sentenças, POS tagger e remoção de *stopwords*, por exemplo. Em seguida, é gerada a matriz de ocorrência de palavra \times sentença, onde cada linha representa uma palavra e colunas as sentenças, que será utilizado para descobrir o conjunto de características sobre produtos que são automaticamente extraídos utilizando uma abordagem baseada em Hu e Liu (2004).

Na abordagem de Hu e Liu (2004), as características de produtos são identificadas como substantivos presentes nas sentenças de várias opiniões. Assume-se que muitas pessoas usarão palavras iguais para descrever algo comum, o que convergirá para representar uma característica de produto. Contudo, no trabalho de Wang *et al.* (2013) foram adicionadas outras perspectivas, como considerar características apenas aqueles substantivos que estão

acompanhados por adjetivo.

Após tais construções, cinco características são recomendadas para o usuário do sistema escolher quais ele julga interessante para serem representadas pelas sentenças que estarão no sumário final. Ao escolhê-las, é aplicada uma fatoração de matrizes não-negativas ponderada (WANG *et al.*, 2013) com base nas características sobre as sentenças com o objetivo de agrupá-las seguindo a relevância desses aspectos. Por fim, é selecionada a sentença com maior probabilidade de cada grupo para fazer parte do sumário e representar uma característica do produto.

Podemos perceber que Wang *et al.* (2013) desenvolveram uma sumarização automática baseada em consulta, dado que o usuário da aplicação escolhe as características que deseja dar de entrada, seguindo as que foram recomendadas, ou até mesmo informar como entrada para o sistema as que deseja observar no sumário final. Importante destacar que o resultado representa uma sumarização parcial, desde que apresenta as características separadas com suas respectivas sentenças mais relevantes identificadas, diferente deste trabalho que apresenta a sumarização completa utilizando as diferentes sentenças que contemplam as características relevantes. Em relação ao reconhecimento das características dos produtos, o que podemos chamar de tópicos no contexto deste trabalho, torna-se limitada desde que se restringe à identificação apenas por substantivos frequentes, sem considerar também seu contexto semântico.

3.1.2 *Multi-document summarization using closed patterns*

No trabalho de Qiang *et al.* (2016) é proposto um modelo de sumarização multi-documentos genérico baseado em padrões, onde a sumarização é realizada utilizando o conceito de padrões fechados (*closed patterns*, em inglês), o que remete aos padrões frequentes identificados, com o intuito de extrair as sentenças mais salientes da coleção original de documentos. Com os padrões fechados, é possível identificar mais informações contextuais semânticas, diferentemente de utilizar um termo individualmente.

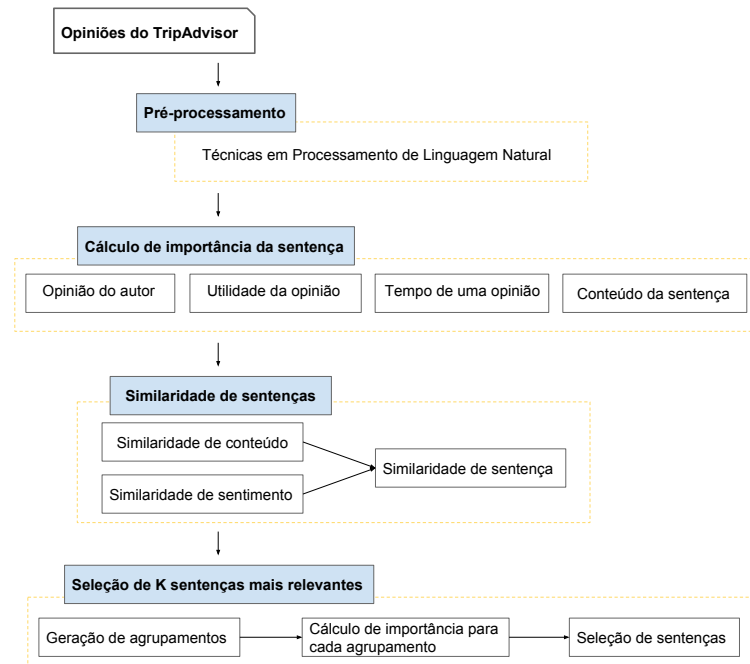
Assumindo um conjunto D de documentos, temos um conjunto $P = p_1, p_2, p_n$ de n padrões para representar todas as sentenças do conjunto D onde um padrão $p = t_1, t_2, t_m$ é uma lista ordenada de m palavras. Primeiro, para a identificação dos padrões fechados, diversos algoritmos podem ser utilizados (CloSpan (YAN *et al.*, 2003), AGraP (FLORES-GARRIDO *et al.*, 2015), SPMW (XIE *et al.*, 2014)). Em seguida, com o conjunto de padrões definidos, cada padrão conterà um conjunto de sentenças pelo qual este está presente. Como terceiro

passo, haverá um ranqueamento de sentenças, onde as sentenças mais bem pontuadas contém, por sua vez, os padrões mais relevantes. O conceito de padrão relevante se dá pelo seu peso, o qual é constituído a partir da presença de termos frequentes. Contudo, dado o resultado de ranqueamento das sentenças, o sumário é gerado levando em consideração duas características: cobertura e redundância. Para garantir tais características, foi utilizada uma abordagem baseada no algoritmo Maximal Marginal Relevance (MMR) (CARBONELL; GOLDSTEIN, 1998), onde são utilizadas medidas de similaridade considerando os padrões fechados para escolher a próxima sentença de acordo com a similaridade entre as previamente selecionadas, almejando selecionar a que apresentar a menor similaridade.

3.1.3 *Opinion mining from online hotel reviews - A text summarization approach*

O trabalho de Hu *et al.* (2017), como um trabalho recente na literatura, demonstra ser um dos mais completos no contexto de trabalho desta dissertação, devido à sua abrangência de áreas aplicadas, como mineração de opiniões e técnicas de sumarização propriamente dita. Sendo o trabalho referência desta dissertação, apresentamos a metodologia desenvolvida por Hu *et al.* (2017) na Figura 6.

Figura 6 – Metodologia de Hu *et al.* (2017)



De acordo com a metodologia, é proposta uma sumarização multi-documento onde seleciona as sentenças mais relevantes do corpo textual. Seu conjunto de documentos são

opiniões de hotéis publicadas pelos usuários na plataforma TripAdvisor, os quais passam por etapas de pré-processamento, identificação de sentimento utilizando abordagem léxica, cálculo de similaridade de sentimento, como também o cálculo de importância das sentenças. Por último, uma etapa de aprendizado não supervisionado foi aplicada para a identificação de k sentenças mais relevantes que deveriam fazer parte do sumário. Como contribuição, sua proposta considera a utilização de metadados e cria heurísticas que agregam no processo de sumarização, como a credibilidade do autor na plataforma, a data da opinião, o cálculo de relevância de uma opinião e, por fim, a identificação de opiniões conflitantes, conforme as etapas de similaridade de conteúdo e sentimento.

Contudo, sendo uma extensão aos resultados de Hu *et al.* (2017), este trabalho de dissertação pretende inserir a etapa de modelagem de tópicos para otimizar a diversidade de temas contemplados nos sumários seguindo a metodologia apresentada na Figura 6. Para mais detalhes, o Capítulo 4 apresentará as etapas desenvolvidas por Hu *et al.* (2017) e a extensão do trabalho será apresentada.

3.2 Modelagem e extração de tópicos

A modelagem de tópicos tem como intuito a extração de palavras que representam um tema ou assunto de interesse identificado em uma sentença, representando esse conjunto de palavras como um conjunto de dados não observáveis: as variáveis latentes. As variáveis latentes são inferidas com base nos dados observáveis, as palavras nesse caso, a partir da sua frequência e coocorrência. Na modelagem de tópicos, um variável latente representa o conceito de tópico que, por sua vez, representa uma distribuição de palavras que estão co-relacionadas sobre um tema em específico. Abaixo apresentamos alguns trabalhos relacionados nesta área.

3.2.1 Mining and summarizing customer reviews

O trabalho de Hu e Liu (2004) foi um dos pioneiros na sumarização de opinião em nível de tópicos, utilizando regras de associação envolvendo substantivos, diferenciando o que são tópicos explícitos e implícitos. Em seu trabalho, no entanto, o foco está em identificar os explícitos e se supõe que usuários tendem a utilizar o mesmo vocabulário para definir um tópico, enquanto que conteúdo irrelevante é geralmente diferente nos diversos documentos. Além da identificação dos tópicos, que em seu contexto representa características de produtos

vendidos em *e-commerce*, pretende-se também analisar o sentimento das sentenças identificando polaridades como positivas ou negativas sobre cada tópico. Por fim, o sumário gerado não é um texto único, mas uma listagem de sentenças de acordo com os tópicos identificados e os respectivos sentimentos envolvidos conforme polaridade negativa ou positiva. Na Figura 7, temos um exemplo de um sumário.

Figura 7 – Exemplo de sumário

Câmera digital

Característica: **qualidade de imagem**

Sentenças **positivas**: 253

- É excelente na qualidade de fotos e vídeos.
- Filma em 4k!

...

Sentenças **negativas**: 6

- As vezes fica borrada.
- O fps não é bom.

...

Característica: **tamanho**

Sentenças **positivas**: 134

- É uma câmera bem compacta.

...

Sentenças **negativas**: 10

- O tamanho dos botões são desproporcionais.

...

Fonte: Adaptado de (HU; LIU, 2004)

Contudo, mesmo que a extração baseada em substantivos frequentes seja a forma mais simples de extração de tópicos, servindo como base para métodos mais complexos, pode apresentar erros devido aos substantivos frequentes que não representam um tópico real. Além do mais, não é possível identificar os tópicos implícitos. Por fim, no âmbito da análise de sentimento, apenas os adjetivos foram utilizados para identificar as polaridades, mesmo que substantivos e advérbios pudessem apoiar na descoberta.

3.2.2 A logic programming approach to aspect extraction in opinion mining

No trabalho de Liu *et al.* (2013), foi proposta uma abordagem aplicando programação lógica para extrair aspectos dos textos. O objetivo era representar as relações sintáticas e o conhecimento de como tais relações estão envolvidas com os aspectos em forma de regras lógicas. Por exemplo, considere a seguinte relação: "se uma palavra *w*, que representa um

substantivo, é modificado por um adjetivo, então w é um aspecto". Agora temos a seguinte frase: "O telefone tem uma ótima tela". Podemos notar que, sendo a palavra "ótima" um adjetivo, a palavra "tela" representa um aspecto. Logo, podemos representar tal conhecimento como uma regra lógica, onde uma palavra w_s pode representar um aspecto se o mesmo é um substantivo e é modificado por um adjetivo w_a , conforme regra abaixo.

$$aspecto(w_s) = relacao(w_s, mod, w_a), adjetivo(w_a), substantivo(w_s) \quad (3.1)$$

Sendo a função "*relacao*" responsável por verificar se uma palavra w_s é modificada por outra palavra w_a , além de w_a ser um adjetivo e w_s um substantivo, verificados pelas funções "*adjetivo*" e "*substantivo*", respectivamente.



Portanto, percebemos que, comparada às abordagens estatísticas, podemos identificar os aspectos não apenas pela frequência dos substantivos, mas pelas relações em que as palavras possuem com outras ao redor. No entanto, o fato de depender apenas das relações sintáticas pode acarretar outro tipo de problema: algumas vezes temos substantivos que são acompanhados por adjetivos, mas não representam significado relevante. Por exemplo: "Esse telefone tem boas coisas para ajudar.". A palavra "coisas", seguindo a regra, representa um aspecto. No entanto, não oferece informação alguma. Dessa forma, a união entre as abordagens sintáticas e estatísticas poderia solucionar tais tipos de problemas.

3.2.3 *Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs*

Neste trabalho, Mei *et al.* (2007) aborda o problema de análise de tópico e sentimento sobre comentários de usuários em sites da *internet* propondo um modelo mistura probabilístico (em inglês, *mixture model*) para capturar o conjunto de tópicos e sentimentos envolvidos, de forma simultânea.

Como podemos ver na Figura 8 acima, as sentenças que correspondiam as opiniões dos usuários sobre um produto foram atribuídas a um conjunto relacionado ao tópico contemplado e à polaridade de sentimento envolvido. Portanto, o sistema primeiro extrai os tópicos envolvidos e, em seguida, associa um sentimento positivo, negativo ou neutro à sentença. Como proposta de Mei *et al.* (2007), estende-se o modelo mistura de distribuições multinomiais para envolver dois modelos de sentimento e, conseqüentemente, identificar o tópico e o sentimento envolvido simultaneamente.

Figura 8 – Exemplo de sentenças relacionadas aos respectivos tópicos e polaridade de sentimento

Query: <i>Dell Laptop</i>			
	positive	negative	neutral
Topic 1 <i>(Price)</i> 	<ul style="list-style-type: none"> • it is the best site and they show Dell coupon code as early as possible 	<ul style="list-style-type: none"> • Even though Dell's price is cheaper, we still don't want it. • 	<ul style="list-style-type: none"> • mac pro vs. dell precision: a price comparis.. • DELL is trading at \$24.66
Topic 2 <i>(Battery)</i> 	<ul style="list-style-type: none"> • One thing I really like about this Dell battery is the Express Charge feature. 	<ul style="list-style-type: none"> • my Dell battery sucks • Stupid Dell laptop battery • 	<ul style="list-style-type: none"> • i still want a free battery from dell.. •

Fonte: Imagem retirada de (MEI *et al.*, 2007).

Segundo Mei *et al.* (2006) e Mei e Zhai (2006), as palavras utilizadas em uma opinião sobre um produto qualquer na *internet* podem ser classificadas em dois grupos: comuns ou aquelas que indicam um tópico (semelhantemente ao conceito de tema ou assunto). Contudo, em Mei *et al.* (2007), são propostas subcategorias à categoria de palavras de tópicos: (i) palavras de tópicos que representam uma opinião positiva; (ii) aquelas que representam uma opinião negativa; (iii) e aquelas com representatividade neutra. Dessa forma, utilizando um modelo de tópico assumindo as componentes para cada categoria definida é possível mapear as sentenças para os respectivos conjuntos de tópico e polaridade de sentimento.

