MODELOS PARA A IDENTIFICAÇÃO DE TÓPICOS EM NOTÍCIAS EXTRAÍDAS DA WEB E FILTRADAS POR QUALIDADE DE DADOS

Luiz Fernando Cagiano Parodi de Frias

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: José Manoel de Seixas

Rio de Janeiro
Março de 2019

MODELOS PARA A IDENTIFICAÇÃO DE TÓPICOS EM NOTÍCIAS EXTRAÍDAS DA WEB E FILTRADAS POR QUALIDADE DE DADOS

Luiz Fernando Cagiano Parodi de Frias

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. José Manoel de Seixas, D.Sc.

Prof. Fernando Guimarães Ferreira, D.Sc.

Prof. Daniel Ratton Figueiredo, D.Sc.

Prof. Marley Maria Bernardes Rebuzzi Vellasco, D.Sc.

Prof. Sergio Lima Netto, D.Sc.

Prof. Alexandre Gonçalves Evsukoff, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2019

Frias, Luiz Fernando Cagiano Parodi de

Modelos para a Identificação de Tópicos em Notícias Extraídas da Web e Filtradas por Qualidade de Dados/Luiz Fernando Cagiano Parodi de Frias. – Rio de Janeiro: UFRJ/COPPE, 2019.

X, 72 p.: il.; 29, 7cm.

Orientador: José Manoel de Seixas

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2019.

Referências Bibliográficas: p. 63 – 72.

1. Modelagem de Tópicos. 2. Qualidade de Dados. 3. Fatoração de Matrizes Não-Negativas. I. Seixas, José Manoel de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Agradecimentos

À Eliane, responsável por prover as condições necessárias para a minha educação. E ao restante da minha família, que sempre esteve presente nos momentos mais importantes.

À Karolina, por todo amor, apoio e, principalmente, paciência durante a preparação deste trabalho.

Ao meu orientador, Seixas, pelos ensinamentos que vou levar para a vida.

Aos amigos de laboratório, Phil, Júnior, Júlio, Breno e cia, pela troca de ideais que enriqueceram este trabalho e por todo o aprendizado ao longo do mestrado.

Ao meu chefe, Eric Leite, pela força e compreensão nessa reta final.

A TWIST, Fernando Ferreira, Laura Moraes e Felipe Grael, por muito: pela orientação ao longo do trabalho final de graduação, que deu origem à esta dissertação; pelos ensinamentos ao longo destes anos; pela amizade e, finalmente, por ceder a infraestrutura necessária para este trabalho.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MODELOS PARA A IDENTIFICAÇÃO DE TÓPICOS EM NOTÍCIAS EXTRAÍDAS DA WEB E FILTRADAS POR QUALIDADE DE DADOS

Luiz Fernando Cagiano Parodi de Frias

Março/2019

Orientador: José Manoel de Seixas

Programa: Engenharia Elétrica

O aumento no número de usuários com acesso a web, das taxas de conectividade e dispositivos móveis, criou uma nova dinâmica para a publicação e disseminação de conteúdo online.

Devido à não uniformidade dos formatos de publicação, a captura de informação por agentes automáticos mostra-se uma tarefa complexa e sujeita à introdução de erros no conteúdo extraído. Dessa forma, a busca por informação relevante torna-se mais complexa, o que enseja o desenvolvimento de técnicas de processamento de texto capazes de extrair informações relevantes de um grande volume de documentos com possíveis problemas de qualidade.

A Modelagem de Tópicos é um conjunto de técnicas que tem por objetivo resumir, explorar e categorizar um conjunto de documentos de maneira não-supervisionada. Como desafios da área estão a garantia de interpretabilidade dos grupos encontrados, além da escolha pelo número ideal de tópicos.

Este trabalho avaliou o uso das medidas de coerência e estabilidade para a escolha do número de tópicos, de forma a garantir a coerência semântica dos grupos, em bases de dados não-anotadas, com notícias sujeitas a problemas de qualidade de dados. Para tanto, foram definidas dimensões e critérios de qualidade a serem atendidos pelos documentos e as medidas de coerência e estabilidade foram avaliadas para diferentes níveis de ruído.

Como resultado, a filtragem por qualidade de dados aumentou a coerência da extração de tópicos, enquanto as medida de coerência e estabilidade ajudaram a diminuir o intervalo possível de escolha para o número de tópicos. No entanto, ainda não foi encontrada uma maneira de conjugar o número de tópicos, estabilidade e coerência para escolher entre uma extração mais generalista ou mais detalhista.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

MODELS FOR IDENTIFYING TOPICS IN NEWS EXTRACTED FROM THE WEB AND FILTERED BY DATA QUALITY

Luiz Fernando Cagiano Parodi de Frias

March/2019

Advisor: José Manoel de Seixas

Department: Electrical Engineering

The increase in the number of users online, connectivity rates and mobile devices has created a new dynamic for the publication and dissemination of online content.

Due to different publication formats, the capture of information by automatic agents is a complex task and prone to the introduction of errors in the extracted content. In this way, the search for relevant information becomes more complex, which leads to the development of text processing techniques that are capable of extracting relevant information from a large volume of documents with possible data quality issues.

Topic Modeling is a set of techniques that aims to summarize, explore and categorize a set of documents using unsupervised learning. As challenges in the area are the interpretability of the clusters, as well as the choice of the best number of topics.

This work evaluates the use of coherence and stability measures for choosing the number of topics, in order to guarantee the groups show semantic coherence, in non-annotated databases, with news with data quality issues.To this end, data quality dimensions and criteria are defined to be met by the documents, and coherence and stability measures are evaluated for different levels of noise.

As a result, filtering the news using data quality criteria increased the consistency of topic extraction, while the measure of coherence and stability helped to narrow the range of choice for the number of topics. However, a way of combining the number of topics, stability and coherence to choose between a more generalist or more detailed extraction has not yet been found.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Desde que foi criada, a rede mundial de computadores vem crescendo de maneira assuntosa [1, 2], tanto em número de usuários quanto em volume publicado, com parte deste conteúdo sendo composto por notícias [3].

A publicação de notícias *online* começou quando a mídia impressa passou a notar o aumento no número de usuários conectados, disponibilizando assim um pequeno conjunto de suas notícias impressas em versão digital [3, 4]. Esta tendência de investimento em conteúdo digital acompanhou o aumento do interesse do público e, em pouco tempo, todas as notícias impressas passaram a ser replicadas digitalmente. Por fim, a consolidação do meio digital de publicação gerou a produção de conteúdo exclusivo para a *web* e forçou muitos jornais a abandonarem suas operações físicas e dedicarem-se exclusivamente à nova forma de publicação [4].

Além do maior número de usuários conectados, o padrão de uso da *internet* por parte dos usuários foi se modificando com o tempo. Enquanto, no início, a televisão e as publicações impressas representavam o principal meio de acesso a notícias [5], esta preferência vem se modificando [6] e, hoje, é seguro afirmar que as notícias publicadas *online* representam o principal meio de acesso à informação [7].

O cenário descrito levou a um aumento no volume de conteúdo disponível *online*, sendo parte dele composto pelas notícias falsas [8–11], o que confere maior dificuldade para encontrar a informação correta desejada.

As ferramentas de busca mais utilizadas permitem a busca por palavra-chave [12]. Desenvolvidas antes do cenário descrito anteriormente surgir, não levam em conta a complexidade da *web* dos dias de hoje. O Google, por exemplo, embora tenha o algoritmo fechado, parece adotar um critério de relevância baseado em número de referências aos documentos que contêm a palavra-chave, mas desconsiderando a credibilidade da fonte. Para o contexto atual, especificamente no universo das notícias, seria interessante contar com outros parâmetros de busca que se adaptem ao cenário existente.

A modelagem de tópicos é um conjunto de técnicas cujo objetivo é extrair, su-

marizar e categorizar informação de um conjunto de documentos, permitindo assim uma nova forma de iteração com a base textual [13]. Através destas técnicas, é possível definir a estrutura semântica de um conjunto de notícias, conferindo uma característica desejável aos buscadores existentes.

1.1 Motivação

A busca por informação é um problema inerente à produção de conhecimento. Com o aumento do volume de notícias publicadas, é natural que exista uma maior dificuldade em encontrar um fato específico.

A Modelagem de Tópicos é um conjunto de técnicas que revela-se como uma alternativa para a busca de informação em bases textuais, possibilitando o agrupamento dos documentos por similaridade semântica, de maneira não-supervisionada. Desde o surgimento dos primeiros trabalhos a área vem ganhando crescente relevância, como mostra a Figura 1.1 com o resultado da pesquisa pelo termo *topic modeling* na plataforma *Scopus* [14]. Entre as áreas de aplicação, a maioria está associada a dados textuais, embora a área de bioinformática vem ganhando destaque.

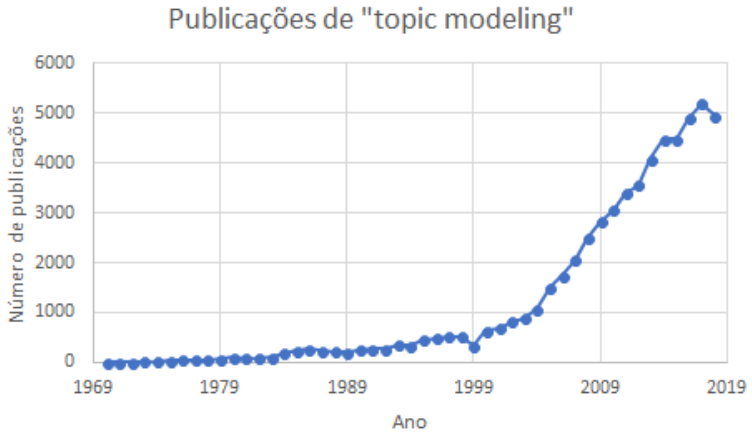


Figura 1.1: Publicações de *topic modeling*.

Já entre os desafios da área, estão: garantir a estabilidade da extração (o grau com que repetidas extrações de tópicos chegam aos mesmos tópicos), garantir a coerência semântica dos tópicos encontrados, ou seja, se os tópicos são interpretáveis por um avaliador humano, e determinar o número de tópicos.