3.3 Mineração de opinião

Com o objetivo de identificar fatores implícitos em uma opinião, o sentimento, a mineração de opinião busca categorizá-las em polaridades, geralmente definidas em: positiva, negativa e neutra. Podendo ser aplicada em diferentes níveis: nível da opinião como um todo, a nível de sentença ou em nível de aspectos ou tópicos. Sendo uma subárea aplicada neste trabalho de dissertação, destacamos, resumidamente, alguns trabalhos relacionados.

O trabalho de Pang *et al.* (2002) foi o pioneiro na classificação binária de sentimentos em opiniões de filmes utilizando algoritmos de máquina supervisionados com a técnica de *Bag-of-words* e unigramas. Para tais experimentos, coletaram informações sobre os filmes e opiniões do *IMDb.com*, os quais foram pré-processados e preparados para serem aplicados em três algoritmos: NB (Naïve Bayes), ME (Maximum Entropy) e SVM (Support Vector Machine), o qual este último representou a melhor acurácia na identificação de sentimento.

Em Cataldi *et al.* (2013), o objetivo é sumarizar opiniões de produtos e serviços a nível de tópicos, onde um produto, como *smartphone*, possui características como: qualidade da tela, duração da bateria, espessura, capacidade de processamento, entre outras. Assim, a sumarização deve levar em consideração os tópicos e identificar a polaridade do sentimento intrínseco. Para tal, primeiramente foi capturado os tópicos mais relevantes e, conseqüentemente, a perspectiva do usuário na opinião. Por fim, é calculado o grau de sentimento de cada tópico em cada opinião.

Zhang *et al.* (2011) utilizaram os algoritmos Naïve Bayes e SVM para classificar os sentimentos a respeito das opiniões sobre restaurantes de uma região específica. Logo, executaram experimentos utilizando opiniões sendo 1500 com polaridade positiva e 1500 negativas e diferentes representações de atributos para treinar o algoritmo foram utilizadas, tais como unigramas e bigramas, bem como suas frequências. Como resultado, o algoritmo Naïve Bayes demonstrou a melhor acurácia, algo em torno de 95.67%, com um conjunto de 900 a 1100 atributos utilizados para treino.

Pak e Paroubek (2010) utilizaram os *emoticons* como indicador de sentimento para os *twitters* e treinaram três algoritmos supervisionados para classificar o sentimento de novos *twitters*: CRF (Conditional Random Fields), Naïve Bayes e SVM. No entanto, Já no trabalho de Davidov *et al.* (2010), além dos *emoticons*, termos como *hashtag* também foram utilizados em algoritmos não supervisionados para agrupar *twitters* semelhantes e assim descobrir suas respectivas polaridades.

No contexto das abordagens utilizando os conjuntos léxicos, o trabalho de Turney (2002) foi o primeiro a criar um algoritmo que extraía bigramas atendendo à certas regras gramaticais predefinidas, estimando suas respectivas polaridades de sentimento calculando PMI (Pointwise Mutual Information) de adjetivos e advérbios para, conseqüentemente, determinar a polaridade de toda a opinião com base na média de sentimento dos bigramas. Hu e Liu (2004) criaram um conjunto de palavras de opinião utilizando WordNet(MILLER, 1995) para predizer o sentimento das sentenças em opiniões com base na prevalência de orientação das palavras. E, posteriormente, Taboada *et al.* (2011) adicionou palavras de intensificação ou negação em relação ao conjunto de palavras anterior para identificar o sentimento de acordo com a análise da relação dessas palavras.

Por fim, Hu *et al.* (2017) desenvolveram um sumarizador multi-documentos onde aplicaram a análise de sentimentos com SOPMI (Semantic Orientation-Pointwise Mutual Infor-

mation)(TURNEY; LITTMAN, 2003) para identificar os escores de sentimentos negativos ou positivos das sentenças. Baseado no PMI (Pointwise Mutual Information) que busca identificar a associação semântica entre as palavras, ou seja, as suas relações de coocorrência em um corpo textual, o cálculo de SOPMI adiciona a orientação semântica, obtendo a relação das palavras e suas polaridades com base no conjunto de adjetivos e advérbios negativos ou positivos. Logo, identifica-se o sentimento de uma sentença baseado na média de todas as orientações semânticas das palavras.

Na Tabela 5, temos o comparativo dos trabalhos relacionados de acordo com as áreas de interesse deste trabalho de dissertação: sumarização automática, mineração de opinião e modelagem de tópicos. Logicamente, alguns trabalhos específicos abordarão apenas uma área de interesse. No entanto, poderemos perceber que nenhum trabalho abrange os três temas de interesse de maneira completa. Portanto, temos como contribuição a aplicação das três grandes áreas com o objetivo de produzir um sumário com diversidade de tópicos.

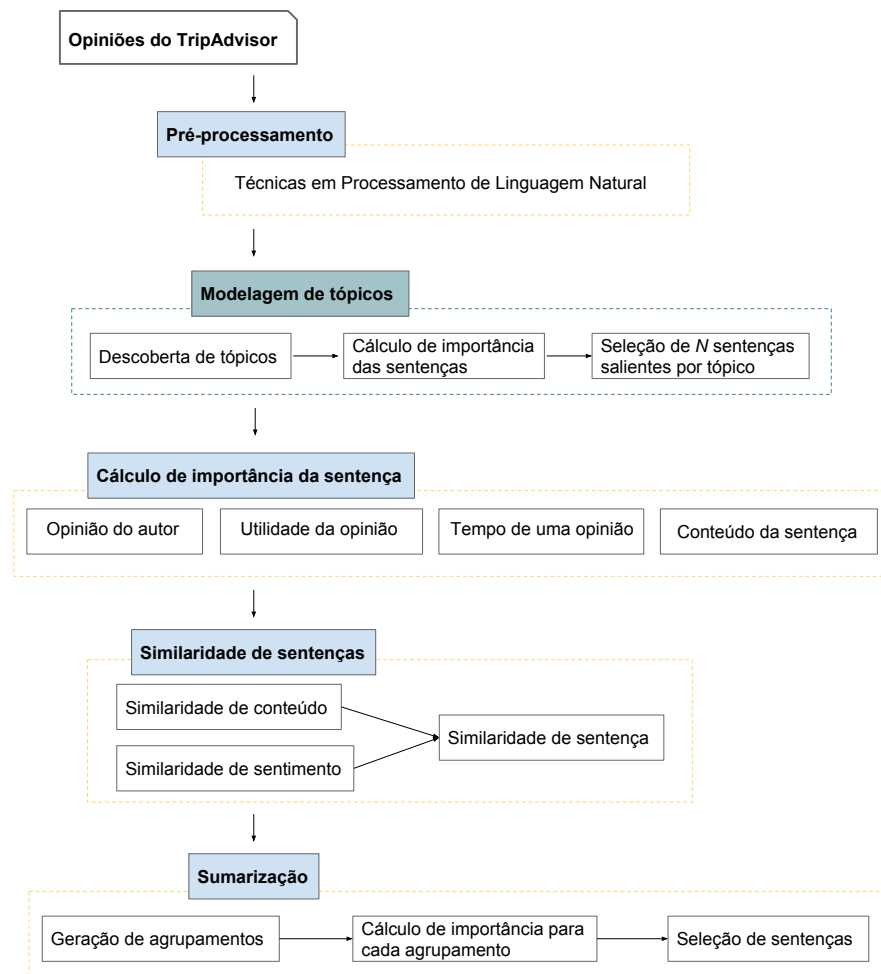
Tabela 5 – Comparativo de trabalhos relacionados.

Autor	Sumarização	Mineração de opinião	Modelagem de tópicos
SumOpinions	Sim	Sim	Sim
Hu <i>et al.</i> (2017)	Sim	Sim	Não
Qiang <i>et al.</i> (2016)	Sim	Não	Parcial
Wang <i>et al.</i> (2013)	Parcial	Não	Sim
Liu <i>et al.</i> (2013)	Não	Não	Sim
Cataldi <i>et al.</i> (2013)	Não	Sim	Sim
Zhang <i>et al.</i> (2011)	Não	Sim	Não
Taboada <i>et al.</i> (2011)	Não	Sim	Não
Pak e Paroubek (2010)	Não	Sim	Não
Davidov <i>et al.</i> (2010)	Não	Sim	Não
Mei <i>et al.</i> (2007)	Parcial	Sim	Sim
Hu e Liu (2004)	Parcial	Sim	Sim
Turney (2002)	Não	Sim	Não
Pang <i>et al.</i> (2002)	Não	Sim	Não

4 SUMOPINIONS: SUMARIZAÇÃO AUTOMÁTICA DE OPINIÕES SOBRE PONTOS TURÍSTICOS

Para o desenvolvimento deste trabalho, utilizamos como referência o trabalho de Hu *et al.* (2017), estendendo-o em relação a sua metodologia ao adicionar a etapa de "Modelagem de tópicos" para a descoberta de tópicos mais discutidos nas opiniões publicadas pelos usuários. Abaixo, temos a Figura 9 apresentando a metodologia geral.

Figura 9 – Metodologia desenvolvida



Como primeira etapa do desenvolvimento na Figura (HU *et al.*, 2017), tivemos a **coleta de opiniões** da plataforma TripAdvisor. Em seguida, executamos a fase de **pré-processamento** sobre os dados previamente coletados, onde foram aplicadas diversas técnicas em Processamento de Linguagem Natural. O intuito dessa fase é preparar os textos em uma estrutura adequada para que possa ser processada nas etapas seguintes da metodologia.

Na terceira fase, como podemos identificar em destaque na Figura 9, a etapa **Modelagem de tópicos**, representa uma nova fase adicionada à metodologia de Hu *et al.* (2017),

sendo uma das contribuições deste trabalho. Seu processo é dividido em três fases principais: (i) representar os documentos originais em tópicos; (ii) aplicar função escore de importância conforme a presença de tópicos; e (iii) o processo de ranqueamento e seleção de sentenças segundo sua importância. Logo, teremos como saída desta etapa as sentenças mais relevantes contemplando os tópicos.

Na fase seguinte, no **cálculo de importância de sentenças**, onde teremos como entrada as sentenças salientes previamente identificadas, foram aplicadas outras heurísticas para identificar a representatividade da sentença no sumário final. Por isso, leva-se em consideração diferentes os seguintes fatores: (i) a credibilidade do autor, (ii) utilidade da opinião, (iii) a data e o (iv) conteúdo da sentença em si (HU *et al.*, 2017). Importante destacar que tais heurísticas serão utilizadas durante o cálculo de importância dos agrupamentos formados pelo algoritmo não supervisionado.

Em seguida, para identificar sentenças conflitantes ou redundantes, uma função de **similaridade** foi definida com base em duas outras: **similaridade de conteúdo** e **de sentimento**. Ambas definirão o quão semelhantes duas sentenças são de maneira que a redundância possa ser minimizada e serão utilizadas em conjunto como função de distância para a formação dos agrupamentos no algoritmo final. Por fim, temos o **processo sumarização** baseado na aplicação do algoritmo não supervisionado *k-medoids*, onde as sentenças serão agrupadas e extraídas para a formação do sumário final, detalhado na Seção 4.6.

Nas seções seguintes, abordaremos cada etapa descrita acima com mais detalhes, contemplando o processo e sua importância neste trabalho.

4.1 Opiniões do TripAdvisor

Como as opiniões dos usuários sobre pontos turísticos na plataforma TripAdvisor são os dados deste trabalho, foi preciso desenvolver um coletor de opiniões¹, projeto *open-source* como uma das contribuições deste trabalho, para que pudesse obter os dados de maneira automática e em quantidade suficiente para a execução do trabalho. Detalhamos os dados coletados nas subseções abaixo.

¹ Disponibilizado em: <https://github.com/InsightLab/poi-crawler>

4.1.1 Informações sobre o autor

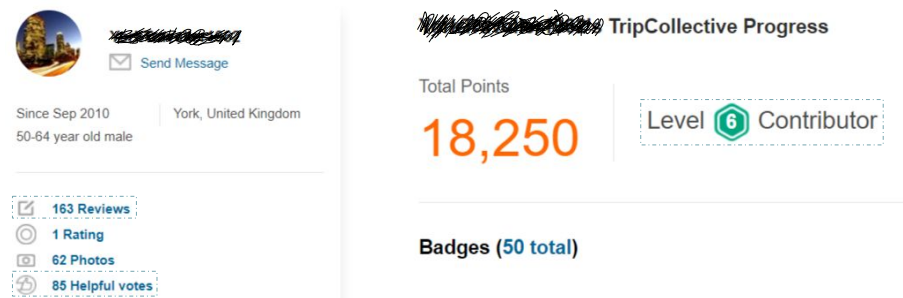
Um autor é todo usuário cadastrado na plataforma TripAdvisor e que tenha compartilhado alguma experiência de viagem na plataforma. Coletamos suas informações para serem utilizadas durante o cálculo da importância de sentenças, explicada com mais detalhes na Seção 4.4. Abaixo detalhamos as informações coletadas e na Figura 10 apresentamos um exemplo da plataforma.

Nível: Uma representação numérica da experiência do usuário na plataforma, onde maior o número, maior é a experiência do usuário utilizando e compartilhando informações na plataforma.

Quantidade de opiniões em geral: Representa a quantidade de opiniões compartilhadas por um usuário, independente do tipo, seja para hotéis ou pontos turísticos, por exemplo.

Quantidade de agradecimentos: Na plataforma é possível agradecer ao autor por uma opinião. Assim, a quantidade de agradecimentos que um usuário recebeu em relação a todas as suas opiniões são contabilizadas.

Figura 10 – Exemplo de informações coletadas sobre um autor: número de opiniões (163), agradecimentos (85) e o nível do autor (6) atribuído pela plataforma.



Fonte: TripAdvisor (2018a).

4.1.2 Informações sobre a opinião

As informações de uma opinião representam os dados mais importantes para a execução do trabalho, dado que contém o texto criado pelo autor e utilizado para a criação do sumário. Contudo, também coletamos dados agregados para o cálculo das heurísticas definidas na Seção 4.4, semelhante aos dados sobre o autor. A data de publicação da opinião, por exemplo,

é importante para ponderar a importância das sentenças. Na Figura 11, apresentamos um exemplo dos dados coletados e detalhamos os itens abaixo.

Título: Título da opinião criado pelo autor. Esse tipo de informação é importante para se ter ideia geral da opinião.