Para este trabalho, a aplicação da modelagem de tópicos será feita com o objetivo de explorar bases de notícias jornalísticas. Como essas notícias ficam em bases de dados dispersas, o primeiro passo é coletar o conteúdo de cada um destes repositórios. Os agentes responsáveis por coletar as notícias a partir dos textos publicados nos portais dos jornais são os *web crawlers*, *softwares* responsáveis por extrair o conteúdo

textual de uma página *web* [15–19]. Devido à falta de padronização entre as formas de publicação das diferentes fontes, esta é uma tarefa complexa e sujeita a erros, que aparecem tipicamente como trechos do documento que não pertencem à notícia e trechos de código da estrutura que compõem a página.

Assim, é necessária uma análise posterior à extração para que se possa avaliar o nível de qualidade dos dados recuperados pelos agentes, a fim de que notícias que não atendam a critérios mínimos pré-estabelecidos possam ser filtradas e, dessa forma, não impactar no resultado do agrupamento.

1.2 Objetivo

Este trabalho aplica técnicas de modelagem de tópicos à base de dados de notícias escritas em língua inglesa, coletadas a partir de *web crawlers* e, portanto, sujeitas a ruído. Geralmente avaliadas em conjuntos de dados controlados e com alta relação sinal-ruído, a aplicação destas técnicas apresenta uma série de dificuldades práticas em conjuntos de dados ruidosos e com maior dimensionalidade.

O objetivo deste trabalho consiste em definir dimensões e critérios de qualidade para a filtragem das notícias ruidosas, a fim de garantir uma extração de tópicos mais coerente semanticamente. Além disso, deseja-se ainda escolher o número de tópicos de forma a garantir a estabilidade da extração e a coerência do resultado final.

1.3 O que foi obtido

Para mitigar o efeito do ruído, este trabalho investigou diferentes dimensões de qualidade que devem ser atendidas pelos documentos textuais que compõem a base. As dimensões de completude e acurácia foram definidas para os diferentes atributos que compõem uma notícia extraída por agentes automáticos.

A filtragem dos documentos que não atenderam aos critérios mínimos de qualidade definidos neste trabalho resultou em uma extração com tópicos mais coerentes semanticamente do que a extração que não sofreu remoção de documentos instáveis, atendendo ao objetivo inicial.

Já para a definição do número de tópicos, as medidas de estabilidade e coerência semântica foram exploradas em uma base de dados anotada para assim avaliar se uma combinação das duas técnicas constitui um método para definição do número de tópicos em bases não-anotadas.

Estas medidas mostraram-se úteis para diminuir o intervalo possível de escolha do número de tópicos, mas não foi encontrada uma forma de conjugar estabilidade e

coerência com a escolha do número de tópicos de forma a obter uma extração mais detalhista ou generalista.

1.4 Organização do documento

O capítulo 2 apresenta o cenário do jornalismo na Era Digital, além de introduzir a modelagem de tópicos e delimitar o escopo deste trabalho. No capítulo 3, são apresentadas as técnicas de modelagem de tópicos. Já o Capítulo 4 descreve o método proposto. Por fim, os resultados do trabalho são apresentados no Capítulo 5, enquanto o Capítulo 6 descreve as conclusões obtidas e os trabalhos futuros.

Capítulo 2

Jornalismo na era digital

Em 59 AC, o imperador de Roma, Júlio César, mandou expor grandes placas para informar os súditos sobre acontecimentos políticos e sociais [20]. Ainda não existia o conceito de imprensa, mas ali surgia o primeiro jornal.

Desde então, a forma como a população é informada pela mídia experimentou diversas transformações. A prensa só seria inventada no século XV, inaugurando a era dos jornais impressos, enquanto a invenção do telégrafo, já no século XIX, seria responsável por maior rapidez na propagação de informação [20].

No entanto, no século XX houve um novo ponto de inflexão em relação à velocidade do tráfego de informações, com a invenção do rádio e da televisão [20]. Embora a *internet* tenha sido inventada também no século XX, a sua popularização e seus efeitos na forma como se consome informação seria experimentada em maior escala apenas no século seguinte [21].

As transformações vividas neste século foram intensas e velozes. A rede mundial de computadores deu voz às pessoas físicas em um mercado até então formado majoritariamente por instituições consolidadas. Que, ao ver o crescimento do número de pessoas conectadas, logo reagiram à demanda por informação *online* com uma frente digital [3].

Com o advento dos dispositivos móveis e das redes sociais, as pessoas passaram a ficar conectadas por mais tempo e em qualquer lugar. Isso aumentou não apenas o volume, mas também a velocidade de propagação da informação, além de democratizar ainda mais a geração de conteúdo.

O volume de informação e a velocidade com que ela se propaga não tem precedentes [2]. Além disso, a democratização dos meios de publicação causada pela invenção e popularização da *internet* possibilitou a propagação de notícias de fontes à margem da imprensa tradicional, por vezes sem credibilidade.

Atualmente é comum encontrar grupos criados especificamente com o propósito de distorcer os fatos, propagando as denominadas *fake news*, assunto que ganhou particular relevância em 2016, quando influenciou diretamente no resultado da elei-

ção norte americana [8] e, mais recentemente, a eleição a presidente do Brasil [22].

O cenário de alto volume de informação, aliado à velocidade de propagação e enorme variedade de fontes, que não são necessariamente críveis, motiva o desenvolvimento de algoritmos computacionais com foco no processamento automático destas notícias.

2.1 Adaptação da mídia impressa

Se nos primórdios da *internet* o acesso era restrito ao computador pessoal, hoje grande parte dos dispositivos eletrônicos possuem sistema embarcado conectados à *web*, com destaque especial para os celulares [23], que têm moldado a indústria digital por permitir que as pessoas estejam constantemente conectadas, em praticamente qualquer lugar.

Com o surgimento desse público, houve uma explosão na demanda por conteúdo *online*, o que impulsionou o surgimento de rádios, jornais e outros meios exclusivamente digitais. Alguns veículos que já possuíam uma plataforma física de publicação entenderam precocemente esta dinâmica e migraram seu conteúdo.

Antes da *internet*, o monopólio da informação estava nas mãos de alguns veículos de comunicação. O surgimento de plataformas como *blogs* e redes sociais democratizaram a geração de conteúdo *online* e, posteriormente, com o aumento de dispositivos móveis cada cidadão pode ser repórter cinematográfico ao testemunhar um evento. Esta dinâmica alterou a forma como a humanidade produz e consome conteúdo [5, 6].

Atualmente, a mídia sofre uma perda de credibilidade generalizada. Com a criação das bolhas das redes sociais, os usuários experimentam um aumento na popularização do conteúdo consumido a partir da posição que ocupam no espectro político, ocasionando descrença em veículos que seguem uma linha editorial do lado oposto deste espectro [24, 25]. Além disso, o fenômeno das notícias falsas abala a confiança no conteúdo publicado [9, 11].

A situação descrita acima mostra o problema de consumir o conteúdo jornalístico *online* nos dias de hoje: além da velocidade e do volume com que as notícias se propagam, ainda há a necessidade de conferir se o que foi publicado pode ou não ser considerado como verdadeiro.

2.1.1 Desenvolvimento tecnológico

Enquanto há três décadas o número de usuários conectados não era relevante, hoje a realidade é completamente diferente. Em números, o crescimento do número de usuários *online* passou de pouco mais de 14 milhões em 1993, para mais de 3

bilhões [1] conectados atualmente.

Além do número de pessoas conectadas, é preciso analisar também a transformação da forma como elas consomem as notícias. Já em 2008, foi constatado que o número de cidadãos norte-americanos que preferem buscar informações na internet havia superado os que preferiam ler jornais impressos [7] e já representava a segunda fonte de informação do país, perdendo apenas para a televisão.

Em 2017, 50% dos americanos adultos se informavam por meio da televisão, enquanto 43% se informavam *online* [5]. Estes números são ainda mais enfáticos para o público entre 18 e 29 anos, 23% e 52%, respectivamente. O número de telespectadores de telejornais vem diminuindo gradativamente [6] e, a julgar pela preferência dos jovens, essa deve ser a tendência para os próximos anos. Além do aumento do número de pessoas conectadas, elas se conectam por mais tempo, devido à evolução dos aparelhos celulares [23].

Esses fatores somados levaram a uma priorização, por parte dos jornais, pelo meio de publicação digital e, hoje, pode-se dizer que os principais jornais norte-americanos possuem seu conteúdo publicado na *web*. A mesma tendência pode ser observada no Brasil, apesar da taxa per capita de acesso à internet ser menor que a norte-americana [7]. Hoje, os principais jornais em circulação no país, como O Globo, Folha de São Paulo, entre outros, contam com planos de assinatura digital.

2.1.2 História do Jornalismo Digital

O começo da história do jornalismo *online* remonta a 1980, quando uma das maiores companhias de serviços de rede, a *CompusServe* [26], realizou um experimento que consistia em colocar alguns dos principais jornais dos Estados Unidos disponíveis em redes de *intranet*. O primeiro deles a ir ao ar foi o *The Columbus Dispatch*, em 1 de Julho de 1980. O experimento também contou com gigantes da comunicação, como *The New York Times* e *The Washington Post*. Nessa época, ainda não existia a *web*, que só seria inventada em 1989, por Sir Tim Berners-Lee, no CERN [21].

Já no Brasil, a história começa com o Jornal do Brasil (JB), jornal carioca fundado em 1891. Em 1995, o JB lançou sua versão *online*, produzida em um PC 386. Sendo, assim, o primeiro jornal brasileiro a participar da *web* (somente em 1996 o jornal O Globo entraria na rede mundial de computadores) [4, 27, 28]. Nessa época, o número de internautas no Brasil era estimado em 30 mil usuários. Mas, mesmo antes disso, em 1993, o JB já deixava evidente seu pioneirismo digital, possuindo um terminal de envio de mensagens sobre economia e política para a Bolsa de Valores do Rio em sua redação [27]. Desde 1 de Setembro de 2010, após 119 anos de publicação e perdendo volume de vendas, ele abandonou a mídia impressa e migrou para o meio digital.

Atualmente, o jornalismo vive uma nova fase. Com a explosão dos dispositivos móveis e redes sociais, o jornalismo vem exibindo um forte componente colaborativo entre a redação e o leitor-repórter [29, 30]. Este fenômeno é relativamente recente e pôde ser observado durante o atentado do dia 11 de Setembro de 2001. À época, as câmeras digitais já eram populares e os jornais não foram capazes de processar o alto volume de dados de imagens e vídeos gravados pelas testemunhas e enviados à redação [31].

Foi então que os jornais entenderam a mudança de cenário e investiram em capacidade técnica. Como consequência disso, os ataques terroristas em Madri, em Março de 2004, e em Londres, em Julho de 2005, foram cobertos com sucesso e desde então a cobertura de grandes incidentes tem sido marcada por informações em tempo real, fruto desta nova dinâmica colaborativa [31].

2.1.3 Busca por notícias

As ferramentas de busca de conteúdo na *web* usam palavras-chave para encontrar os documentos, adotando critérios de relevância para ordenar os resultados. Embora este funcionamento seja perfeitamente razoável para encontrar parte das informações, a busca por notícias envolve uma maior complexidade, visto a dinâmica temporal e o relacionamento semântico dos documentos. Dessa forma, novas ferramentas fazem-se necessárias [32, 33].

Algumas funcionalidades desejadas na busca por notícias, seriam:

- Credibilidade da fonte: o critério de relevância adotado pelas ferramentas de busca geralmente levam em consideração o quanto certo documento é referenciado por outros documentos. Isto não significa necessariamente que certo documento foi feito por uma fonte crível.

- Tópico: o que confere sentido a uma palavra é o contexto em que ela está inserida. Uma forma de contornar a ambiguidade inerente às palavras-chave frequentemente adotada pelos usuários das ferramentas de busca é digitar outras palavras que delimitem o contexto, de forma a guiar a busca. Para notícias, isso poderia ser feito de maneira mais criteriosa, escolhendo o tópico à qual a notícia pertenceria, por exemplo.

- Raio de proximidade semântica: os resultados de uma ferramenta de busca tradicional geralmente envolvem variações diretas da busca pela palavra-chave, com pouca margem para conteúdo relacionado. Uma forma de se contornar isto é definir um raio de conteúdo relacionado ao que se deseja, possibilitando extrapolar o conteúdo pesquisado para outros assuntos.

- Outros idiomas: a busca por palavra-chave é limitada pelo idioma das palavras que se está buscando. Para procurar por repercussão internacional de uma certa notícia é necessária a busca por outros idiomas.

- Entidades: caso o objetivo da pesquisa seja um personagem, o uso do seu nome como palavra-chave pode não recuperar os resultados desejados. Como um exemplo, considere a palavra "presidente". A cada eleição a entidade referenciada por ela pode sofrer uma mudança, e uma pesquisa por "presidente"provavelmente recuperará todos os resultados.

- Sentimento associado: as notícias podem possuir conteúdo positivo, negativo ou neutro sobre um determinado tema [34]. As ferramentas atuais de busca não permitem filtrar por este parâmetro, o que dificulta a busca por repercussões positivas e negativas de um certo acontecimento, por exemplo.

- Evolução dos fatos: algumas notícias referem-se a eventos que ocorreram anteriormente. A busca por palavras-chave estabelecem critérios de relevância que tendem a favorecer itens mais recentes, deixando os eventos antigos em segundo plano. Idealmente este é um parâmetro a ser configurado em uma ferramenta de busca de notícias.

2.2 Modelagem de Tópicos

Como evidenciado ao longo do capítulo, o cenário atual é de alto volume de informação publicada, aliado à velocidade de publicação, e com uma grande variedade de possíveis fontes. Some-se a isso, têm-se ainda a perda de credibilidade das fontes e a probabilidade de que determinada informação seja falsa. Além disso, as ferramentas de busca tradicionais não provêm parâmetros suficientes para adaptar-se à complexidade deste cenário.

Assim, é necessário o desenvolvimento de soluções computacionais para o processamento automático dessas publicações, de forma a originar novas ferramentas para exploração desses conjuntos de dados. Modelagem de tópicos refere-se ao conjunto de técnicas que têm por objetivo descobrir a estrutura temática em um conjunto de documentos textuais e classificá-los em grupos similares semanticamente, de forma a categorizá-los para o leitor [13, 35, 36].

Capítulo 3

Revisão bibliográfica

A modelagem de tópicos procura agrupar documentos por similaridade semântica. Para tanto, assume que os documentos são distribuições de tópicos e que um tópico é uma distribuição de palavras. O modelo, então, estima estas distribuições a partir da estatística do *corpus* [37].

Neste trabalho, a base de dados utilizada será formada por notícias jornalísticas. Como as notícias encontram-se dispersas nos bancos de dados de cada fonte, é necessário o desenvolvimento de um agente responsável por coletar o conteúdo textual delas através dos portais *web* em que estão hospedadas. Este agente chama-se *web crawler* [15].

Essa coleta de dados é uma tarefa complexa e sujeita a erros, devido às diferentes estruturas de marcação encontradas em diferentes fontes e, por isso, é natural a presença de ruído nestes dados [18]. Durante a revisão bibliográfica feita para este trabalho, não foram encontrados estudos que abordassem a modelagem de tópicos a partir de uma base de dados ruidosos. Isso ocorre porque, em geral, os trabalhos que apresentam resultados práticos recorrem a bases anotadas, com alta relação sinal-ruído [38].

3.1 Efeitos do ruído no agrupamento textual

Agarwal et al [39] avaliaram o efeito da presença de ruído no agrupamento de dois conjuntos de dados anotados, 20 newsgroups e Reuters-21578. Para isso, foram introduzidos erros ortográficos seguindo um padrão e um nível de ruído, responsável pela probabilidade de alterar ou não a grafia de uma palavra. Variando o nível de ruído de 10% em 10%, e avaliando a acurácia de dois modelos supervisionados, Máquina de Vetor Suporte (SVM) [40] e *Naive Bayes* (NB) [41], o resultado do trabalho mostrou que, mesmo com 70% de ruído, ou seja, 7 em cada 10 palavras escritas de maneira errada, não há queda significativa na acurácia da classificação. Além disso, avaliou-se que, quanto maior o vocabulário utilizado, maior a acurácia

do modelo. Este resultado dá a entender que o modelo aprendeu o padrão do ruído.

Para o contexto deste trabalho, uma análise visual dos documentos mostra que a introdução de ruído mais comum nos documentos que compõem a base de dados acontece por falhas na identificação do conteúdo principal da notícia por parte, dentro da estrutura HTML da página. Isso leva a introdução de trechos de código JavaScript, CSS e marcações HTML. Portanto, é pouco provável que o ruído tratado neste trabalho apresente um padrão, o que pode levar a resultados diferentes do encontrado em [39].

Outros trabalhos que tratam de ruídos em base de dados textuais [42, 43] abordam problemas diferentes do encontrado neste trabalho e não servem como referência.

3.2 Qualidade de dados

Os dados possuem características multi-dimensionais, onde cada dimensão representa um aspecto de qualidade. Estas dimensões são definidas de acordo com o contexto e problema abordado e a mesma dimensão pode assumir diferentes definições em problemas distintos [44]. Por exemplo, em um determinado contexto a dimensão de completude pode corresponder a informação binária, se existe ou não o dado. Já em outro contexto, esta dimensão pode refletir a quantidade de dado disponível.

Pesquisadores também apresentam formas diferentes de classificar as dimensões de qualidade, sendo a mais difundida as classificações objetiva, ou seja, o que se pode medir, e subjetiva, para dimensões com percepção qualitativa de qualidade. Assim como a mesma dimensão pode apresentar definições distintas em diferentes contextos, a sua classificação também pode mudar [44].

A área encontra grande aplicabilidade na indústria [45], onde, no contexto atual, as decisões devem ser baseadas em dados [46] e a má qualidade deles implica em perda de dinheiro e risco para os negócios [47].

Assim, como uma forma de melhorar o nível de qualidade dos dados, o processo a ser seguido tipicamente consiste no fluxo de: delimitar as dimensões propostas na literatura para o contexto tratado; identificar as dimensões objetivas; elaborar formas de medir as dimensões; monitorar as medidas de qualidade e melhorar os resultados.

Diferentes *frameworks* de melhoria da qualidade de dados foram propostos. O *Total quality data management* (TQDM) [48] é o programa pioneiro e, mais recentemente, a ISO 8000 [49].

3.2.1 Delimitação de contexto

No estudo da bibliografia não foram encontrados trabalhos que abordassem os problemas de qualidade que podem ser vistos na base de dados de notícias utilizada neste trabalho para fazer as análises. Assim, primeiramente foram estudadas medidas gerais de qualidade de dados, sem delimitação de contexto. A referência consultada foi Batini 2016 [50], onde o autor define *clusters* de qualidade, que agrupam medidas similares.

As principais dimensões estudadas foram as de medidas que levam em consideração a natureza dos dados textuais e dados provenientes de uma extração com um fluxo de dados, detalhados a seguir.

Acurácia: definida pelo autor como a proximidade entre o dado analisado e o dado verdadeiro. Como o dado verdadeiro pode mudar com o tempo, este grupo divide-se em dois: a acurácia estrutural, onde é considerada uma janela de tempo em que o valor permanece como verdadeiro, e a acurácia temporal.

Este primeiro subgrupo subdivide-se em outros dois: sintática, léxica. Por acurácia sintática entende-se como o respeito pelas regras sintáticas da linguagem. Já o segundo trata da presença de uma palavra no dicionário.

O segundo subgrupo é particularmente importante no domínio deste trabalho. Notícias são altamente voláteis, tornando-se menos relevantes à medida que novas matérias são publicadas. Este componente ainda é variável dependendo do objetivo da análise. Imagine que o objetivo seja agrupar todas as notícias referentes a um assunto histórico. Pela natureza do problema, é natural imaginar que o peso de publicações mais antigas dentro do grupo analisado deve ser o mesmo das publicações mais recentes. Este cenário muda quando o objetivo da análise é sumarizar as notícias para informar o leitor. Neste caso, em geral, as notícias mais recentes têm maior relevância.

A acurácia temporal ainda necessita lidar com a mudança de domínio inerente à evolução dos fatos. Imagine que as notícias a serem analisadas estejam em um intervalo de tempo suficiente para que certos fatos mudem. Para isso, considere, por exemplo, que houve uma eleição presidencial entre as janelas de tempo sob análise. Com isso, o termo "presidente"presente nas notícias das janelas podem não referir-se mais à mesma pessoa. Nessa dimensão, há ainda a dinâmica da língua, que determina a modificação do dicionário de acordo com a época.

Consistência: este grupo de qualidade trata da coesão e coerência textuais. Entende-se coesão como a facilidade com que se passa a ideia desejada, enquanto coerência indica a capacidade do leitor de colocar em contexto aquilo que foi lido.

Completude: corresponde à extensão com que os dados são suficientes para a tarefa. No caso de textos, pode-se pensar se o texto escrito foi o suficiente para

passar a informação desejada, embora isso seja condicionado ao grau de detalhe em que a informação é necessária.