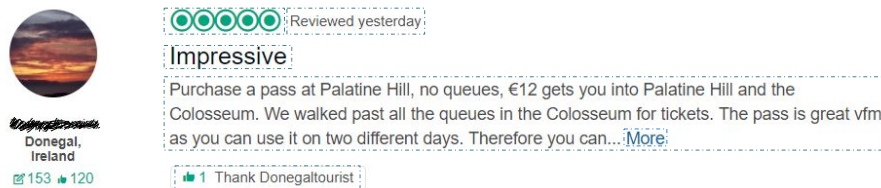
Data: Data em que a opinião foi publicada. Dessa forma, conseguimos atribuir pesos maiores às opiniões mais recentes, dando mais relevância as sentenças do texto.

Nota: A nota atribuída, pelo autor, referente à sua experiência no ponto turístico. Com essa nota conseguimos identificar se o autor está alinhado com as notas de outros autores.

Opinião: O relato de experiência do autor durante a visita no ponto turístico. Portanto, o texto propriamente dito e utilizado para a construção do sumário.

Quantidade de agradecimentos: A quantidade de agradecimentos deixados por outros usuários em relação a opinião. Com esse tipo de informação podemos qualificar a utilidade da opinião.

Figura 11 – Informações coletadas sobre a opinião de um autor para um ponto turístico.



Fonte: TripAdvisor (2018b).

4.2 Pré-processamento

Muitas das informações coletadas, conforme explanado na seção anterior, precisam passar por um processamento com o objetivo de preparar o conteúdo textual dos documentos para o processo de sumarização, representado pela Figura 12. Portanto, os documentos são processados por uma pipeline de Processamento de Linguagem Natural (NLP), incluindo as seguintes tarefas:

Segmentação de sentenças: recebe como entrada um texto e devolve as sentenças que o compõem. O trabalho é realizado por algoritmos capazes de distinguir quando símbolos de pontuação são delimitadores de sentenças (RATNAPARKHI, 1998).

Segmentação de tokens: recebe como entrada uma sentença e produz uma sequência de termos não vazios (tokens). De maneira simplificada, o texto é quebrado por espaços e os elementos resultantes geram tokens quando possuem caracteres comuns de palavras ou números, ou são

divididos em mais de um token quando iniciam ou terminam por pontuação (RATNAPARKHI, 1998).

Análise morfosintática (Part-of-speech tagging): encontra as classes gramaticais das palavras (substantivo, verbo, adjetivo, etc) e demais atributos, como gênero, número, tempo verbal, etc. É realizada por algoritmos que aprendem a rotular sequências de tokens. O processo de Part-of-speech tagging (POS tagging) utilizado foi de Porter (1980).

Tarefas comuns a área de Mineração de textos também são necessárias na sumarização, incluindo:

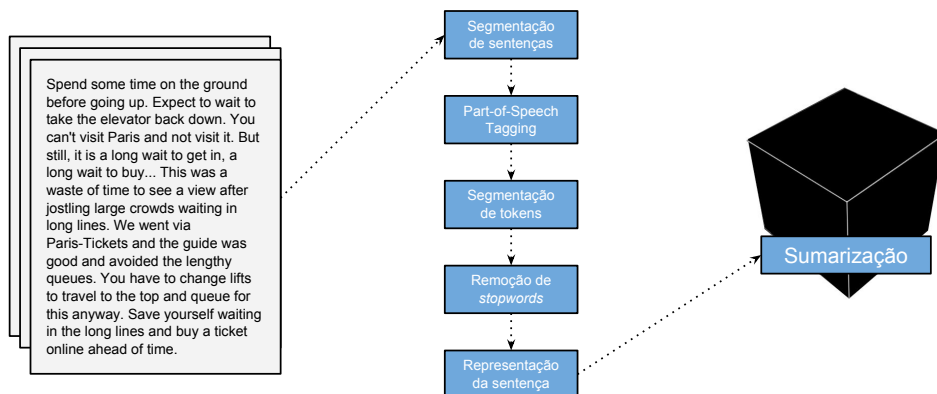
Remoção de stopwords: recebe uma sequência de tokens e remove os termos que são muito comuns (stopwords) no idioma ou no domínio. Essas palavras possuem pouco valor semântico e não são úteis para discriminar o texto. A lista de stopwords utilizadas foram as disponibilizadas pela documentação do MySQL².

Redução ao radical: recebe um termo e remove a variação gramatical por conjugação, gênero, número, etc, resultando apenas o radical.

Normalização: transforma os caracteres dos tokens para maiúsculo ou minúsculo. Outras tarefas comuns a essa etapa incluem a remoção de caracteres especiais e números.

Representação de sentença: as sentenças foram representadas apenas pelas palavras anotadas como substantivos, adjetivos e advérbios, dado que são as classes gramaticais relevantes no processo de análise de sentimento e sentido da frase.

Figura 12 – Representação da etapa de pré-processamento dos dados.



² <http://dev.mysql.com/doc/refman/5.5/en/fulltext-stopwords.html>

Na Figura 12, representamos o conjunto de processos gerais aplicados ao texto de uma opinião. No entanto, cada processo apresentado não é obrigatório em todas as etapas da metodologia, fazendo-se uso apenas de um subconjunto das tarefas conforme a necessidade de cada etapa.

4.3 Modelagem e descoberta de tópicos

Como extensão ao trabalho de Hu *et al.* (2017) e uma das contribuições deste trabalho, esta fase tem como objetivo extrair os tópicos mais importantes presentes nas opiniões sobre os pontos turísticos e calcular a importância de uma sentença em relação aos tópicos descobertos. Portanto, esta seção contempla a aplicação do algoritmo LDA (Latent Dirichlet Allocation) (BLEI *et al.*, 2003) para a descoberta e extração de tópicos presentes no *corpus*, bem como a aplicação da função *score* de importância e a seleção das sentenças principais.

Para a aplicação do LDA, há bibliotecas como *Gensim*³ e *Scikit-learn*⁴ que implementam o algoritmo. Para o contexto deste trabalho, *Scikit-learn* foi a biblioteca escolhida devido à sua maturidade e documentação disponibilizada. No entanto, ambas as bibliotecas possuem referências de implementação semelhantes, baseadas no trabalho de Hoffman *et al.* (2010)⁵.

Para executar o LDA, foi aplicado um processo de transformação das opiniões de sua forma textual para uma representação em vetor de frequências utilizando *TF-IDF* (Term Frequency-Inverse Document Frequency) (RAMOS *et al.*, 2003). Além disso, foram definidos quatro parâmetros principais como entrada ao algoritmo: o número de tópicos que devem ser extraídos, o decaimento da taxa de aprendizagem, o balanceamento que pondera as iterações iniciais e o número de iterações do algoritmo sobre os documentos originais.

Logo, com os hiper-parâmetros definidos (para informações sobre os valores, descrevemos o processo no Capítulo 5), treinamos o modelo com as opiniões mais recentes para descobrir a distribuição de tópicos. Em seguida, obtendo a estrutura dos tópicos, foram selecionadas as cinco palavras mais frequentes de cada tópico como forma de representá-los, servindo de entrada para a função de *score* que calcula a importância da sentença com base na presença de tópicos. Seja o conjunto de n palavras de uma sentença s , $W_s = \{w_1, w_2, \dots, w_n\}$, e o conjunto de cinco palavras representantes de um tópico t , $W_t = \{w_1, w_2, \dots, w_5\}$, foi verificada a intersecção de palavras entre os dois conjuntos como uma função de *score*:

³ <https://radimrehurek.com/gensim/models/ldamodel.html>

⁴ <http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html>

⁵ O código utilizado está disponível em: https://github.com/blei-lab/online_lda

Contagem de palavras: seja $I_t = |W_s \cap W_t|$, I_t , o resultado representa quantas palavras estão presentes no conjunto W_s e W_t , indicando quanto do conteúdo de uma sentença s contém sobre um tópico t .

Importante destacar que a contagem não leva em consideração o quão relevante é cada palavra de um tópico para uma sentença, apenas se o termo está presente ou não, pois devido ao tamanho das sentenças, a frequência de cada palavra seria baixa. Do contrário, sentenças maiores poderiam ser privilegiadas, desde que há maior probabilidade de repetição de palavras sobre um tópico. No entanto, desde de que a variação de tamanho das sentenças não deve ser considerada para contabilizar a presença de tópicos, a identificação do conjunto intersecção de palavras se mostrou a mais adequada, sem necessitar de outras abordagens de similaridade como Jaccard (NIWATTANAKUL *et al.*, 2013) ou similaridade Cosseno (HUANG, 2008). Podemos então representar a função escore no Algoritmo 1 abaixo:

Algoritmo 1: Função de escore para sentenças com base no tópico

Entrada: W_s ; W_t

Saída: Uma tupla em forma de (sentença, escore)

início

 contaPalavras = 0;

para cada $w \in W_t$ **faça**

Se $w \in W_s$ **então**

 contaPalavras += 1;

fim

fim

fim

Retorna (W_s , contaPalavras)

Assumindo que, no pior caso, o algoritmo execute cinco iterações no laço, uma para cada palavra do vetor de tópicos, e 18 verificações no vetor de palavras, sendo o tamanho máximo de 18 termos em uma sentença conforme o Algoritmo 2, temos uma complexidade $O(1)$.

No entanto, antes que o algoritmo 1 pudesse ser aplicado, as sentenças deveriam ser filtradas para eliminar as que estão desestruturadas ou não representam relevância. Pois é importante destacar que podemos encontrar erros de pontuações na escrita do usuário, onde as sentenças não estarão bem definidas. Do contrário, poderemos ter sentenças curtas, sendo irrelevantes para a compreensão do sumário final. Portanto, definiu-se o Algoritmo 2 de filtragem

abaixo.

Algoritmo 2: Filtragem de sentenças

Entrada: Conjunto universo de sentenças de todos os documentos $U_{s,d}$

Saída: Matriz de sentenças filtradas $U_{w,s}$

início

$U_{w,s} = []$;

para cada $s \in U_{s,d}$ **faça**

$W_s = \text{tokeniza}(s)$;

Se $4 \leq \text{tamanho}(W_s) \leq 18$ **então**

$U_{w,s} += W_s$

fim

fim

fim

Retorna $U_{w,s}$

Por simplicidade, assumindo que a função de tokenização divide a sentença em palavras ao encontrar espaços em branco, possuirá complexidade $O(m)$, o qual representa o custo de separar m palavras da uma sentença. Assim, conclui-se que o algoritmo 2 possui complexidade $O(n * m)$, onde n representa a quantidade de sentenças.

Dessa forma, para cada iteração de uma sentença, aplicamos o processo de tokenização para capturar o vetor de termos e verificamos se a mesma possui a estrutura adequada, ou seja, um tamanho mínimo de 4 e máximo de 18 palavras. Para o valor mínimo de 4, como o objetivo é identificar sentenças com relevância em termos de tópicos, observou-se que sentenças com menos de quatro termos não seria suficiente para expressar uma ideia e resultasse em conteúdo relevante para o sumário. Por outro lado, 5 é o número de palavras representantes de um tópico, logo foi escolhido um número menor que cinco e que tivesse um mínimo de estrutura para representar um conteúdo. Contudo, temos sentenças muito grandes, seja por tentar expressar muitas ideias em uma só sentença ou mesmo por ter erros de escrita por parte do usuário, onde as pontuações não foram corretamente utilizadas. Portanto, o valor máximo de 18 por representar um número próximo à média de palavras por sentenças. Por fim, a matriz de palavras cuja sentenças obedecem aos critérios é retornada para que possa ser aplicado o Algoritmo 1 de relevância ao contexto de tópicos.

Logo, calculado o escore de acordo com a presença dos tópicos abordados, as sentenças foram ranqueadas em ordem decrescente e selecionadas as 30 primeiras sentenças mais relevantes de cada tópico.

4.4 Cálculo de importância da sentença

Após calcular o escore das sentenças conforme os tópicos descobertos, outro conjunto de características foram utilizadas para identificar o escore de importância de uma sentença, conforme o trabalho de Hu *et al.* (2017), como: i) a representatividade do autor; ii) importância da opinião; iii) o quão recente é a opinião; e o iv) conteúdo das sentenças em uma opinião.

Essa etapa tem influência direta na seleção final das sentenças para a formação do sumário, desde que o escore de importância será utilizado para calcular a relevância dos agrupamentos formados, detalhado na Seção 4.6. Portanto, seguem as definições das funções para calcular o escore de importância das sentenças.

Definição 4.4.1 (Credibilidade do autor) *A credibilidade de um autor a representa o quão próxima a nota deixada por ele se encontra em relação às outras opiniões, onde estar perto da média é ser mais confiável. Assim, baseada no erro absoluto médio entre as notas deixadas pelo autor a e a média geral de um hotel h (objeto de estudo no trabalho de Hu *et al.* (2017)), podemos definir a credibilidade como:*

$$AC_a = 1 - \left(\frac{\sum_{h=1}^{H_a} \frac{|r_h^a - ar_h|}{5}}{H_a} \right) \quad (4.1)$$

Onde r_h^a representa a pontuação de um hotel h dada pelo autor a , ar_h a pontuação média de um hotel h e H_a o número de relatos de experiência. Destaco que o denominador de valor 5, considerado no somatório, representa a pontuação máxima que um hotel pode receber na plataforma TripAdvisor.

Definição 4.4.2 (Escore de recomendação do autor) *Considere arn_a a média de recomendações de um autor a , define-se o escore de recomendação ARS_a para este autor como:*

$$ARS_a = \begin{cases} 1, & \text{se } \frac{\log_2(arn_a+1)}{2} \geq 1 \\ \frac{\log_2(arn_a+1)}{2}, & \text{caso contrário} \end{cases} \quad (4.2)$$

Observa-se que sendo $arn_a \geq 3$, o escore do autor a será igual a 1, o que se pode interpretar como confiáveis as opiniões deixadas pelo usuário a .

Definição 4.4.3 (Representatividade de um autor) Com AC_a e ARS_a representando a credibilidade e o escore de recomendação de um autor a , respectivamente, podemos definir a representatividade do autor como:

$$RCA_a = \frac{(AC_a + ARS_a)}{2} \quad (4.3)$$

Portanto, a sua representatividade será a média entre a credibilidade e o escore de recomendação. Assim, poderemos dar um peso maior às opiniões em que o autor tem boa representatividade.