Define-se ainda acessibilidade como a facilidade de acesso aos dados desejados e legibilidade, como a facilidade de leitura do texto sob análise.

Por fim, a última dimensão corresponde à confiança nos dados. Esta é uma dimensão que ganha crescente importância no contexto da comunicação atual, onde a desconfiança nos veículos tradicionais da mídia vem se intensificando e veículos independentes publicam uma série de notícias falsas, que pela primeira vez na história podem ter influenciado o resultado de uma eleição [8, 22]. Tanto a propagação como a detecção automática de notícias falsas por modelo estatísticos é um assunto em aberto, mas ativamente explorado [9–11].

3.3 Pré-processamento

Para a aplicação de métodos computacionais em texto, os autores fazem uso do mesmo fluxo geral de pré-processamento, que será responsável por representar os textos em formato vetorial [36, 38, 51]. Primeiramente, os documentos são separados em suas menores unidades, os chamados *tokens* [52]. Em seguida, os *tokens* são filtrados por regras que podem variar de acordo com a aplicação. Faz-se então a opção por representar o *token* em sua forma original ou por uma redução ao seu lema ou radical. Por fim, é feita a representação numérica dessas unidades.

3.3.1 Tokenização

Para representar vetorialmente um documento textual, é necessário que antes se identifiquem os *tokens*, ou seja, as suas unidades básicas [52]. Um *token* pode ser uma palavra, mas não necessariamente o é. Por exemplo: *they'll* é representado por dois *tokens*: *they* e *ll*. Este exemplo mostra também que o processo de tokenização é específico para cada idioma.

Para que se possa identificar os *tokens*, é necessário anteriormente separar o texto em sentenças, o que é feito com o algoritmo Punkt [51], que mostra-se robusto a abreviaturas. Já o algoritmo de tokenização utilizado é o Peen Treebank [51], adequado para língua inglesa.

Identificação de n-grams

O número de *tokens* pode ser reduzido através da identificação de padrões de repetição, utilizando o modelo de frases [53]. Este modelo captura palavras que coocorrem frequentemente. Assim, enquanto na representação original "São Paulo"é representado por dois *tokens*, "São"e "Paulo", este modelo transforma a palavra em apenas

um: "São_Paulo".

Diferentemente da divisão do texto em *tokens*, esta etapa nem sempre é adotada.

Filtragem dos *tokens*

Outra etapa frequentemente adotada na literatura é a filtragem dos *tokens* pouco relevantes. A relevância está geralmente associada ao valor semântico agregado pela presença da palavra. Por exemplo, palavras como "que", "a", "o"e etc, estão presentes em grande parte das sentenças do texto e, por isso, são consideradas pouco relevantes. Esses termos são chamados de *stopwords* e formam listas pré-definidas, específicas para cada idioma. Assim como as *stopwords*, a pontuação do texto também é filtrada. Os trabalhos utilizados como referência implementam ambos os passos [36, 38, 51].

Além disso, pode-se definir outras regras específicas para o problema, como filtrar numerais, datas etc.

Transformação dos *tokens*

Após a etapa de filtragem, podem ser realizadas as etapas de lematização ou radiciação [51]. Ou seja, as palavras são reduzidas aos seus lemas e radicais, respectivamente, o que implicaria em uma redução de dimensionalidade da base de dados. Embora a redução de dimensionalidade costume melhorar o resultado do modelo de agrupamento [54], Schofield et al (2016) [55] avaliariam que a lematização e a radiciação prejudicam a estabilidade do modelo de tópicos e levam a tópicos menos coerentes.

O processo completo de tokenização pode ser visto na Figura 3.1.

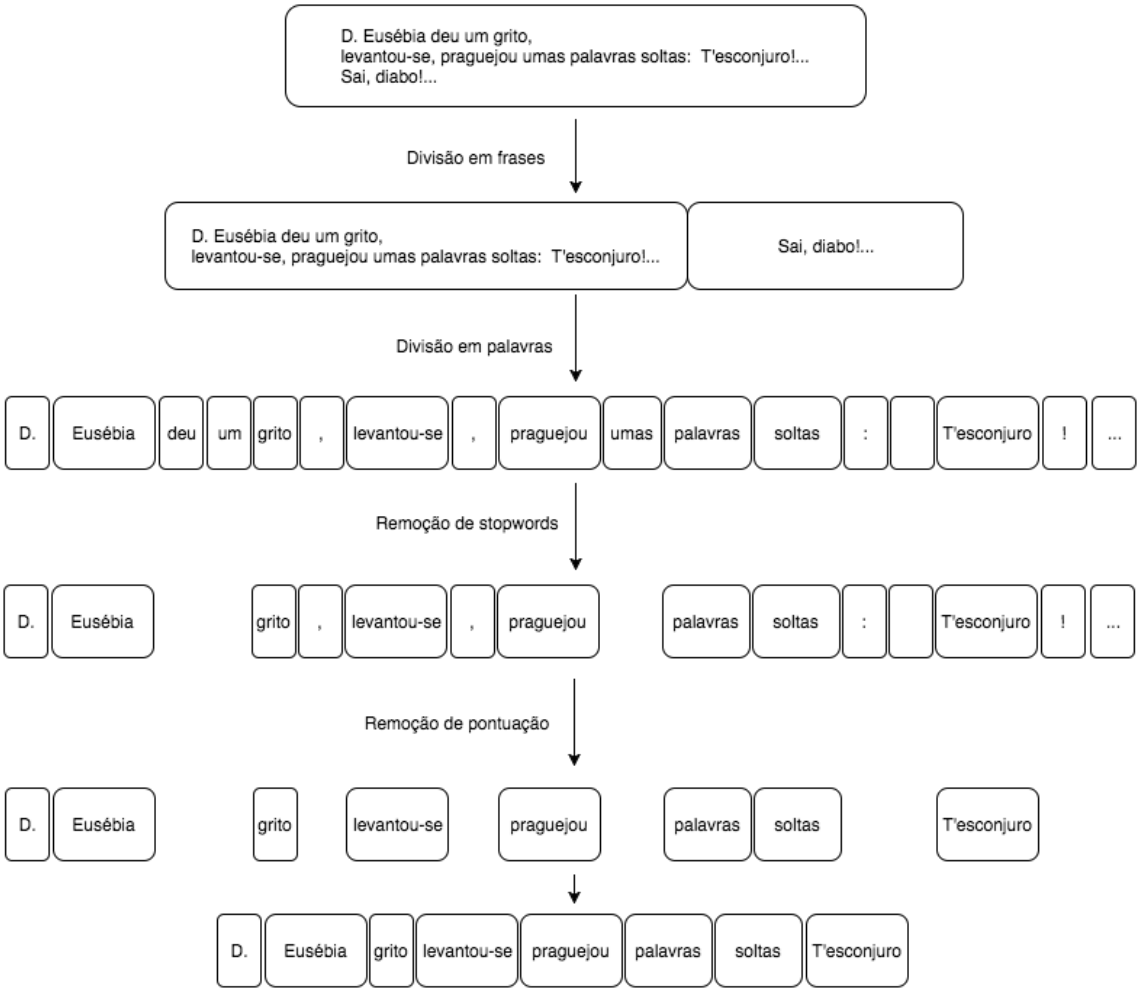


Figura 3.1: Ilustração do algoritmo de tokenização.

3.4 Representação vetorial

Os documentos precisam então ser transformados de texto para um espaço onde são representados algebricamente, o que é chamado de VSM (do inglês, *Vector Space Model*) [56]. Cada documento é transformado em um vetor onde os elementos são as representações algébricas dos *tokens*. Esta representação pode ser feita de diversas formas [57, 58].

Para Modelagem de Tópicos, duas abordagens são utilizadas: *bag of words* [59] e *term frequency, inverse document frequency* (tf-idf) [58].

Enquanto a primeira representação é apenas uma contagem da frequência de cada palavra no documento, a segunda tenta capturar a importância semântica do *token* para cada documento, ao ponderar a frequência com que cada *token* aparece no documento (*term frequency*) com a frequência com que cada termo aparece no *corpus*. Esta ponderação serve para equilibrar a importância do termo no documento com a importância do termo em toda a base, reduzindo assim termos muito

frequentes em todo o conjunto de documentos. Por exemplo, imagine que deseja-se modelar um conjunto de notícias que foram obtidas ao pesquisar pela mesma palavra-chave. É natural que esta palavra apareça em todas as notícias, mas ela não agrega informação para a separação de tópicos. A princípio este termo poderia ser adicionado à lista de *stop words* na etapa de filtragem. A representação tf-idf relativiza a importância desse tipo de termo de forma automática, sem que seja necessária uma análise manual acerca da presença desse tipo de ocorrência.

Matematicamente, uma base de documentos textuais, com n documentos e K *tokens*, é representada por uma matriz D de dimensão $n \times K$, onde cada posição dessa matriz corresponde ao peso tf-idf, dado por $tf_{ij} \times idf_i$, onde:

$$tf_{ij} = \frac{n_{ij}}{\sum_{k=1}^{K} n_{kj}} \tag{3.1}$$

$$idf_i = \log \frac{|D|}{|d : t_i \epsilon d|} \tag{3.2}$$

sendo n_{ij} o número de vezes que o termo i aparece no documento j e $|D|$ o número de documentos na base de dados.

Por fim, em [60] o autor mostra que a normalização L2 das matriz D melhora o resultado dos modelos de recuperação da informação. Esse procedimento ajusta a importância semântica dos termos para a mesma escala. Isso porque, caso um mesmo termo esteja presente em dois documentos de tamanhos distintos, a sua representação algébrica será maior para o documento de menor comprimento, uma vez que o fator *tf* é inversamente proporcional a esta variável. A normalização ajusta essa diferença.

Problemas com a representação

Cada coluna da matriz $D_{n \times K}$ é reservada para um termo, o que torna a representação de problemas textuais naturalmente esparsa. Outra característica é a perda da ordem original dos termos, o que evidencia a maior fragilidade da representação tf-idf. Uma vez que uma mesma palavra, que pode aparecer diversas vezes em um documento, é reduzida a apenas uma aparição por documento, os diferentes contextos em que ela era apresentada são perdidos.

Delimitação dos termos

A matriz de documento pode ter sua dimensão reduzida ao estabelecer critérios de corte para o vocabulário. Critérios frequentemente utilizados são a filtragem por número mínimo de ocorrências do termo no documentos (por exemplo, um termo em que aparece menos de 3 vezes em todos os documentos da base) e número máximo

de ocorrências, para evitar justamente termos que são comuns a grande parte da base e não agregam informação para o modelo de tópicos. Não foram encontrados trabalhos que abordassem como a mudança de vocabulário afeta o resultado da extração de tópicos.