Definição 4.4.4 (Utilidade da opinião) Assumindo que crn_i denota o número de recomendações para uma opinião i e $\max(crn)$ o número máximo de recomendações entre todas as opiniões, a utilidade da opinião define-se como:

$$CH_i = \frac{crn_i}{\max(crn)} \quad (4.4)$$

Definição 4.4.5 (Recência da opinião) Define-se t como o intervalo de tempo entre a data de publicação da opinião e a data da consulta. Em relação à dm , representa o intervalo de tempo entre a data da primeira e última opinião. Portanto, a recência da opinião apresenta-se como:

$$CR_i = \exp\left(\frac{-t}{dm}\right) \quad (4.5)$$

Saber o quão recente é uma opinião ajuda a dar mais relevância para as opiniões que foram publicadas próximas ao período de consulta.

Definição 4.4.6 (Escore de sentença) Para calcular o escore de uma sentença j , CSS_j é formada em conjunto com outras três funções: i) $LOC(s_j)$, sendo igual a 1, se a sentença s_j é o título ou a primeira sentença da opinião, ou 0, caso contrário; ii) $IP(s_j)$ igual a 1 se s_j contém alguma palavra/frase indicadora (olhar Figura 13 abaixo), senão igual a 0; iii) $NW(s_j)$ representa a razão entre o número de palavras presentes em s_j e o número máximo de palavras entre todas as sentenças da opinião. Importante citar que, para o conjunto de palavras/frases indicadoras, o artigo utiliza-se de um conjunto previamente definido na literatura.

Figura 13 – Palavras/frases indicadoras utilizadas no artigo.

however	nevertheless	though	goal	summaries
although	summary	results	as a result	conclusion
invention	even though	intent	intention	discussion
conclusions	even if	purpose	in summary	all in all
discussions	objective	finally	Not with standing	result

Portanto, temos que o escore de uma sentença j é definido como:

$$CSS_j = w_1 \times LOC(s_j) + w_2 \times IP(s_j) + w_3 \times NW(s_j) \quad (4.6)$$

Contudo, o conjunto de pesos w_1, w_2 e w_3 representam os pesos para a posição de uma sentença no texto, a palavra/frase indicadora e o número de palavras presentes na sentença, respectivamente.

Definição 4.4.7 (Importância da sentença) Por fim, dada uma opinião i escrita por um autor a , a importância de uma sentença s_j representada como $SI_{a,i,j}$ é:

$$SI_{a,i,j} = \frac{RCA_a + CH_i + CR_i}{3} \times CSS_j \quad (4.7)$$

Sendo a função de importância da sentença a relação de resultados de todas as outras funções de escore de importância, ela será utilizada para calcular a importância dos agrupamentos de sentenças formados no processo de sumarização utilizando um algoritmo não supervisionado. Portanto, essa frase não possui uma saída de resultados, mas representa um processo intermediário na construção do sumário, detalhado na Seção 2.5.2.

4.5 Similaridade de sentenças

Semelhante ao cálculo de importância da sentença, descrito na seção anterior, tem como entrada as sentenças que estão no processo de sumarização, durante a fase de agrupamento das sentenças. Consequentemente, tal similaridade de sentença é a função de distância utilizada no algoritmo não supervisionado. Contudo, ela é dividida em duas outras similaridades: de sentimento e de conteúdo.

Para o cálculo de similaridade de conteúdo, foi necessário calcular a dissimilaridade entre dois termos utilizando *Normalized Google Distance* (NGD) (CILIBRASI; VITANYI, 2007), a qual identifica a distância semântica entre pares de termos.

Definição 4.5.1 (Normalized Google Distance) *Portanto, para duas sentenças s_j e s_k , primeiramente são extraídos todos os pares de substantivos possíveis entre as duas sentenças para, em seguida, calcular a similaridade de cada par, como segue:*

$$SIM_{ngd}(n_x, n_y) = 1 - NGD(n_x, n_y) \quad (4.8)$$

Após o cálculo de similaridade de cada par de substantivo, é possível calcular a similaridade de conteúdo entre duas sentenças.

Definição 4.5.2 (Similaridade de conteúdo) *Assumindo que $m \times n$ representa o produto cartesiano entre os conjuntos de substantivos de duas sentenças e a função $count_\beta$ retorna quantos pares satisfazem o limite inferior de similaridade $\beta = 0.65$, temos a similaridade de conteúdo como:*

$$ContentSim(s_j, s_k) = \frac{count_\beta(s_j, s_k)}{m \times n} \quad (4.9)$$

Já na similaridade de sentimento é utilizado o método SOPMI (*Semantic Orientation-Pointwise Mutual Information*) (TURNEY; LITTMAN, 2003) para calcular a influência dos adjetivos no escore de sentimento.

Definição 4.5.3 (Sentimento de uma sentença) *Contudo, para calcular o escore de sentimento de uma sentença, definimos:*

$$O(s_j) = \frac{\sum O(a_x)}{\text{\#número de adjetivos em } s_j} \quad (4.10)$$

Definição 4.5.4 (Influência de adjetivos na análise de sentimento) *Sendo $O(a_x)$ a função que retorna um escore indicando a influência que um adjetivo a_x tem na análise de sentimento, con-*

forme a relação com os adjetivos positivos A_{post} e negativos A_{neg} , temos

$$O(a_x) = \sum_{t_l \in A_{pos}} SOPMI(a_x, t_l) - \sum_{t_l \in A_{neg}} SOPMI(a_x, t_l) \quad (4.11)$$

Definição 4.5.5 (Polaridade de sentimento) Contudo, baseado na equação $O(s_j)$ e assumindo r uma constante de valor 0,2 (conforme Hu et al. (2017)), podemos calcular a polaridade da sentença da seguinte maneira:

$$SP(s_j) = \begin{cases} 1, se O(s_j) > r \\ 0.5, se |O(s_j)| \leq r \\ 0, se O(s_j) < -r \end{cases} \quad (4.12)$$

Definição 4.5.6 (Similaridade de sentimento) Com a polaridade das sentenças definidas, agora é possível calcular a similaridade de sentimento entre duas sentenças como:

$$SentiSim(SP(s_j), SP(s_k)) = \begin{cases} 1, se SP(s_j) = SP(s_k) \\ 0.5, se SP(s_j) = 0.5 ou SP(s_k) = 0.5 \\ 0, se |SP(s_j) - SP(s_k)| = 1 \end{cases} \quad (4.13)$$

Definição 4.5.7 (Similaridade de sentenças) Por fim, conclui-se que a similaridade entre duas sentenças, s_j e s_k , é a relação entre a similaridade de conteúdo e a similaridade de sentimento:

$$SenSim(s_j, s_k) = ContSim(s_j, s_k) \times SentiSim(s_j, s_k) \quad (4.14)$$

4.6 Sumarização

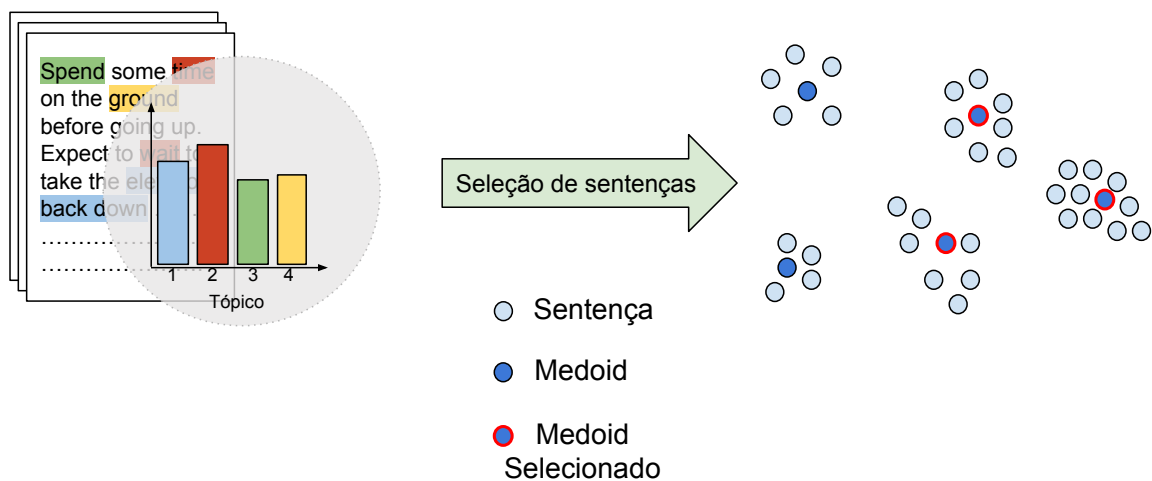
Em última fase, tendo como entrada o conjunto de sentenças representativas de cada tópico, de acordo com a Seção 4.3, o sumário de opiniões será construído utilizando-se de uma abordagem não supervisionada, o algoritmo k -medoids (KAUFMAN; ROUSSEEUW, 1987). Assim, semelhante ao processo aplicado no trabalho de Hu et al. (2017), P agrupamentos de

sentenças serão criados, onde o fator de distância entre as sentenças será o cálculo de similaridade apresentado na Seção 4.5.

Em seguida, após a formação dos agrupamentos, K sentenças serão selecionadas, sendo $K < P$, representando os medoids dos agrupamentos mais relevantes. Contudo, para que fossem identificados os K medoids mais relevantes, foi aplicado o cálculo de importância de sentenças, discutido na Seção 4.4, em cada agrupamento e, por fim, somadas as importâncias para descobrir a relevância do grupo como um todo.

Desse modo, as sentenças extraídas dos agrupamentos formarão o sumário final e, conseqüentemente, promover uma seleção de sentenças com base na extração dos tópicos tende a proporcionar sumários que contemplem um conjunto maior de temas, conforme Figura 14. Sendo o objetivo deste trabalho maximizar a diversidade de tópicos, os resultados serão demonstrados no Capítulo 5 de resultados.

Figura 14 – Processo de sumarização com seleção das sentenças representativas de K medóides.



5 EXPERIMENTOS

Para a realização dos experimentos foram coletadas opiniões sobre dois pontos turísticos: Torre Eiffel (Paris, França) e Parque Central (Nova Iorque, Estados Unidos). Tais pontos de interesse foram escolhidos devido à sua importância no turismo mundial e a grande disponibilidade de opiniões no TripAdvisor. Os resultados refletem a aplicação das três grandes áreas de interesse, onde temos a sumarização automática com mineração de opinião e a modelagem de tópicos com o objetivo de maximizar a relevância do sumário final com base na diversidade de temas contemplados. Nas seções seguintes, apresentaremos o conjunto de dados utilizados e as abordagens de avaliação para demonstrar os resultados. Importante destacar que desenvolvemos o sistema de sumarização do trabalho de Hu *et al.* (2017), seguindo todos os processos apresentados. Em seguida, ampliamos seu trabalho utilizando modelagem de tópicos para atingir os nossos objetivos. Portanto, os resultados abaixo refletem uma comparação entre os resultados desta dissertação e o trabalho de Hu *et al.* (2017).

5.1 Dataset

Na Tabela 6, apresentamos algumas informações sobre o conjunto de dados utilizados: opiniões, quantidade de autores, sentenças e a média de sentenças por opinião.

Tabela 6 – Informações das opiniões coletadas da plataforma TripAdvisor.

Ponto de Interesse	Opiniões	Autores	Sentenças	Vocabulário	Média de sentenças
Torre Eiffel	27.881	27.069	106.310	24.050	3.91
Parque Central	38.923	23.618	139.332	30.248	3.72

5.2 Configuração dos experimentos

Segundo o trabalho de Hu *et al.* (2017), a etapa de similaridade de conteúdo utiliza a função *NGD*, o que exige a quantidade de *hits* de palavras em plataformas de busca, como *Google Search* ou *Bing*. Portanto, mesmo sendo desenvolvido um coletor, há diversas restrições enfrentadas na coleta, como o bloqueio de acesso ou a limitação de coletas paralelas, o que torna o processo demorado. Portanto, como a quantidade de palavras do vocabulário é grande em cada conjunto de dados, foram selecionadas apenas as 200 opiniões mais recentes de cada

ponto turístico com o objetivo de diminuir o vocabulário para o cálculo do *NGD*, mas que ainda sejam representativas para a geração de sumários. Este processo, por exemplo, resultou em 718 sentenças e um vocabulário de 1.745 palavras para os dados da Torre Eiffel.

Em seguida, para que os resultados pudessem ser comparados, os experimentos foram executados sobre duas abordagens de sumarização: sumários baseado no trabalho de Hu *et al.* (2017) e outro conjunto de sumários conforme a metodologia deste trabalho, apresentada no Capítulo 4. Em ambas as abordagens, foram definidas três configurações de geração: i) sumários com 10 sentenças; ii) sumários com 20 sentenças; e iii) sumários com 30 sentenças. Para todas as configurações, 100 sumários são gerados de maneira independente, sem qualquer influência entre os processos de geração.

5.3 Resultados

Para os 100 sumários gerados de cada abordagem, aplicamos três etapas de avaliação: a diversidade de tópicos, análise de redundância e a avaliação da dificuldade de leitura. Os resultados em diversidade de tópicos buscam avaliar a cobertura dos sumários em relação aos tópicos descobertos, ou seja, têm o objetivo de identificar os resultados do sumário em relação ao que contempla dos temas principais sobre os pontos turísticos. Para a análise de redundância, verificamos qual abordagem tende a gerar sumários menos redundantes, levando em consideração a similaridade das sentenças que formam um sumário. Portanto, o intuito é verificar o quão repetitivo o sumário é em termos de sentenças e tópicos. Por fim, analisamos os resultados também com base em métricas de qualidade de leitura, identificando o grau de escolaridade necessário para uma leitura proveitosa do sumário.

5.3.1 *Diversidade de tópicos*

Como objetivo principal do trabalho, pretende-se gerar sumários que apresentem uma maior diversidade de tópicos. Ou seja, maximizar a relevância do sumário com base na proporção de informações relevantes presentes, os tópicos. Portanto, os experimentos abaixo demonstram como a diversidade dos tópicos foi maximizada.

Para descobrir os tópicos presentes no *corpus* de cada ponto turístico, aplicamos a etapa de extração de tópicos nas 200 opiniões selecionadas. Contudo, para aplicar o algoritmo de identificação, foi preciso encontrar os seguintes hiper-parâmetros do modelo LDA: o número de

tópicos que devem ser extraídos, o decaimento da taxa de aprendizagem, o balanceamento que pondera as iterações iniciais e o número de iterações do algoritmo sobre os documentos originais.

Abaixo, segue o conjunto de valores testados para cada ponto turístico:

- **Número de tópicos:** 5, 7, 10 e 15;
- **Taxa de decaimento no aprendizado:** 0.5, 0.7 e 0.9;
- **Valor de ponderação:** 10, 30 e 50;
- **Número de iterações:** 50, 100 e 150;

O processo de *GridSearch* consiste em executar o algoritmo LDA com todas as combinações possíveis de parâmetros. Em seus resultados, os valores mais apropriados para ambos os pontos turísticos foram 5, 0.5, 10 e 150 para o número de tópicos, taxa de decaimento, balanceamento e número de iterações, respectivamente. Contudo, o parâmetro número de tópicos foi ampliado para 10 no caso da Torre Eiffel e 7 tópicos para o Central Parque, possibilitando uma maior variedade de tópicos descobertos sem implicar em diferenças significativas nas métricas de qualidade utilizadas¹ em relação aos resultados de 5 tópicos.

Importante destacar que a decisão de 10 tópicos para a Torre Eiffel e 7 para o Parque Central é baseado em experimentos e, logicamente, depende muito do conjunto de dados. Como temos um cenário muito dinâmico de tópicos abordados pelos usuários em cada ponto turístico, temos essa diferença. Portanto, os números escolhidos foram de acordo com os resultados de experimentos e análise dos tópicos gerados, buscando aumentar a diversidade dos tópicos de forma que os tópicos não tivessem muita intersecção de assuntos. Como resultados das extrações, temos a Tabela 7 para a Torre Eiffel e a Tabela 8 sobre o Parque Central.

¹ <http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html#sklearn.decomposition.LatentDirichletAllocation.score> e <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html#sklearn.decomposition.LatentDirichletAllocation.perplexity>

Tabela 7 – Tópicos extraídos sobre a Torre Eiffel

Tópico/Palavras	1	2	3	4	5
0	queue	tickets	long	queues	booked
1	paris	tower	night	eiffel	day
2	tower	eiffel	tickets	paris	views
3	visit	tower	loved	place	queue
4	view	long	people	lines	enjoy
5	tower	eiffel	minute	good	trocadero
6	line	amazing	structure	security	fantastic
7	tour	early	wait	times	guide
8	tour	booked	advance	eiffel	time
9	visit	paris	place	view	beautiful

Tabela 8 – Tópicos extraídos sobre o Parque Central

Tópico/Palavras	1	2	3	4	5
0	park	new	central	york	tour
1	visit	day	place	just	bit
2	walk	park	great	place	nice
3	taking	visit	carriage	ice	wonderful
4	park	central	bike	ride	bikes
5	park	beautiful	central	ice	zoo
6	park	city	people	just	great

Essa estrutura de tópicos se deve ao resultado do algoritmo LDA (Latent Dirichlet Allocation) utilizado, segundo a Seção 4.3. Como explanado, um tópico não possui um nome ou estrutura, mas é representado por uma distribuição de termos presentes no corpo textual. Portanto, acima temos as cinco palavras mais frequentes de cada tópico extraído de acordo com o ponto turístico de interesse.

Seguindo esse conjunto de tópicos, um processo de identificação foi aplicado em todos os sumários para quantificar os tópicos abordados. A ideia do processo consiste em contabilizar a quantidade de termos de cada tópico presentes nas sentenças do sumário, exigindo uma quantidade mínima de três palavras para um tópico ser considerado. Por exemplo, se uma sentença de um sumário sobre a Torre Eiffel, ao ser *tokenizado* em palavras, apresentar os termos "tower", "eiffel" e "trocadero", então ele aborda o Tópico 5 da Tabela 7. Portanto, apresentamos os resultados dos experimentos abaixo sobre a cobertura de tópicos dos 10 primeiros sumários

na Tabela 9 e na Tabela 10 para Torre Eiffel e Parque Central, respectivamente:

Tabela 9 – Cobertura de tópicos para os 10 primeiros sumários sobre a Torre Eiffel.

Trabalho	Configuração	Sumário									
		1	2	3	4	5	6	7	8	9	10
Hu et al. 2017	10 sentenças	0	0	2	0	1	0	0	1	0	0
SumOpinions		1	3	2	2	2	2	2	2	4	3
Hu et al. 2017	20 sentenças	5	3	2	2	2	2	4	1	1	2
SumOpinions		5	3	4	4	5	4	4	5	4	4
Hu et al. 2017	30 sentenças	3	6	5	5	4	3	2	4	5	3
SumOpinions		3	5	7	7	5	5	5	6	5	5

Tabela 10 – Cobertura de tópicos para os 10 primeiros sumários sobre o Parque Central.

Trabalho	Configuração	Sumário									
		1	2	3	4	5	6	7	8	9	10
Hu et al. 2017	10 sentenças	0	0	0	0	0	1	2	0	1	0
SumOpinions		2	2	1	1	1	1	2	1	1	0
Hu et al. 2017	20 sentenças	0	1	3	1	1	1	1	1	1	0
SumOpinions		2	2	3	1	1	2	2	3	2	1
Hu et al. 2017	30 sentenças	1	1	1	4	2	3	1	4	2	1
SumOpinions		2	1	4	2	2	3	2	3	2	4

Como podemos observar nos resultados acima nas Tabelas 9 e 10, os sumários gerados a partir do processo proposto neste trabalho abrange, em geral, uma quantidade consideravelmente maior de tópicos para os 10 primeiros sumários. Contudo, esse comportamento se mantém nos 100 sumários gerados, como podemos ver nas Figuras 15 e 16 sobre a Torre Eiffel, onde o número do sumário gerado é apresentado no eixo x e o número de tópicos contemplados pelos sumários no eixo y .

Figura 15 – Comportamento da diversidade de tópicos sobre os 100 sumários gerados com 20 sentenças sobre a Torre Eiffel.

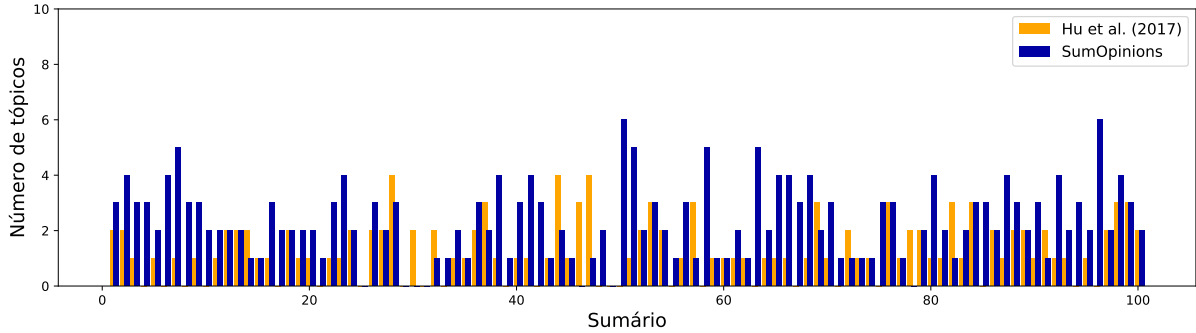
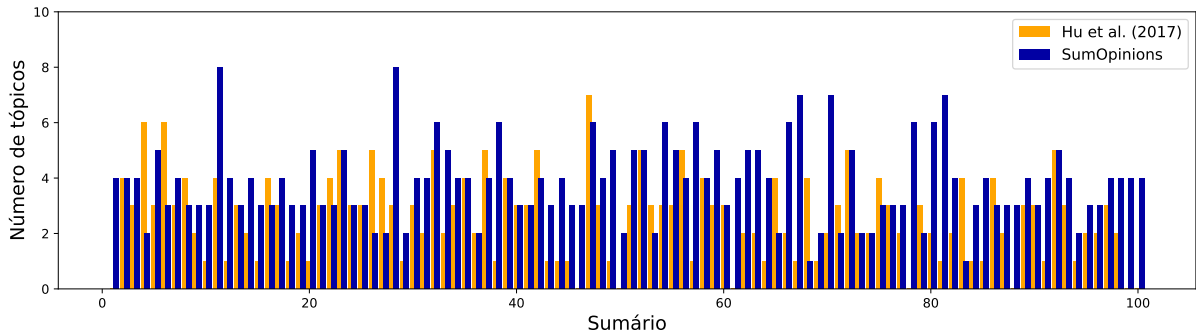


Figura 16 – Comportamento da diversidade de tópicos sobre os 100 sumários gerados com 30 sentenças sobre a Torre Eiffel.



Semelhantemente, apresentamos os resultados sobre o Parque Central nas Figuras 17 e 18. Podemos identificar resultados semelhantes aos identificados sobre a Torre Eiffel.

Figura 17 – Comportamento da diversidade de tópicos sobre os 100 sumários gerados com 20 sentenças sobre o Parque Central.

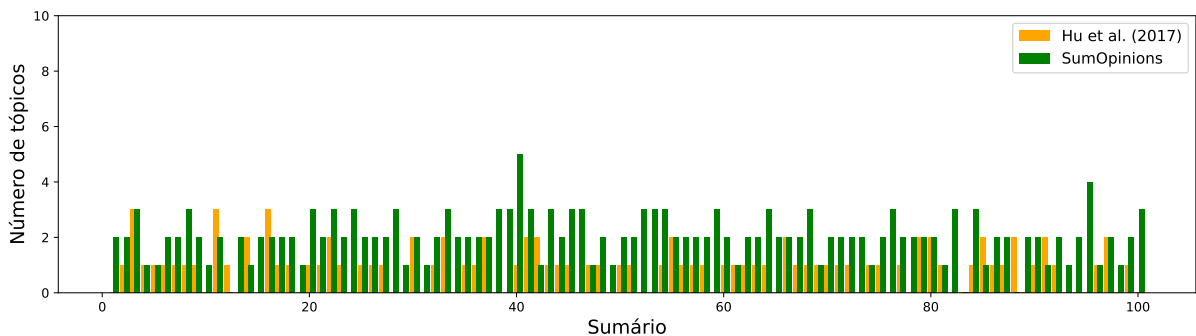
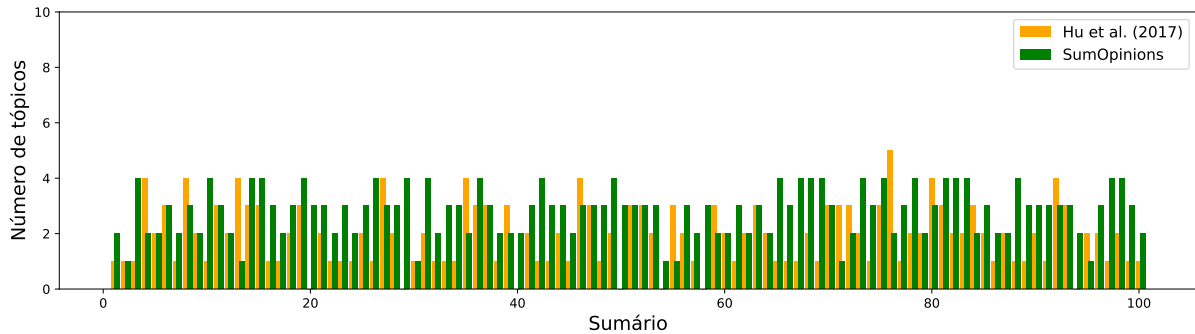


Figura 18 – Comportamento da diversidade de tópicos sobre os 100 sumários gerados com 30 sentenças sobre o Parque Central.



Os resultados sobre o comportamento da diversidade de tópicos nos sumários com 10 sentenças não foram apresentados por representar menor relevância diante dos resultados. No entanto, apresenta o mesmo comportamento. Em suma, podemos demonstrar os resultados estatísticos com a Tabela 11 abaixo:

Tabela 11 – Resultados sobre a cobertura de tópicos nos sumários

Trabalho	Configuração	Média		Desvio Padrão		Moda	
		Torre Eiffel	Parque Central	Torre Eiffel	Parque Central	Torre Eiffel	Parque Central
Hu et al. 2017	10 sentenças	0.65	0.32	0.817	0.56	0	0
SumOpinions		2.17	1.06	1.058	0.69	2	1
Hu et al. 2017	20 sentenças	2.40	0.92	1.208	0.74	2	1
SumOpinions		4.24	2.03	1.040	0.82	4	2
Hu et al. 2017	30 sentenças	4.05	1.91	1.472	1.12	4	1
SumOpinions		5.55	2.82	1.177	0.88	5	3

Destacados os melhores resultados, podemos notar que o SumOpinions atinge os melhores valores em relação à cobertura de tópicos com números relevantes na média e na moda. No caso do desvio padrão, que reflete o comportamento de geração dos sumários, temos que, com exceção do caso da configuração de 10 sentenças, apresentando baixo desvio padrão devido ao baixo valor da média, o comportamento de geração do SumOpinions tende a mostrar melhores resultados. E para demonstrar as diferenças, na página seguinte temos uma comparação de sumários gerados sobre a Torre Eiffel conforme as abordagens de experimentos definida. Na coluna esquerda, é apresentado um sumário resultante do trabalho de Hu *et al.* (2017) e, à direita, encontra-se o sumário gerado pelo SumOpinions.

It is fantastic! Yes, it's crowded and hot. Seeing Eiffel up close is a treat. It was indeed magnificent despite being really hot, crowded and long waits. However, the real magic is climbing to the top. We couldn't go to the top because of the limited space but we were able to go to the mid-station which was so much fun and offered Panoramic views of the city. Firstly there's a bag search to get into the inner square where you then have to queue again to buy your tickets. The views from the top are fantastic and well worth the extra effort. Unfortunately, the violence and terror that has rocked many parts of... Wow.

Tower is just fantastic. Is there something left to say about the Eiffel Tower? No you waste time and extra money for a time to the top. It certainly is an amazing structure but I'd just wish they would limit amount of entries more. At night, the Eiffel lights up – and Paris comes to life with colour (pictures attached). I was unprepared for how truly magnificent the views from up in the tower are. I booked a tour with Viator Tours for the Louvre, Eiffel Tower and Notre Dame. We had pre booked timed tickets which saved us joining the rather long ticket cue. Be aware security is mad and most folk appear unable to stand in line and wait their turn. Loved the tower but the aggressive trinket vendors spoil the atmosphere.