3.5 Modelagem de tópicos

A Modelagem de Tópicos procura agrupar documentos textuais por estrutura temática, descobrindo a semântica escondida no *corpus* sob análise [13]. Um modelo de tópicos confere uma probabilidade de pertinência a um determinado tema para cada termo e documento presente na base de dados.

A literatura costuma dividir os modelos de tópicos em probabilísticos e não probabilísticos [61]. O primeiro grupo assume que os dados seguem uma determinada distribuição de probabilidade e, dado um processo gerador para os dados, os parâmetros da distribuição são estimados a partir da estatística do *corpus* [36]. Já o segundo grupo é reservado para os modelos de fatoração de matrizes, onde os dados são agrupados através da minimização de uma função custo [62].

A Alocação Latente de Dirichlet (LDA) [36] é a principal representante do primeiro grupo. Este modelo conta com variações [63, 64], e foi originado a partir da Análise de Probabilidade Latente Semântica (pLSA) [65]. Já no segundo grupo, dos modelos não probabilísticos, encontra-se a Fatoração de Matrizes Não-negativas (NMF) [66, 67]. Embora pLSA e NFM modelem o problema de maneira distinta, ambos os métodos se equivalem [68], quando o segundo modelo minimiza a divergência de Kullback-Leiber (KL).

Em 1990 foi proposta a *Latent Semantic Analysis* (LSA) [69], como uma técnica cujo objetivo era a representação de conhecimento através da redução de dimensionalidade do conjunto de documentos decompondo a matriz documento-termos em valores singulares (SVD) [70]. Por consequência, os tópicos obtidos por este modelo eram assumidos como ortogonais entre si.

Como a LSA possuía a habilidade de extrair conteúdo pela associação de termos que ocorrem em contextos similares, este modelo era uma ferramenta poderosa para a área de *information retrieval* [71], onde passou a ser conhecido como *Latent Semantic Indexing* (LSI) [72]. Neste contexto, outros estudos mostravam a efetividade da LSI para capturar documentos semanticamente relacionados, além da possibilidade de tratar sinônimos [73],

Posteriormente, em 1999, foi apresentado o *Probabilistic Latent Semantic Indexing* (PLSI) [74], onde o autor modela as coocorrências entre documento e termo como uma mistura de distribuições multinomiais condicionalmente independentes. Este modelo não era mais limitado a tópicos ortogonais, mas possuía algumas limi-

tações: para documentos novos, ou seja, apresentados após o treinamento, não era possível determinar a probabilidade de pertinência a um tópico. Outra limitação era a quantidade de parâmetros a serem estimados, que crescia linearmente com o número de documentos, o que faz o modelo ser propenso a *overfitting* [36].

Em 2002, foi proposta a *Latent Dirichlet allocation* (LDA) [36], uma generalização para o PLSI que resolvia as duas limitações mencionadas anteriormente. Esse modelo tornou-se popular e talvez seja o mais utilizado até hoje e, desde então, diversas extensões vêm sendo desenvolvidas [63, 75–78].

Já a NMF é uma maneira esparsa e eficiente de representar sinais, imagens e dados em geral [62, 66] e apresenta-se como uma alternativa aos modelos probabilísticos. Esta técnica foi proposta em 1994 [79], onde o autor discute as possíveis aplicações da técnica para modelar fenômenos físicos, devido a restrição dos fatores serem não-negativos. No entanto, a aplicação da NMF em dados textuais foi discutida apenas em 2003 [80], onde o autor destaca a a maior flexibilidade do modelo em relação a fatoração SVD, que supõe que um documento pode pertencer a apenas um tópico. Esta técnica foi estendida para permitir correlação entre tópicos [81] e rastreabilidade temporal de tópicos [82].

3.5.1 Modelos de Tópicos

De uma maneira geral, dado um conjunto de documentos, os modelos de tópicos desejam chegar à seguinte fatoração [37]:

$$D_{documentos \times termos} = \Phi_{termos \times topicos} \Theta_{topicos \times documentos} \tag{3.3}$$

Para tanto, as duas principais famílias de modelos utilizadas, são: os modelos probabilísticos, cujo exemplo principal é o LDA e os modelos de fatoração de matrizes, com a NMF como principal modelo.

Alocação Latente de Dirichlet (LDA)

Esta técnica concebe um documento como uma mistura de tópicos e é melhor descrita pelo seu modelo gerador [83], o processo aleatório cujo modelo assume ser a origem dos documentos da base. Este processo consiste em:

- Escolha de um tópico da distribuição de tópicos de um documento.
- Escolha de uma palavra do tópico e associação da mesma ao documento.

O objetivo do modelo pode ser pensado como o processo gerador inverso, ou seja, encontrar a estrutura escondida que gerou a coleção de documentos observada [13].

Fatoração de Matrizes não Negativas (NMF)

Neste modelo, a matriz original de documento-termos é aproximada como produto de duas matrizes de dimensionalidade reduzida, ambas com elementos não-negativos: uma delas relaciona os documentos aos tópicos, enquanto outra relaciona os termos aos tópicos. Os elementos não-negativos conferem uma interpretabilidade direta da força de pertinência dos termos e documentos aos tópicos em questão.

Matematicamente, seja $V = (v_1, \ldots, v_n)$ uma matriz de entrada que contém uma coleção de n vetores-coluna de dados. Fatora-se V em duas matrizes:

$$V = WH + E \tag{3.4}$$

onde, $V \in \mathbb{R}^{p \times n}$, $W \in \mathbb{R}^{p \times k}$, $H \in \mathbb{R}^{k \times n}$ e E é o residual da aproximação. Nesta fatoração, os elementos assumem valores não-negativos.

Dessa forma, o modelo minimiza a função custo D:

$$\min_{W,H} D(V|WH) \tag{3.5}$$

sujeito a $W \geq 0$ e $H \geq 0$.

As funções custo mais populares são:

- Distância euclidiana

$$D(V|WH) = ||V - WH||^2 \tag{3.6}$$

- Divergência de Kullback-Leibler (KL)

$$D(V|WH) = -\sum_{i=1}^{m}\sum_{j=1}^{n}(V_{ij}(\log(\frac{(WH)_{ij}}{V_{ij}}) + 1) + (WH)_{ij}) \tag{3.7}$$

Em [84], o autor compara a performance de ambos os modelos supracitados e conclui que a NMF produz tópicos mais coerentes que a LDA. Ele atribui este resultado à menor escolha de parâmetros para o treinamento da NMF.

3.6 Estabilidade

A aplicação dos modelos mencionados acima mostrou diversos desafios, como a identificação do número de tópicos e a estabilidade e interpretabilidade dos resultados.

Estabilidade, no contexto de modelagem de tópicos, refere-se à consistente replicação de soluções similares no mesmo conjunto de dados. Ou seja, estimações de distribuições de probabilidade similares mesmo com diferentes inicializações do modelo [36, 85].

O número de tópicos pode ser interpretado como o grau de aproximação dos grupos semânticos. Quanto menor, mais gerais são esses grupos e mais documentos eles contêm. Conforme o número de tópicos aumenta, esses grupos vão se dividindo e os documentos se redistribuem. Assim, não é possível dizer que existe um número correto para a escolha do número de tópicos, mas observa-se que algumas dessas escolhas levam a agrupamentos instáveis.

No trabalho que originou a LDA [36], percebeu-se este problema e foi proposto calcular a similaridade entre duas extrações através da divergência KL entre as suas respectivas distribuições de tópicos.

Já Brunet et al (2004) [86] propuseram uma medida de estabilidade para extrações feitas a partir de modelos de fatoração de matrizes, que consistia em:

- Executar múltiplas extrações de tópicos com inicializações aleatórias do modelo.
- Montar a matriz de conectividade para cada extração, onde matriz de conectvidade é a matriz $M_{documentos \times documentos}$, composta por elementos M_{ij} que recebem o valor 1, caso os documentos compartilhem o mesmo tópico, e 0, caso contrário
- Obter a matriz de consenso, resultado da média entre as matrizes de conectividade.
- O nível de estabilidade desta extração é então obtido, através do coeficiente de correlação cofenético [87] calculado para esta matriz.

A interpretação segue da seguinte forma: quanto maiores os valores da matriz de consenso, maior a concordância entre as matrizes de conectividade, o que significa que os documentos compartilharam o mesmo grupo em grande parte das diferentes inicializações e mais estável é o *corpus*. Esta medida é aplicável apenas para fatoração de matrizes e não para modelos probabilísticos.

Já a medida de estabilidade proposta por Greene et al utiliza a distribuição das palavras dos tópicos [85] e é inovadora por poder ser aplicada tanto para modelos probabilísticos como para fatoração de matrizes, possibilitando a comparação de desempenho desses modelos. Além disso, foi desenvolvida de forma a não sofrer os efeitos das principais características dos conjuntos de dados textuais: alta esparsidade e dimensionalidade.

O método adotado em [85] segue da seguinte forma: para duas extrações de tópicos, com inicializações distintas, e para o mesmo número de tópicos, a primeira extração produz o conjunto S_1, de tópicos R_{11}, R_{12} *e* $R13$:

$R_{11} = sport,\ win,\ award$

Tabela 3.1: Ilustração do cálculo de similaridade entre dois tópicos

R_i	R_j	Jaccard	AJ
album	sport	0.000	0.000
album, music	sport, best	0.000	0.000
album, music, best	sport, best, win	0.200	0.067
album, music, best, award	sport, best, win, medal	0.143	0.086
album, music, best, award, win	sport, best, win, medal, award	0.429	0.154

$R_{12} = bank,\ finance,\ money$

$R_{13} = music,\ album,\ band$

Enquanto a segunda extração S_2 produz os tópicos $R_{21}, R_{22}\ e\ R23$.

$R_{21} = finance,\ bank,\ economy$

$R_{22} = music,\ band,\ award$

$R_{23} = win,\ sport,\ money$

Primeiramente é calculada a concordância entre as duas extrações, o que é feito utilizando a similaridade média de Jaccard (AJ), que consiste na média da similaridade de Jaccard entre os conjuntos completos, depois entre os conjuntos sem os respectivos primeiros elementos e assim por diante, até restar apenas o último elemento. Matematicamente, considere a medida AJ entre os tópicos R_{ia} e R_{ib} de t palavras como

$$\begin{aligned} AJ(R_{ia}, R_{ib}) &= \frac{1}{t}\sum_{d=1}^{t} Jaccard(R_{ia}, R_{ib}),\ onde \\ Jaccard(R_{ia}, R_{ib}) &= \frac{|R_{ia} \bigcap R_{ib}|}{|R_{ia} \bigcup R_{ib}|} \end{aligned} \tag{3.8}$$

Este processo é ilustrado na tabela 3.1, onde fica evidenciada a diferença entre a medida proposta e a similaridade de Jaccard. Enquanto a segunda não leva a ordem das palavras em consideração ao calcular a similaridade entre os grupos, a nova medida consegue levar em consideração o peso das palavras.

O processo é então repetido para todos os tópicos de ambas as extrações, resultando na matriz que compara cada tópico de uma extração a todos os demais da outra extração. Para os tópicos $R_{11}, R_{12}\ R13,\ R_{21}, R_{22}\ e\ R23$, a matriz resultante é dada por:

	R_{21}	R_{22}	R_{23}
R_{11}	0.00	0.07	0.50
R_{12}	0.50	0.00	0.07
R_{13}	0.00	0.61	0.00

onde cada elemento ij representa a similaridade de Jaccard média entre os tópicos i e j. Calcula-se então a concordância entre as duas extrações S_1 e S_2, que deram origem aos tópicos $R_{11}, R_{12}e\ R_{13}$ e $R_{21}, R_{22}e\ R_{23}$ como a média entre as maiores concordâncias entre tópicos linha-a-linha: $agree(S_1, S_2) = \frac{0.50+0.50+0.61}{3} = 0.54$. Finalmente, a estabilidade é a soma das concordâncias entre todas as extrações de tópicos com inicializações distintas, para um mesmo número de tópicos. Esta medida apresenta valor entre zero, quando não há nenhuma concordância entre as extrações, e um, quando as extrações encontram sempre os mesmos termos.

O algoritmo seguido para obter a nota final de estabilidade para um modelo de k tópicos é o seguinte:

- Gerar aleatoriamente τ amostras do conjunto de dados, cada uma com $\beta \times n$ documentos, onde β é o percentual de documentos que deseja-se na amostra.
- Para cada valor de k entre k_{min} e k_{max}:
  - Aplicar o modelo de tópicos no conjunto completo dos n documentos para gerar os k tópicos, e assim gerar o *ranking* de referência S_0.
  - Para cada amostrar X_i:
    - Aplicar o algoritmo à amostra para gerar k tópicos e representar a saída como o S_i.
    - Calcular a concordância entre S_0 e S_i.
  - Calcular a média entre as concordância para todas as amostras.

Nível de instabilidade

Apesar de haver um limite inferior teórico e igual a zero, não necessariamente uma extração precisaria atingir esse nível para ser considerada instável. Greene et al. encontram este nível gerando aleatoriamente documentos textuais que não estabelecem nenhuma relação entre si. O resultado obtido é um valor pouco menor que 0.2, revelando-se assim o valor limite da instabilidade.

De forma similar à coerência, esta medida não possui calibração. Assim, embora seja possível afirmar que uma extração é instável e uma extração produziu um resultado mais estável que a outra, não é possível afirmar o quão mais estável.

De fato, comparar extrações é um desafio. O processo seguido para medir a estabilidade de diferentes extrações consiste em, para uma mesma base, escolher um intervalo desejado para o número de tópicos e fazer a extração para cada um deles. No final, comparam-se os diferentes níveis de estabilidade obtidos. Em geral, o resultado apresenta uma tendência de queda do nível do estabilidade para um aumento do número de tópicos, sem necessariamente implicar que o nível de instabilidade

seja atingido. Isso não significa dizer automaticamente que quanto menos tópicos melhor. Significa apenas que modelos com um maior número de tópicos são mais complexos e vêm acompanhados de uma maior variação no resultado da extração, visto que a separação relativa entre os grupos semânticos é menor.

3.7 Coerência

Os modelos de tópicos utilizam a estatística presente no *corpus* para agrupar os documentos em grupos. Desde o advento da LSI, foi observado que estes grupos remetem a grupos de conteúdo semântico relacionado. Portanto, argumentou-se que esses modelos são capazes de agrupar semanticamente os documentos [72]. Embora isso tenha se evidenciado ao longos dos experimentos posteriores, não há garantia por parte desses modelos que o resultado da extração será coerente do ponto de vista semântico.

Diz-se que um tópico é coerente quando suas principais palavras fazem sentido para um avaliador humano. As medidas de coerência foram desenvolvidas visando avaliar o resultado final da extração de tópicos, a fim de garantir a interpretabilidade do resultado. Newmann et al (2010) [88] avaliaram diferentes medidas de coerência e avaliaram seu desempenho calculando a correlação entre os resultados dessas medidas com notas de coerência conferidas a pares de palavra por avaliadores humanos.

De uma maneira geral, as medidas de coerência procuram calcular:

$$coherence(V) = \sum_{(w_i,w_j)\epsilon V} score(w_i, w_j) \tag{3.9}$$

onde a função *score* confere uma nota de coerência para um par de palavras, ou a uma palavra e um conjunto de palavras vizinhas. O que difere entre as medidas é a forma como essas notas são calculadas e a base de dados utilizada para fazer o cálculo.

Uma dessas medidas é a *point-wise mutual information* (PMI), representada na equação 3.10.

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)} \tag{3.10}$$

Onde, $p(w_i)$ representa a probabilidade de ocorrência da palavra w_i, $p(w_i, w_j)$ representa a probabilidade de coocorrência e ϵ é uma constante para que o logarítimo não seja indefinido.

Dependendo da implementação, essas probabilidades podem ser estimadas em janelas móveis de diferentes tamanhos. O que também varia de acordo com a imple-

mentação é a base utilizada para estimar esses valores: se a própria base utilizada para a extração dos tópicos, ou uma base externa, como a Wikipedia [89].

Com base nas diferentes possibilidades, Röder et al (2015) [90] exploram o espaço de variações possíveis para as medidas propostas na literatura. O trabalho avaliou diversas possíveis combinações para as medidas de coerência, comparando a correlação destas medidas com o sentimento de avaliadores humanos em bases anotadas, assim como seus respectivos tempos de processamento. O resultado encontrado aponta a medida denominada por C_V como o melhor método para língua inglesa, obtendo correlação de 0.821 com o julgamento humano, quando a Wikipedia foi utilizada como base para a estimação de probabilidades. Para chegar a esse número, avaliadores humanos conferiram notas de qualidade para conjuntos de palavras, formando bases anotadas de coerência. Dessa forma, foram comparadas as classificações dos avaliadores com os resultados das medidas.

Uma das propriedades da coerência é que não existe um valor em que a classificação de um par de palavras muda de "coerente"para "incoerente". Estas medidas, normalmente, são utilizadas para comparar dois pares, afirmando se um é mais coerente que outro. Outra particularidade é que esta medida não possui calibração. Assim, além de não ser possível falar se um par de palavras é "coerente", também não é possível afirmar que um par de palavras é muito mais coerente que outro.

Outra propriedade é que nem sempre as medidas de coerência possuem limites inferior e superior. No entanto, a medida de melhor resultado, C_V, faz a estimação de coerência com similaridade de cosseno entre dois vetores com elementos entre 0 e 1, o que força o resultado a estar entre 0 e 1 [90].

3.8 Escolha do número de tópicos

O número de grupos em uma base textual tem como intuição natural funcionar como o nível de detalhamento dos temas. A escolha por um número pequeno de temas leva ao agrupamento em temas gerais, enquanto um aumento do número de tópicos leva a um detalhamento maior desses temas [13].

O problema da escolha do número de grupos é recorrente na literatura [62, 85, 91, 92] e diversas medidas foram propostas para a escolha ótima deste número, como a silhueta [93], Davies-Bouldin [94], entre outras [95–97]. Especificamente, para fatoração de matrizes não-negativas, um método bastante difundido é analisar a variação da soma dos quadrados residuais (RSS) para cada número de *clusters* e escolher o ponto de inflexão como o número ótimo de grupos [85]. No entanto, a alta dimensionalidade das bases textuais invalida o uso de todas estas medidas.

Por isso, as medidas de estabilidade e coerência apresentam-se como alternativas importantes. Por medirem o resultado da extração, especificamente as palavras dos

tópicos encontrados, ambas as técnicas são independentes do modelo e não afetadas pela alta dimensionalidade das bases textuais.

3.9 Visualização

A visualização de bases de dados textuais não é uma tarefa simples, devido à alta dimensionalidade do problema. Foram desenvolvidas formas de reduzir a dimensionalidade do *corpus* para visualização em duas dimensões, como a aplicação de Análise de Componentes Principais (PCA) [98, 99] e o desenvolvimento de uma técnica chamada *t-Distributed Stochastic Neighbor Embedding* (t-SNE) [100]. Esta última é o estado da arte e é a maneira mais difundida de visualizar dados multidimensionais.

Recentemente, uma nova técnica apresentou resultados competitivos com o t-SNE, mas com performance computacional muito superior: a *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP) [101].

3.10 Outros desafios

Existem ainda outros desafios em aberto, mas que não serão tratados aqui neste trabalho. Um deles é atribuir uma descrição curta ao tópico. Como explicado anteriormente, cada tema encontrado fica associado a uma distribuição de termos ordenados de forma decrescente de probabilidade de pertinência ao tópico. Como essa distribuição é sobre todos os termos do *corpus*, tipicamente na ordem de milhares, cada tema deveria receber uma descrição breve com base nas principais palavras, para facilitar a comunicação do resultado do modelo de tópicos.

Outro desafio é o rastreamento temporal dos tópicos. Como as notícias são dados dinâmicos, uma característica desejável é relacionar os tópicos ao longo do tempo. Em [82], este assunto é explorado e a NMF é estendida para estimar a relação entre sucessivas execuções.

Capítulo 4

Método Proposto

O método proposto neste trabalho consiste em: seleção das notícias, com posterior análise de qualidade de dados para que seja possível eliminar aquelas que não atendem aos critérios mínimos estabelecidos; pré-processamento e representação vetorial dos documentos; escolha do número de tópicos e análise do resultado. Cada etapa será detalhada a seguir.

4.1 Seleção dos documentos

Os documentos são recuperados do banco de dados que contém os documentos extraídos dos portais de notícias pelos *web crawlers*, a partir da busca por palavra-chave.