Os sumários acima representam os resultados de textos com a configuração de 10 sentenças. Salientamos que, em ambos os sumários, destacamos em negrito a presença de palavras e sentenças que agregam informações ao leitor, especialmente contemplando um tópico extraído do *corpus*, conforme a Tabela 7. Em geral, os sumários resultantes deste trabalho tendem a refletir textos com melhor estrutura e conteúdo devido a fase de filtragem das sentenças e cobertura de tópicos, abordada na Seção 4.3. Assim, evita-se cenários em que sentenças com palavras soltas ou que não refletem um sentido completo sejam utilizadas. Percebe-se que há uma variação de tópicos igual ou maior no texto à direita, onde as frases tendem a fornecer um sentido completo e objetivo em relação ao tópico em questão. Podemos observar na Figura 19 e Figura 20 os resultados da diversidade de tópicos para Torre Eiffel, onde o eixo x representa os tópicos extraídos conforme seu número representativo e o eixo y reflete o número de sumários que contemplam cada tópico.

Figura 19 – Distribuição de tópicos sobre os 100 sumários gerados com 20 sentenças para a Torre Eiffel.

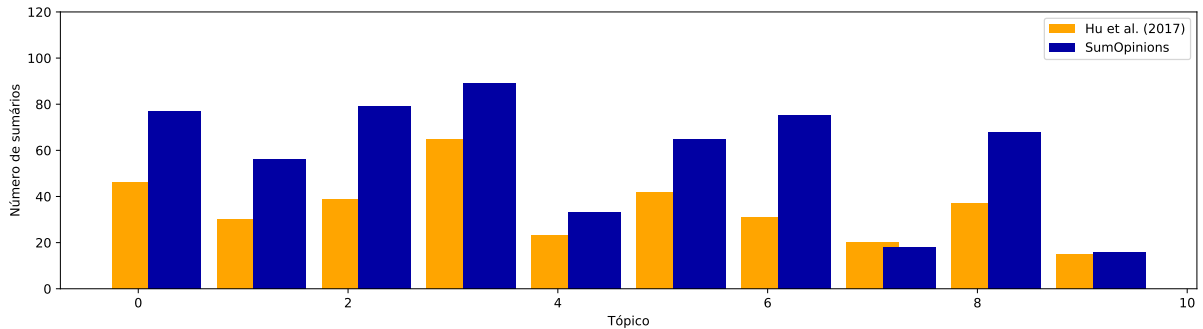
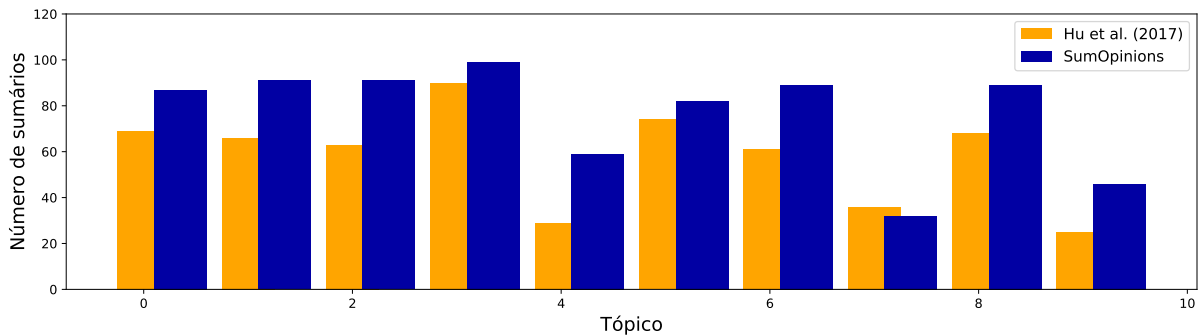
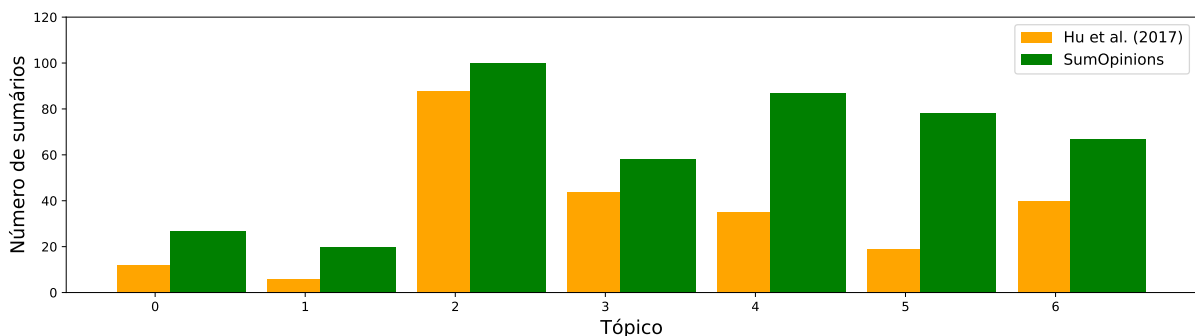


Figura 20 – Distribuição de tópicos sobre os 100 sumários gerados com 30 sentenças para a Torre Eiffel.



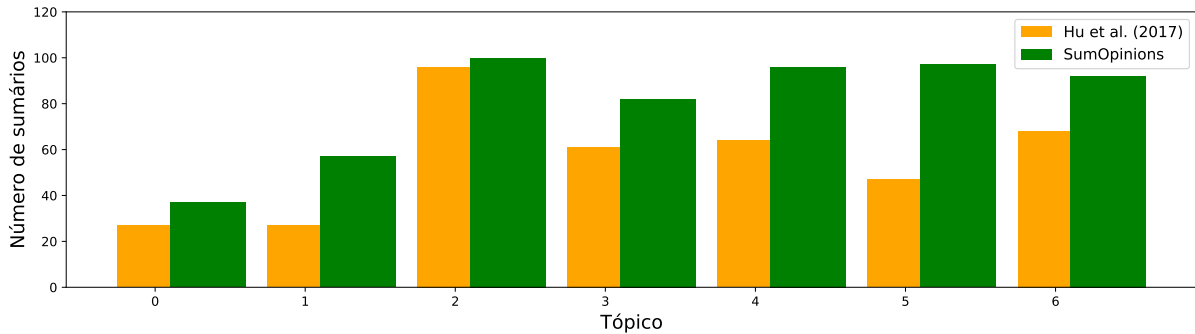
Semelhantemente, apresentamos os resultados da distribuição de tópicos para os sumários do Parque Central nas Figuras 21 e 22.

Figura 21 – Distribuição de tópicos sobre os 100 sumários gerados com 20 sentenças para o Parque Central.



Percebemos acima que a proporção de tópicos abordados nos sumários resultantes deste trabalho é maior em relação ao de Hu *et al.* (2017), figurando a presença de sentenças diversificadas em termos de assunto, bem como maximizando a presença de todos os tópicos descobertos. Logo, os resultados se apresentam de maneira satisfatória, superando os resultados do trabalho referência.

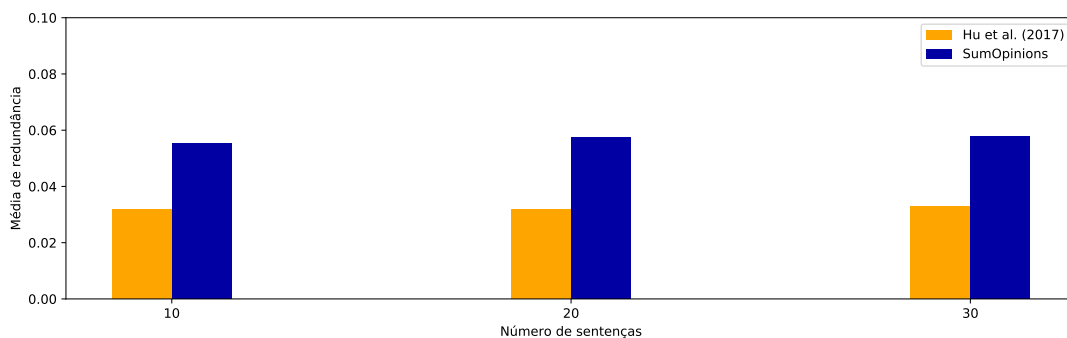
Figura 22 – Distribuição de tópicos sobre os 100 sumários gerados com 30 sentenças para o Parque Central.



5.3.2 Análise de redundância

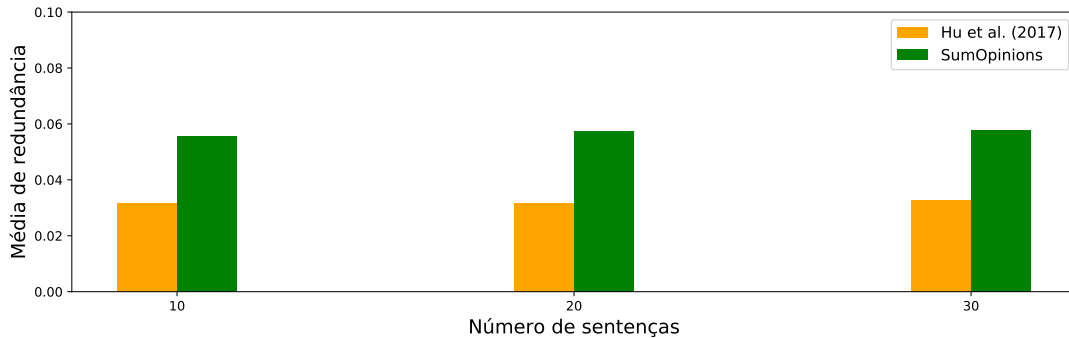
Nesta etapa de experimentos buscamos identificar o quão redundante são os sumários baseada na similaridade das sentenças. Dessa forma, para cada sumário, aplicamos a similaridade Jaccard (NIWATTANAKUL *et al.*, 2013) para todas as combinações de sentenças de cada sumário gerado. Em geral, isso representa o quão similar são duas sentenças com base nas palavras utilizadas. Abaixo os resultados na Figura 23 e 24 comparando a redundância para todos os 100 sumários da Torre Eiffel e do Parque Central, respectivamente, de acordo com a configuração da quantidade de sentenças.

Figura 23 – Comparativo de redundância de acordo com a quantidade de sentenças para a Torre Eiffel.



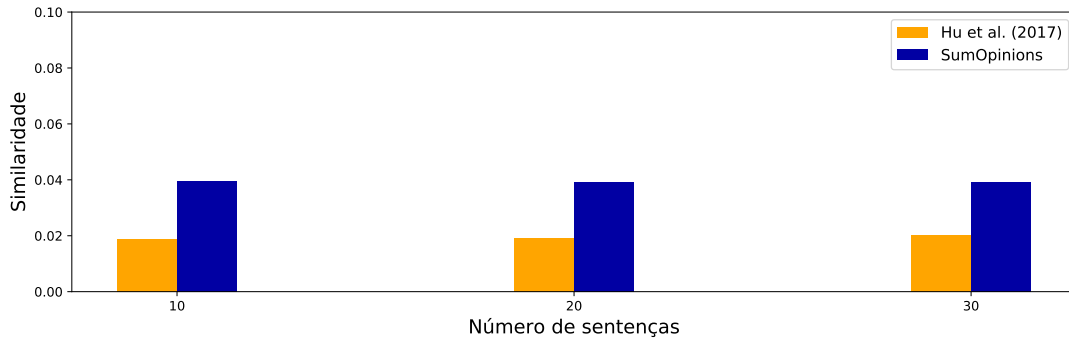
Podemos perceber que a média geral de redundância, representado pelo eixo y, dos sumários de SumOpinions reflete quase que o dobro em relação aos sumários de (HU *et al.*, 2017), o que podemos inferir que ele tende a gerar sumários com maior redundância, dado que as sentenças tendem a repetir certas palavras. No entanto, a média em todos os casos é muito baixa, o que implica em sentenças diversificadas na composição dos sumários, independente da abordagem. Assim, reflete também em resultados constantes para as diferentes quantidades de sentenças na configuração dos sumários.

Figura 24 – Comparativo de redundância de acordo com a quantidade de sentenças para o Parque Central.



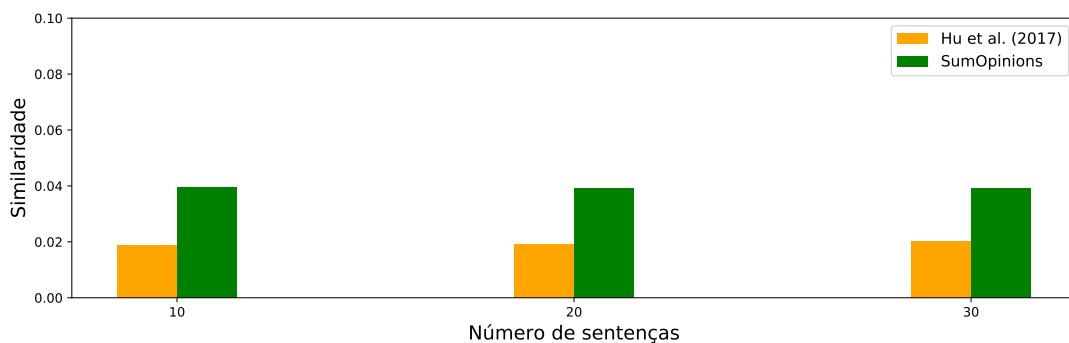
Por outro lado, para os mesmos 100 sumários, podemos aplicar a mesma similaridade considerando a composição das sentenças em relação aos tópicos. Portanto, sendo um conjunto T contendo todas as palavras dos tópicos, de acordo com a Tabela 7, comparamos com os termos de cada sentença para verificar a similaridade geral dos sumários em relação aos tópicos, de acordo com a Figura 25.

Figura 25 – Comparativo da similaridade de sentenças em relação aos tópicos para a Torre Eiffel.



Considerando o conjunto T de palavras dos tópicos referentes a Tabela 8, demonstramos resultados semelhantes ao anterior para o Parque Central na Figura 26.

Figura 26 – Comparativo da similaridade de sentenças em relação aos tópicos para o Parque Central.



Diante dos resultados, é importante refletir que quando se otimiza a diversidade com base nos tópicos, não implica em maior dissimilaridade das sentenças, pois os tópicos podem compartilhar de palavras semelhantes ou ocorrer de duas sentenças distintas expressarem sobre um mesmo tópico. Além dos mais, conforme apresentado no Algoritmo 1, as sentenças são ranqueadas conforme a presença dos tópicos, resultando em um conjunto menor de sentenças escolhidas para fazer parte do sumário, o que explica o aumento de similaridade. Além disso, o fato do sumário de Hu *et al.* (2017) apresentar menor similaridade de sentenças não significa ter sentenças mais relevantes nos sumários contemplando os assuntos principais. Ou seja, a ideia dos tópicos não é apenas amenizar a redundância, mas amenizar as sentenças irrelevantes, como sentenças de conteúdo isolado, assuntos desnecessários ou até pessoais, ao mesmo tempo que maximiza a importância do sumário com sentenças contemplando os assuntos gerais: os tópicos extraídos.