É comum a compra de notícias escritas por agências [102], o que faz com que a mesma notícia seja replicada por múltiplos jornais. Essas notícias atuam como ruído na base de dados, enviesando o resultado ao conferir maior importância para documentos e termos duplicados.

Por isso, após a recuperação das notícias, a primeira etapa será a filtragem por notícias duplicadas, o que não pode ser feito apenas comparando o conteúdo exato entre as notícias. Isso porque, por vezes, a estrutura do documento *web* de uma das fontes é mais complexa e o *web crawler* tem dificuldade de recuperá-la integralmente. Assim, não basta eliminar documentos com o mesmo conteúdo.

O método adotado neste trabalho para filtragem dos documentos é a transformação *tf-idf* do *corpus* e posterior cálculo da similaridade entre os documentos pela similaridade de cosseno:

$$\cos\Theta = \frac{\vec{a}.\vec{b}}{\|\vec{a}\| \left\|\vec{b}\right\|} \tag{4.1}$$

onde a e b são as representações vetoriais dos documentos. Documentos cuja similaridade seja maior do que 0.9, valor definido empiricamente a partir de inspeção

visual do resultado, são candidatos a eliminação. Os documentos duplicados são, então, agrupados e o algoritmo remove o documento de menor comprimento.

4.2 Análise da qualidade dos dados

Extrair automaticamente as notícias dos portais em que estão hospedadas, preservando seu conteúdo original e excluindo toda a sorte de informação não relevante, como propagandas, links relacionados, e outras seções alheais ao conteúdo, é uma tarefa complexa e sujeita a erros.

Assim como as notícias duplicadas, notícias com problemas na estrutura semântica ou sintática também atuam como ruído na base de dados. Por isso, é necessária uma validação do resultado gerado por esses agentes.

Este trabalho confere uma dimensão de qualidade às páginas extraídas ao calcular o nível de ruído presente nas notícias. Ao definir e medir diferentes dimensões de qualidade é possível remover as notícias que não atendem a critérios mínimos e, assim, introduzem ruído na base de dados.

4.2.1 Origem dos problemas de qualidade

A partir de uma análise visual da base de dados, é possível observar que, praticamente a totalidade dos problemas de qualidade dos dados ocorrem quando o agente tem dificuldade de detectar os limites do conteúdo da notícia. Em geral, isso ocorre devido às mudanças na estrutura de marcação da página *web*, o que frequentemente leva o conteúdo extraído a conter trechos de código e seções de conteúdo recomendado.

Ambos os problemas impactam no resultado da análise. O primeiro aumenta a dimensionalidade do problema sem acrescentar nenhuma informação semântica, atuando como puro ruído. Já o segundo aumenta a informação semântica de forma indevida, visto que frequentemente o conteúdo recomendado pertenceria a outro tópico e a presença das palavras do título deste segundo documento atua como um viés para a classificação.

Outro problema comum são as notícias curtas, geralmente notas publicadas nos jornais. Como os modelos de tópicos são em geral complexos, textos curtos não possuem estatística suficiente para a estimação de seus parâmetros, atuando assim como ruído na base. Existem trabalhos desenvolvidos especificamente para textos curtos, cuja proposta é a simplificação dos modelos de tópicos, reduzindo o número de parâmetros necessários [103]. Neste trabalho, os textos curtos serão excluídos da análise.

4.2.2 Aplicação das definições ao contexto

A abordagem adotada neste trabalho foi delimitar as dimensões de qualidade propostas em [50] para o contexto da aplicação tratada. O autor define as seguintes dimensões de qualidade: acurácia, que por sua vez é divida em estrutural e temporal; consistência; acessibilidade; completude; legibilidade; e confiança nos dados.

Embora as dimensões de qualidade acima tratem de documentos textuais, este trabalho situa-se no contexto de notícias extraídas de forma automática por *web crawlers*. Assim, é necessária a adaptação das definições acima para definir medidas de qualidade objetivas.

- Acurácia estrutural

No contexto deste trabalho, faria sentido *a priori* supor acurácias sintática e léxica como completas, assim como a consistência, haja vista que os dados são notícias jornalísticas. No entanto, como mencionado anteriormente, a extração automática destes documentos está sujeita a falhas, tornando razoável supor que pode haver introdução de ruído na extração.

Como este ruído geralmente apresenta-se sob a forma de trechos de marcação HTML, CSS e JavaScript, foi treinado um classificador de k-vizinhos [104] para identificação da presença desses trechos no corpo da notícia. Para o treinamento, foram selecionadas 500 notícias com presença de trechos de código e 500 notícias apenas com o texto da notícia. Foram utilizadas como características a contagem de palavras reservadas de JavaScript e CSS [105, 106], e a contagem de pontuação (ponto e vírgula e chaves). O modelo apresentou $98 \pm 0.2\%$ de acurácia nos 5 *folds* de treinamento, e 97% de acurácia no conjunto de teste, composto por 200 documentos (100 de cada classe) que não participaram dos *folds* de treinamento.

Assim, caso o modelo identifique a presença de trechos de código, a nota de acurácia do documento é automaticamente definida como zero. Visto que esse modelo pode apresentar falsos negativos, uma segunda forma de identificar trechos de código é procurar pelas palavras do documento no dicionário. Esta forma é eficaz, visto que é computacionalmente barata, e razoavelmente efetiva em identificar o tipo de problema tratado.

Uma outra alternativa seria calcular a coerência semântica para trechos do documentos e assim encontrar trechos de baixo valor. Mas, como será visto mais adiante, os modelos de coerência são computacionalmente pesados e sujeitos a falhas de classificação.

Por isso, a nota final de acurácia segue da seguinte forma:

$$acuracia = \min(nota\ do\ modelo, 1 - \frac{quantidade\ de\ palavras\ erradas}{quantidade\ total\ de\ palavras}) \quad (4.2)$$

onde a nota de modelo é zero, caso um trecho de código seja identificado no documento e, um, caso contrário. Esta é uma medida potencialmente drástica, uma vez que o trecho de código pode ser pequeno frente ao tamanho do texto. Mas como a avaliação deste tipo de ruído não é o objetivo deste trabalho e não é possível avaliar o tamanho do impacto de deixar estes documentos na base, levar a nota de acurácia a zero é uma forma de eliminar o potencial problema.

Mesmo que o documento não apresente trechos de código, é esperado que a nota de acurácia raramente seja igual a 1, principalmente devido a neologismos e estrangeirismos. Como este trabalho situa-se no contexto de notícias, e não documentos históricos, não será aplicada nenhuma correção para nota de acurácia, considerando a evolução do dicionário ao longo do tempo.

Outro problema comum na extração das matérias é a publicação de notícias em outros idiomas, diferentes daqueles que se espera do jornal. Isso ocorre geralmente para jornais com um perfil mais internacional, que adotam o inglês como língua oficial, mas correspondentes podem escrever no idioma local. Nesse caso, foi utilizado o classificador Naive Bayes do projeto *Compact Language Detector 2* [107], para verificar se o idioma em que o texto foi escrito corresponde ao idioma da análise.

- Acurácia temporal

Os modelos de tópicos usam a matriz de documentos por termos para encontrar os grupos semânticos, sem levar a dimensão temporal em consideração. No entanto, o contexto das notícias tende a evoluir com o tempo. Alguns trabalhos tentam estender estes modelos para relacionar como os tópicos evoluem no tempo [82].

Apesar de importante para o leitor, não é trivial medir a acurácia temporal de uma maneira uniforme. Por exemplo, considere uma notícia sobre a divulgação de um evento. Neste caso, o valor das notícias publicadas após o eventos ter acontecido diminui. Já para outro assunto, pode ser que a notícia seja relevante por mais tempo. Em ambos os casos, não pode-se dizer que a notícia perde totalmente seu valor. Desta forma, esta dimensão não será tratada como medida objetiva neste trabalho. Em vez disso, pode-se analisar de maneira subjetiva a acurácia temporal ao fim da extração.

Vamos adotar acurácia como sinônimo de acurácia estrutural ao longo do texto.

- Consistência

No grupo de consistência, pode-se pensar na dimensão de coesão como o poder de comunicação de fonte. Jornais cujo objetivo é escrever para classes C e D possuem mais facilidade de leitura e tendem a deixar mais clara a mensagem a se passar, ou seja, tem melhor legibilidade e coesão.

A coesão refere-se à forma como o raciocínio é construído de forma a passar a mensagem e não necessariamente existe uma forma melhor do que a outra [50]. Leitores diferentes podem discordar sobre qual construção é mais coesa. Assim, coesão é uma medida subjetiva.

Já para a Legibilidade foram propostas algumas medidas na literatura [108]. Uma delas, o índice de legibilidade Gunning fog, possui a seguinte expressão:

$$0.4 \left\lfloor \left(\frac{quantidade\ de\ palavras}{quantidade\ de\ sentencas} \right) + 100 \left(\frac{quantidade\ de\ palavras\ complexas}{quantidade\ de\ palavras} \right) \right\rfloor \quad (4.3)$$

onde as palavras complexas são aquelas com três sílabas ou mais.

Embora essa seja uma dimensão que abre possibilidades de análise, como a iteração do leitor com fontes de acordo com o nível de legibilidade de cada uma delas, o que está sendo medido, neste caso, está no campo da qualidade da informação e não apresenta utilidade imediata para identificar problemas na extração dos documentos, que constitui o objetivo principal das medidas de qualidade definidas neste capítulo. Portanto, esta dimensão não será utilizada ao longo deste trabalho.

- Acessibilidade

Pode-se pensar na dimensão de acessibilidade no contexto de *web crawlers* como sendo a facilidade com que os dados estão disponíveis por meio de uma extração automática. Jornais que possuem controle de acesso não são acessíveis por esses extratores e assim não podem ser considerados acessíveis, por exemplo.

Não há necessidade de definição para esta dimensão, uma vez que, por construção, fontes inacessíveis serão ignoradas, enquanto fontes acessíveis serão extraídas por agentes mais ou menos complexos, de acordo com a estrutura do portal.

- Completude

Em [109], o autor constatou que o LDA falhava em agrupar textos curtos. Isso ocorre porque a elevada esparsidade desse tipo de texto não permite ao modelo inferir seus parâmetros corretamente durante o treinamento. Uma solução para este problema é reduzir a complexidade do modelo, como feito em [103]. No entanto, como a maioria dos documentos textuais tratados no contexto deste trabalho são textos longos, este tipo de modelo não será abordado. Para garantir que o resultado do agrupamento não seja afetado pela presença de textos curtos na base de dados, o documento que tiver menos do que 100 palavras, após passar pelo pré-processamento, receberá nota zero de completude. Este limite foi definido empiricamente e carece de maior análise.