5.3.3 Resultados da qualidade de leitura

Um dos pontos importantes para atestar a qualidade do sumário é identificar a dificuldade de leitura do mesmo, pois uma leitura difícil pode implicar em uma construção inadequada do texto. Contudo, identificar a qualidade resultante da sumarização é um aspecto relevante por representar uma leitura agradável e, conseqüentemente, obter um leitor envolvido com a compreensão do texto. Portanto, as métricas em dificuldade de leitura foram aplicadas nos 100 sumários com 20 e 30 sentenças, dada maior relevância se comparado a um sumário de 10 sentenças. Nas Figuras 27 e 28, estando no eixo *x* as abordagens para serem comparadas e no eixo *y* a métrica de qualidade de leitura, de acordo com a Subseção 2.5.3.2, apresentamos os resultados.

Figura 27 – Dificuldade de leitura dos sumários com 20 sentenças sobre a Torre Eiffel.

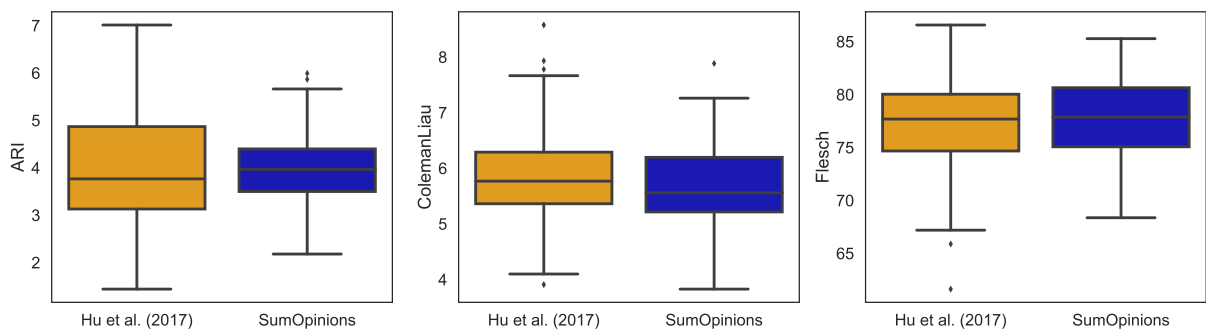
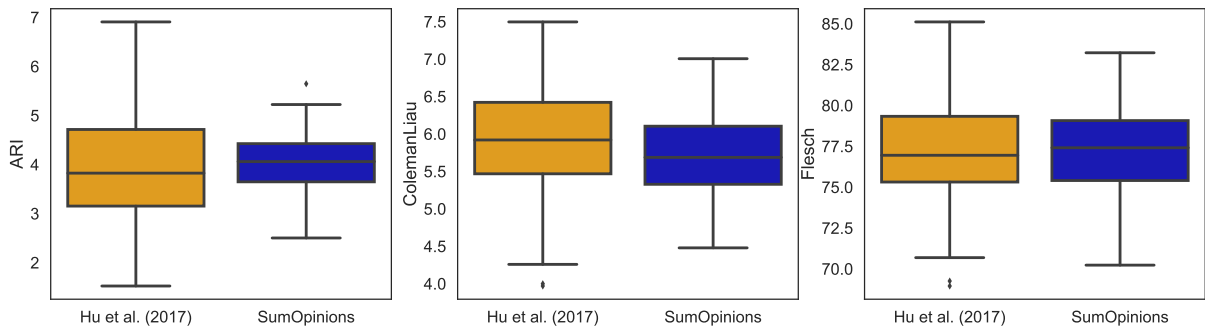


Figura 28 – Dificuldade de leitura dos sumários com 30 sentenças sobre a Torre Eiffel.



Em geral, os resultados foram semelhantes nas três métricas de leitura para ambas as abordagens, em termos de média e dificuldade de leitura. Para Flesch, admitindo-se um resultado médio igual a 77, temos *7th grade*, o que implica em fácil leitura. Para ColemanLiau, sendo uma média de 6, temos *6th grade*, representando uma dificuldade semelhante ao de Flesch. Já para ARI, com média igual a 4, temos resultados para idades de 8 a 9 anos, também representando textos de fácil leitura. Para maiores detalhes sobre os escores, olhar as tabelas de escores na Subseção 2.5.3.2.

Mesmo sem diferenças relevantes, é interessante notar que a configuração de 30 sentenças demonstra um menor desvio padrão em relação aos outros cenários, independente da abordagem. Esse resultado é possivelmente explicado pela maior presença de sentenças relevantes em termos de tamanho e estrutura, dado que todas são selecionados com base na definição de sentença mais importante de um agrupamento.

Contudo, há algumas diferenças presentes nos resultados acima: a divergência entre os valores com maior frequência e a diferença dos intervalos de valores em cada métrica. Percebe-se que, nos resultados deste trabalho, o intervalo do eixo y tende a ser maior em alguns casos, indicando que há maior frequência de sumários que atingem os resultados satisfatórios. Além disso, tanto o fato dos intervalos de valores serem menores, quanto a maior presença de valores centralizados, demonstram que geralmente os resultados são mais homogêneos, ou seja, já se espera resultados semelhantes e de boa qualidade aplicando a metodologia desenvolvida. Portanto, os resultados apresentados são satisfatórios, semelhantemente ao que foi apresentado na relação da diversidade de tópicos.

Nas Figuras 29 e 30 são apresentados os resultados para as opiniões sobre o Parque Central. Podemos perceber que os resultados são semelhantes aos anteriores sobre a Torre Eiffel.

Figura 29 – Dificuldade de leitura dos sumários com 20 sentenças sobre o Parque Central.

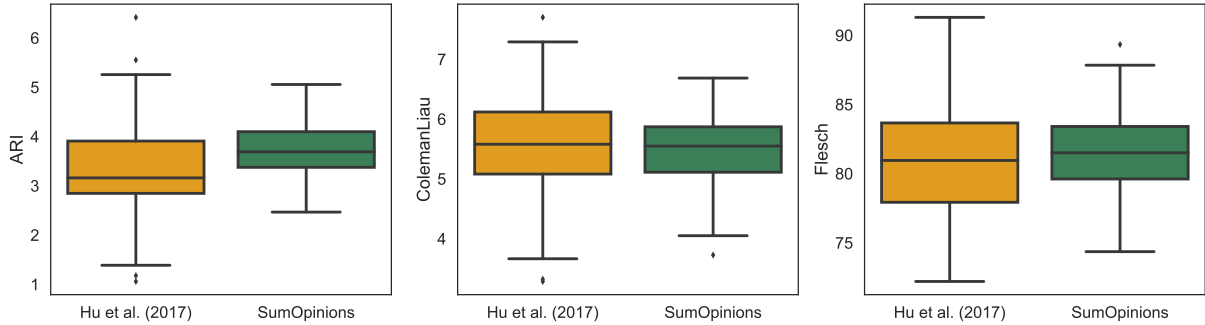
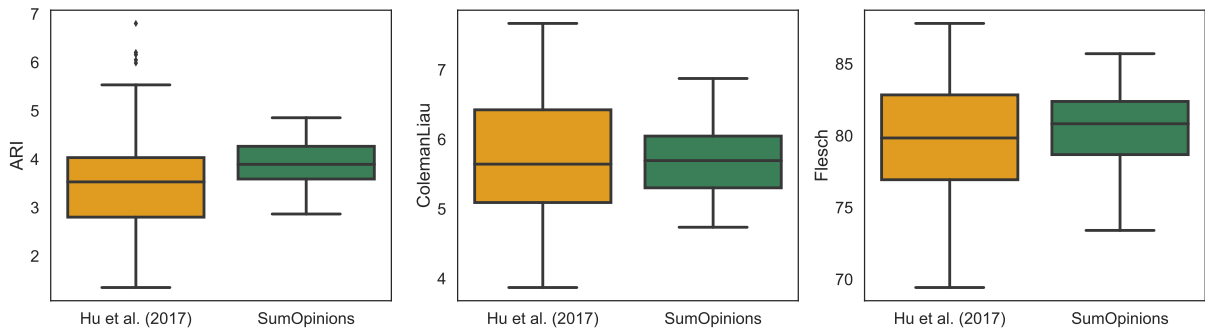


Figura 30 – Dificuldade de leitura dos sumários com 30 sentenças sobre o Parque Central.



6 CONCLUSÕES E TRABALHOS FUTUROS

De acordo com os resultados apresentados no Capítulo 5, a diversidade de tópicos foi contemplada e maximizada com a utilização de métodos em modelagem de tópicos integrados à sumarização extrativa. O princípio se deu com base na descoberta de sentenças relevantes em termos de tópicos, os quais foram previamente identificados no *corpus* utilizado. Importante destacar que a modelagem de tópicos influencia na relevância dos sumários gerados desde que é possível abordar os temas de maiores destaques nas opiniões, além de minimizar a presença de conteúdos irrelevantes.

Utilizando-se do algoritmo LDA para a descoberta dos tópicos, foi possível pontuar a importância das sentenças do corpo textual com base na estrutura e conteúdo referente aos tópicos. Com base nesse processo, foram obtidos resultados superiores, em termos de cobertura dos tópicos, se comparados ao trabalho de (HU *et al.*, 2017), o qual não se preocupa, explicitamente, em promover uma sumarização baseada em tópicos. Além disso, é possível obter sumários específicos para um conjunto de tópicos, permitindo que o sistema sumarizador utilize apenas um subconjunto dos tópicos descobertos para extrair as sentenças. O que possibilita também uma sumarização baseada em tópicos que possa receber consultas do usuários, onde o mesmo pode selecionar os tópicos que julga importante no sumário para a sua avaliação.

De acordo com a análise de redundância na Subseção 5.3.2, percebemos que a utilização de modelagem de tópicos para otimizar os tópicos envolvidos no sumário final não, necessariamente, diminui a redundância dos sumários. Isso vem do fato de que certos tópicos podem conter palavras semelhantes, mas que em sua totalidade representa um tema distinto. Tais resultados dependem muito de decisão e experimento, onde é possível decidir como a função de escore vai pontuar uma sentença em relação ao tópico, por exemplo. Contudo, é válido dizer que tende a gerar sumários com boa cobertura de tópicos e com comportamento constante em relação aos tópicos contemplados, gerando sumários relevantes no ponto de vista em agregar informações para o usuário.

No entanto, como o processo de construção do sumário ainda utiliza em grande parte a metodologia de Hu *et al.* (2017), o resultado da cobertura de tópicos ainda pode ser otimizada, pois além da fase de "Modelagem de tópicos" adicionada neste trabalho, é possível estender todas as outras fases da metodologia de forma a qualificar e ponderar os tópicos, principalmente na fase de construção do sumário, descrito na Seção 4.6. Portanto, essa é uma atividade futura que este trabalho pretende abordar.

Além disso, diversas heurísticas foram definidas para considerar a importância do conteúdo e do sentimento das sentenças. Logo, faz-se interessante a necessidade de elaborar heurísticas que também considerem a importância de tópicos presentes.

Um ponto interessante como trabalho futuro na geração dos sumários seria, baseado na identificação de sentimento, construir sumários que consigam balancear informações de teor positivos e negativos de maneira parametrizada. Como resultado, poderemos ter uma sumarização regulada ao contexto do usuário.

Ademais, na geração de sumários há diversos fatores de qualidade que precisam ser aplicados para garantir um resultado adequado. Para isso, este trabalho aplica métricas de qualidade de leitura em busca de identificar a razoabilidade de entendimento do texto por parte do usuário. Portanto, como apresentado no Capítulo 5, os resultados se mostraram positivos em relação ao nível de escolaridade necessário, como também se comparado aos resultados apresentados pelo artigo referência utilizado.

No entanto, outro fator importante de qualidade é a coerência do texto, de modo a avaliar a qualidade em que as sentenças foram extraídas e unidas, dado que se trata de uma sumarização extrativa. No entanto, como tal tipo de teste não foi contemplado nesse trabalho, pretende-se como atividade futura.

Por fim, uma das etapas mais onerosas do trabalho se trata da aplicação da medida de similaridade de conteúdo, pois se utiliza da função NGD (Normalized Google Distance). Sua dificuldade se trata da necessidade de coleta dos *hits* de cada palavra e dos pares possíveis existentes no vocabulário em algum mecanismo de busca, tais como *Google Search*, *Bing* ou *Yahoo search*. Portanto, como trabalho futuro, pretende-se encontrar outro método que atinja resultados positivos com menor custo e dependência.

REFERÊNCIAS

- ACHANANUPARP, P.; HU, X.; SHEN, X. The evaluation of sentence similarity measures. In: SPRINGER. **International Conference on data warehousing and knowledge discovery**. Berlin, 2008. p. 305–316.
- ADY, M.; QUADRI-FELITTI, D. *et al.* Consumer research identifies how to present travel review content for more bookings. **Retrieved form <http://webcache.googleusercontent.com/search>**, 2015.
- ALGHAMDI, R.; ALFALQI, K. A survey of topic modeling in text mining. **Int. J. Adv. Comput. Sci. Appl.(IJACSA)**, Citeseer, v. 6, n. 1, 2015.
- ALGULIEV, R. M.; ALIGULIYEV, R. M.; HAJIRAHIMOVA, M. S.; MEHDIYEV, C. A. Mcmr: Maximum coverage and minimum redundant text summarization model. **Expert Systems with Applications**, Elsevier, v. 38, n. 12, p. 14514–14522, 2011.
- ALIGULIYEV, R. M. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. **Expert Systems with Applications**, Elsevier, v. 36, n. 4, p. 7764–7772, 2009.
- ARI. **Tabela de escores para a métrica ARI**. 2018. Disponível em: https://en.wikipedia.org/wiki/Automated_readability_index. Acesso em: 2 jul. 2018.
- ARTIN, E. **The gamma function**. [s.l.]: Courier Dover Publications, 2015.
- BING, L.; LI, P.; LIAO, Y.; LAM, W.; GUO, W.; PASSONNEAU, R. J. Abstractive multi-document summarization via phrase selection and merging. **arXiv preprint arXiv:1506.01597**, 2015.
- BLEI, D. M.; LAFFERTY, J. D. Topic models. In: **Text Mining**. [s.l.]: Chapman and Hall/CRC, 2009. p. 101–124.
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.
- BRANTS, T.; CHEN, F.; TSOCHANTARIDIS, I. Topic-based document segmentation with probabilistic latent semantic analysis. In: ACM. **Proceedings of the eleventh international conference on Information and knowledge management**. Virginia, 2002. p. 211–218.
- CAO, Z.; WEI, F.; DONG, L.; LI, S.; ZHOU, M. Ranking with recursive neural networks and its application to multi-document summarization. In: **Twenty-ninth AAAI conference on artificial intelligence**. [s.l.: s.n.], 2015. p. 2153–2159.
- CARBONELL, J.; GOLDSTEIN, J. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: ACM. **Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval**. [s.l.], 1998. p. 335–336.
- CATALDI, M.; BALLATORE, A.; TIDDI, I.; AUFAURE, M.-A. Good location, terrible food: detecting feature sentiment in user-generated reviews. **Social Network Analysis and Mining**, Springer, v. 3, n. 4, p. 1149–1163, 2013.

- CHUNG, N.; KOO, C. The use of social media in travel information search. **Telematics and Informatics**, Elsevier, v. 32, n. 2, p. 215–229, 2015.
- CILIBRASI, R. L.; VITANYI, P. M. The google similarity distance. **IEEE Transactions on knowledge and data engineering**, IEEE, v. 19, n. 3, 2007.
- COHN, T.; LAPATA, M. Sentence compression beyond word deletion. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1**. [s.l.], 2008. p. 137–144.
- COLEMAN, M.; LIAU, T. L. A computer readability formula designed for machine scoring. **Journal of Applied Psychology**, American Psychological Association, v. 60, n. 2, p. 283, 1975.
- CONDORI, R. E. L. **Sumarização automática de opiniões baseada em aspectos**. Tese (Doutorado) — Universidade de São Paulo, 2014.
- DAS, D.; MARTINS, A. F. A survey on automatic text summarization. **Literature Survey for the Language and Statistics II course at CMU**, v. 4, p. 192–195, 2007.
- DAVIDOV, D.; TSUR, O.; RAPPOPORT, A. Enhanced sentiment learning using twitter hashtags and smileys. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 23rd international conference on computational linguistics: posters**. [s.l.], 2010. p. 241–249.
- DUMAIS, S. T. Latent semantic analysis. **Annual review of information science and technology**, Wiley Online Library, v. 38, n. 1, p. 188–230, 2004.
- DURRETT, G.; BERG-KIRKPATRICK, T.; KLEIN, D. Learning-based single-document summarization with compression and anaphoricity constraints. **arXiv preprint arXiv:1603.08887**, 2016.
- ERKAN, G.; RADEV, D. R. Lexpagerank: Prestige in multi-document text summarization. In: **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing**. [s.l.: s.n.], 2004.
- FLESCHE, R. A new readability yardstick. **Journal of applied psychology**, American Psychological Association, v. 32, n. 3, p. 221, 1948.
- FLORES-GARRIDO, M.; CARRASCO-OCHOA, J.-A.; MARTÍNEZ-TRINIDAD, J. F. Agrap: an algorithm for mining frequent patterns in a single graph using inexact matching. **Knowledge and Information Systems**, Springer, v. 44, n. 2, p. 385–406, 2015.
- GANESAN, K.; ZHAI, C.; HAN, J. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 23rd international conference on computational linguistics**. [s.l.], 2010. p. 340–348.
- GILLICK, D.; FAVRE, B. A scalable global model for summarization. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing**. [s.l.], 2009. p. 10–18.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. **Deep learning**. [s.l.]: MIT press Cambridge, 2016. v. 1.

- HOFFMAN, M.; BACH, F. R.; BLEI, D. M. Online learning for latent dirichlet allocation. In: **advances in neural information processing systems**. [s.l.: s.n.], 2010. p. 856–864.
- HOFMANN, T. Probabilistic latent semantic analysis. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence**. [s.l.], 1999. p. 289–296.
- HU, M.; LIU, B. Mining and summarizing customer reviews. In: ACM. **Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining**. [s.l.], 2004. p. 168–177.
- HU, Y.-H.; CHEN, Y.-L.; CHOU, H.-L. Opinion mining from online hotel reviews—a text summarization approach. **Information Processing & Management**, Elsevier, v. 53, n. 2, p. 436–449, 2017.
- HUANG, A. Similarity measures for text document clustering. In: **Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand**. [s.l.: s.n.], 2008. p. 49–56.
- KAUFMAN, L.; ROUSSEEUW, P. **Clustering by means of medoids**. [s.l.]: North-Holland, 1987.
- KIM, H. D.; GANESAN, K.; SONDHI, P.; ZHAI, C. **Comprehensive review of opinion summarization**. [s.l.], 2011.
- KLEMA, V.; LAUB, A. The singular value decomposition: Its computation and some applications. **IEEE Transactions on automatic control**, IEEE, v. 25, n. 2, p. 164–176, 1980.
- LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
- LAPATA, M.; BARZILAY, R. Automatic evaluation of text coherence: Models and representations. In: **IJCAI**. [s.l.: s.n.], 2005. v. 5, p. 1085–1090.
- LI, L.; ZHOU, K.; XUE, G.-R.; ZHA, H.; YU, Y. Enhancing diversity, coverage and balance for summarization through structure learning. In: ACM. **Proceedings of the 18th international conference on World wide web**. [s.l.], 2009. p. 71–80.
- LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. **Text Summarization Branches Out**, 2004.
- LIN, C.-Y.; HOVY, E. From single to multi-document summarization: A prototype system and its evaluation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. [s.l.], 2002. p. 457–464.
- LIN, H.; BILMES, J. A class of submodular functions for document summarization. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1**. [s.l.], 2011. p. 510–520.
- LIU, B. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

- LIU, Q.; GAO, Z.; LIU, B.; ZHANG, Y. A logic programming approach to aspect extraction in opinion mining. In: IEEE. **Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on**. [s.l.], 2013. v. 1, p. 276–283.
- LOPYREV, K. Generating news headlines with recurrent neural networks. **arXiv preprint arXiv:1512.01712**, 2015.
- LYRA, D. H.; AMARAL, C. L. F. Apreensibilidade e legibilidade de artigos científicos de um periódico nacional. **Tekhne e Logos**, v. 3, n. 3, p. 90–101, 2012.
- MANI, I. **Advances in automatic text summarization**. [s.l.]: MIT press, 1999.
- MANI, I.; MAYBURY, M. T. Automatic summarization. J. Benjamins Publishing Company New York, USA, 2001.
- MCDONALD, R. A study of global inference algorithms in multi-document summarization. In: SPRINGER. **European Conference on Information Retrieval**. [s.l.], 2007. p. 557–564.
- MEI, Q.; LING, X.; WONDRA, M.; SU, H.; ZHAI, C. Topic sentiment mixture: modeling facets and opinions in weblogs. In: ACM. **Proceedings of the 16th international conference on World Wide Web**. [s.l.], 2007. p. 171–180.
- MEI, Q.; LIU, C.; SU, H.; ZHAI, C. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In: ACM. **Proceedings of the 15th international conference on World Wide Web**. [s.l.], 2006. p. 533–542.
- MEI, Q.; ZHAI, C. A mixture model for contextual text mining. In: ACM. **Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining**. [s.l.], 2006. p. 649–655.
- METZLER, D.; KANUNGO, T. Machine learned sentence selection strategies for query-biased summarization. In: **Sigir learning to rank workshop**. [s.l.: s.n.], 2008. p. 40–47.
- MIHALCEA, R.; TARAU, P. A language independent algorithm for single and multiple document summarization. In: **Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts**. [s.l.: s.n.], 2005.
- MILLER, G. A. Wordnet: a lexical database for english. **Communications of the ACM**, ACM, v. 38, n. 11, p. 39–41, 1995.
- MOON, T. K. The expectation-maximization algorithm. **IEEE Signal processing magazine**, IEEE, v. 13, n. 6, p. 47–60, 1996.
- NENKOVA, A.; MCKEOWN, K. A survey of text summarization techniques. In: **Mining text data**. [s.l.]: Springer, 2012. p. 43–76.
- NENKOVA, A.; MCKEOWN, K. *et al.* Automatic summarization. **Foundations and Trends® in Information Retrieval**, Now Publishers, Inc., v. 5, n. 2–3, p. 103–233, 2011.
- NIWATTANAKUL, S.; SINGTHONGCHAI, J.; NAENUDORN, E.; WANAPU, S. Using of jaccard coefficient for keywords similarity. In: **Proceedings of the International MultiConference of Engineers and Computer Scientists**. [s.l.: s.n.], 2013. v. 1, n. 6.

- PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. **The PageRank citation ranking: Bringing order to the web.** [s.l.], 1999.
- PAK, A.; PAROUBEK, P. Twitter as a corpus for sentiment analysis and opinion mining. In: **LREc.** [s.l.: s.n.], 2010. v. 10, n. 2010, p. 1320–1326.
- PANG, B.; LEE, L. *et al.* Opinion mining and sentiment analysis. **Foundations and Trends® in Information Retrieval**, Now Publishers, Inc., v. 2, n. 1–2, p. 1–135, 2008.
- PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.** [s.l.], 2002. p. 79–86.
- PATERSON, M.; DANČÍK, V. Longest common subsequences. In: SPRINGER. **International Symposium on Mathematical Foundations of Computer Science.** [s.l.], 1994. p. 127–142.
- PONTIKI, M.; GALANIS, D.; PAPAGEORGIOU, H.; ANDROUTSOPOULOS, I.; MANANDHAR, S.; MOHAMMAD, A.-S.; AL-AYYOUB, M.; ZHAO, Y.; QIN, B.; CLERCQ, O. D. *et al.* Semeval-2016 task 5: Aspect based sentiment analysis. In: **Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016).** [s.l.: s.n.], 2016. p. 19–30.
- PORTER, M. F. An algorithm for suffix stripping. **Program**, MCB UP Ltd, v. 14, n. 3, p. 130–137, 1980.
- QIANG, J.-P.; CHEN, P.; DING, W.; XIE, F.; WU, X. Multi-document summarization using closed patterns. **Knowledge-Based Systems**, Elsevier, v. 99, p. 28–38, 2016.
- QIU, G.; LIU, B.; BU, J.; CHEN, C. Opinion word expansion and target extraction through double propagation. **Computational linguistics**, MIT Press, v. 37, n. 1, p. 9–27, 2011.
- RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. **Proceedings of the IEEE**, Ieee, v. 77, n. 2, p. 257–286, 1989.
- RADEV, D. R.; ALLISON, T.; BLAIR-GOLDENSOHN, S.; BLITZER, J.; CELEBI, A.; DIMITROV, S.; DRABEK, E.; HAKIM, A.; LAM, W.; LIU, D. *et al.* Mead-a platform for multidocument multilingual text summarization. 2004.
- RADEV, D. R.; JING, H.; STYŠ, M.; TAM, D. Centroid-based summarization of multiple documents. **Information Processing & Management**, Elsevier, v. 40, n. 6, p. 919–938, 2004.
- RAMOS, J. *et al.* Using tf-idf to determine word relevance in document queries. In: **Proceedings of the first instructional conference on machine learning.** [s.l.: s.n.], 2003. v. 242, p. 133–142.
- RATNAPARKHI, A. Maximum entropy models for natural language ambiguity resolution. 1998.
- RUSH, A. M.; CHOPRA, S.; WESTON, J. A neural attention model for abstractive sentence summarization. **arXiv preprint arXiv:1509.00685**, 2015.
- SHIMADA, K.; TADANO, R.; ENDO, T. Multi-aspects review summarization with objective information. **Procedia-Social and Behavioral Sciences**, Elsevier, v. 27, p. 140–149, 2011.

- STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. **Handbook of latent semantic analysis**, v. 427, n. 7, p. 424–440, 2007.
- TABOADA, M.; BROOKE, J.; TOFILOSKI, M.; VOLL, K.; STEDE, M. Lexicon-based methods for sentiment analysis. **Computational linguistics**, MIT Press, v. 37, n. 2, p. 267–307, 2011.
- TOUTANOVA, K.; BROCKETT, C.; GAMON, M.; JAGARLAMUDI, J.; SUZUKI, H.; VANDERWENDE, L. The pythy summarization system: Microsoft research at duc 2007. In: **Proc. of DUC**. [s.l.: s.n.], 2007. v. 2007.
- TRIPADVISOR. **Informações sobre o autor na plataforma**. 2018. Disponível em: <https://www.tripadvisor.com/members/YorkTraveller1962>. Acesso em: 28 jun. 2018.
- TRIPADVISOR. **Informações sobre uma opinião na plataforma**. 2018. Disponível em: https://www.tripadvisor.com/Attraction_Review-g187147-d188151-Reviews-Eiffel_Tower-Paris_Ile_de_France.html. Acesso em: 28 jun. 2018.
- TURNEY, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 40th annual meeting on association for computational linguistics**. [s.l.], 2002. p. 417–424.
- TURNEY, P. D.; LITTMAN, M. L. Measuring praise and criticism: Inference of semantic orientation from association. **ACM Transactions on Information Systems (TOIS)**, ACM, v. 21, n. 4, p. 315–346, 2003.
- WAN, X.; YANG, J.; XIAO, J. Single document summarization with document expansion. In: **AAAI**. [s.l.: s.n.], 2007. p. 931–936.
- WANG, D.; ZHU, S.; LI, T. Sumview: A web-based engine for summarizing product reviews and customer opinions. **Expert Systems with Applications**, Elsevier, v. 40, n. 1, p. 27–33, 2013.
- XIE, F.; WU, X.; ZHU, X. Document-specific keyphrase extraction using sequential patterns with wildcards. In: IEEE. **Data Mining (ICDM), 2014 IEEE International Conference on**. [s.l.], 2014. p. 1055–1060.
- YAN, X.; HAN, J.; AFSHAR, R. Clospan: Mining: Closed sequential patterns in large datasets. In: SIAM. **Proceedings of the 2003 SIAM international conference on data mining**. [s.l.], 2003. p. 166–177.
- ZHANG, Z.; YE, Q.; ZHANG, Z.; LI, Y. Sentiment classification of internet restaurant reviews written in cantonese. **Expert Systems with Applications**, Elsevier, v. 38, n. 6, p. 7674–7682, 2011.