- Confiança

Mais uma vez, esta dimensão não precisa ser definida, uma vez que, por construção, as fontes de notícias falsas não terão agentes associados. Caso o leitor não dê credibilidade a uma das fontes, poderá filtrá-la do resultado da análise.

Assim, com base na definição das medidas de qualidade, é possível criar critérios objetivos para os documentos capturados. Cada notícia extraída possui os atributos: data de publicação, idioma, título e corpo. Como o título não impacta no agrupamento dos documentos, ele é ignorado para efeito da análise de qualidade. Cada um desses atributos será medido por completude e acurácia.

4.3 Filtragem com base na qualidade dos dados

A implementação das medidas de qualidade será feita associando-se uma expressão matemática que deverá ter resultado entre zero, conferido à ausência total de qualidade, e um, quando a qualidade é completa.

Cada atributo receberá uma nota de qualidade. Dessa forma, de modo a combinar cada uma dessas notas para conferir uma única nota à notícia, é necessária uma ponderação das notas dos atributos.

Para isso, a data recebeu o peso 1, enquanto corpo e idioma receberam um peso igual a 5. Esses pesos foram definidos de maneira empírica, de modo a não permitir que problemas de completude da data, ou seja, a ausência de informação de horário em que a notícia foi publicada, possam enviesar o resultado de qualidade e passar a impressão errada que a extração possui pouca qualidade.

A ponderação das notas de qualidade por atributo para cada documento permite associar uma nota de qualidade para a notícia. Adotando critérios mínimos a serem atendidos, é possível então remover os documentos que não satisfaçam as condições de qualidade pré-estabelecidas, o que ajuda a reduzir o nível de ruído na base de dados.

Ainda assim, existem ruídos que não podem ser eliminados totalmente. Uma análise *ad-hoc* dos documentos mostra que um conjunto deles é contaminado por textos das propagandas presentes nas páginas dos jornais e também por manchetes de notícias que o jornal mostra como relacionadas à notícia.

Esse tipo de ruído pode impactar no resultado da extração, uma vez que as palavras contidas no título podem ter alta probabilidade de pertencer a outro tópico e, por isso, enviesar a associação do documento ao tópico errado.

4.4 Pré-processamento e Representação vetorial

Este passo segue o roteiro explicado no capítulo anterior. Inicialmente os documentos são pré-processados, sendo separados em sentenças e termos. O modelo de frases é utilizado para fazer a identificação de bigramas e trigramas frequentes. Por fim, os termos são então filtrados utilizando a lista pré-definida de *stopwords* para língua inglesa.

A representação algébrica das palavras é feita por uma transformação tf-idf que transforma o *corpus* em uma matriz $D_{documentos \times termos}$. São eliminados os termos que aparecem menos de 4 vezes em toda a base e em mais de 80% dos documentos. Dependendo do número de termos, pode-se ainda estabelecer um corte para limitar o tamanho máximo do vocabulário. Por fim, a matriz é normalizada para que os documentos apresentem comprimento unitário.

4.5 Escolha do número de tópicos

Para escolher o número de tópicos serão empregadas duas medidas utilizadas na literatura: a estabilidade e a coerência. Embora sirvam ao mesmo propósito, ambas trabalham com conceitos diferentes: enquanto a primeira mostra-se capaz de classificar uma determinada extração como instável, a segunda tenta capturar o grau de interpretabilidade da extração.

4.5.1 Estabilidade

Para avaliar a estabilidade de uma determinada escolha por um número de tópicos, é seguido o método estudado no capítulo anterior. São selecionadas 50 amostras aleatórias com 80% dos documentos da base, e a estabilidade é calculada com base nos 10 primeiros termos de cada tópico.

4.5.2 Coerência

Já para a coerência, é calculada a nota para cada um dos tópicos encontrados na base amostrada, ou seja, ao optar-se por n tópicos, são n notas de coerência calculadas. Uma possibilidade seria obter a média dessas notas, para cada extração, e comparar a distribuição das médias, assim como feito para a medida de estabilidade.

No entanto, a média das notas de coerência pode esconder a presença de tópicos incoerentes. Dessa forma, será usada a medida de SP [110] para balancear o tópico menos coerente e a coerência média de uma determinada extração. Chamamos esta medida de SPT e ela é definida como:

$$SPT = \sqrt{\sqrt{\min(C)avg(C)}\left(\frac{\min(C) + avg(C)}{2}\right)} \tag{4.4}$$

onde $C = c_1, ..., c_n$ é o conjunto das n notas de coerência, para uma extração de n tópicos, $\min(C)$ representa a menor nota de coerência deste conjunto e $avg(C)$ representa a média das n notas. Assim como as notas individuais de coerência, a medida SPT também está entre 0 e 1 e quanto maior, mais coerente a extração.

Embora o estado da arte da medida de coerência seja obtido ao utilizar a estatística da Wikipedia, esta implementação é mais custosa computacionalmente [90]. Portanto, optou-se neste trabalho por calcular as probabilidades utilizando a estatística do *corpus* sob análise.

4.6 Escolha do modelo

Como o objetivo final da Modelagem de Tópicos consiste em sumarizar o conteúdo do conjunto de documentos em tópicos interpretáveis, a medida de coerência apresenta-se como a melhor alternativa para avaliação de modelos, apesar de computacionalmente custosa. Além da etapa de avaliação do número de tópicos, feita por duas medidas custosas computacionalmente, medir a coerência de múltiplas extrações para selecionar o melhor modelo deixaria o problema intratável do ponto de vista prático.

Neste trabalho serão avaliados os algoritmos de acordo com o número de documentos mal classificados na base anotada e o de melhor performance será adotado também para a base cega.

O modelo de tópicos utilizado neste trabalho é a NMF com inicialização composta por fatores iniciais baseados na aproximação esparsa da fatoração SVD [84]. Este modelo minimiza uma divergência de escolha adicionando o seguinte termo de regularização:

$$\alpha(L1_{ratio}||vec(W)||_{L1} + L1_{ratio}||vec(H)||_{L1} + 0.5(1 - L1_{ratio})||W||^2_{L2} + 0.5(1 - L1_{ratio})||H||^2_{L2}) \tag{4.5}$$

onde α controla se haverá alguma regularização e $L1_{ratio}$ controla o balanço entre as penalizações das normas L1 e L2 das matrizes H e W.

Os modelos avaliados na base anotada serão obtidos variando a divergência que se deseja minimizar e variando os parâmetros de regularização.

4.7 Análise do resultado

O resultado será analisado qualitativamente a partir da inspeção das 10 principais palavras de cada tópico encontrado. Por fim, a proximidade entre os tópicos e como os documentos associados a eles foram classificados de acordo com a NFM serão inspecionados visualmente com a utilização da UMAP [101].

Capítulo 5

Resultados e Discussões

O objetivo do trabalho é avaliar o uso conjunto das medidas de coerência e estabilidade para a escolha do número de tópicos em bases de dados filtradas por critérios de qualidade de dados, preservando a estabilidade da extração e coerência semântica dos tópicos encontrados.

Para isso, primeiramente, as medidas serão analisadas na base anotada de notícias da BBC Sports [38]. Esta base de dados atende aos critérios mínimos de qualidade de dados definidos neste trabalho.

Posteriormente, as medidas serão avaliadas em uma base de dados não-anotada, composta por notícias que não necessariamente atendem aos critérios de qualidade definidos. Dessa forma, serão avaliados os níveis de estabilidade e coerência, com e sem filtragem por qualidade de dados.

5.1 Base anotada

A base de dados anotada utilizada neste trabalho é a BBC Sports [38]. Ela conta com 737 notícias de esporte publicadas entre 2004 e 2005, separadas em 5 tópicos: atletismo, cricket, futebol, rugby e tênis. Todas as notícias estão escritas em língua inglesa, possuem tamanho acima do mínimo definido nos critérios de qualidade definidos anteriormente, e não estão sujeitas a presença de ruído. Assim, não faz-se necessária a etapa de análise da qualidade de dados.

Pré-processamento

Os documentos presentes na base passaram pelo pré-processamento padrão: após serem divididos em *tokens*, são removidas as *stopwords*, palavras com menos de 3 caracteres e numerais. Após essa etapa, é realizada a identificação de bigramas e trigramas frequentes.

Com isso, os *tokens* passam por uma transformação algébrica, para que os documentos sejam representados matricialmente. É feita uma transformação tf-idf, com remoção de palavras que ocorrem em 4 ou menos notícias, ou em pelo menos 80% dos documentos. É feita então uma escolha pelos 1000 principais termos e, finalmente, a matriz de documentos e termos é normalizada para que os documentos fiquem com comprimento unitário.

Comparação entre os modelos

Após a etapa de pré-processamento, e dado que sabe-se o número anotado de tópicos e o tópico à qual cada documento pertence, é possível calcular o número de classificações incorretas de cada modelo. Como um modelo de tópicos encontra o peso da associação de cada documento com cada um dos grupos encontrados, assume-se que o maior peso é o tópico encontrado para aquele documento.

Dessa forma, foram testadas diferentes variações da NMF, assim como a LDA. Para a NMF, variaram-se os algoritmos entre o aditivo (representado pela sigla "cd") e o multiplicativo ("mu"), além da escolha pelas divergência de Frobenius e Kullback-Leiber (apenas quando o algoritmo é multiplicativo), assim como diferentes graus de regularização (parâmetros "alpha"e "L1"). Os modelos e seus respectivos resultados são apresentados na Tabela 5.1.

Observa-se que os modelos de melhor performance foram variações de diferentes graus de regularização do modelo aditivo que minimiza a divergência de frobenius. É interessante notar que o mesmo algoritmo apresentou o segundo pior resultado ao sofrer regularização acentuada pela norma L2. As variações de Kullback-Leiber obtiveram performance intermediária.

Os 4 primeiros modelos da Tabela apresentam performance semelhante. Uma análise mais detalhada evidencia que os dois primeiros modelos cometem os mesmos erros de classificação, enquanto o quarto modelo também comete os mesmos erros, com a adição de mais um documento mal classificado. Já o terceiro modelo, que não apresenta qualquer regularização, divergiu das classificações feitas pelo primeiro modelo, como mostra a Tabela 5.2.

O modelo escolhido foi a NMF com algoritmo aditivo minimizando a divergência de frobenius, com regularização majoritariamente L1 (parâmetros $\alpha = 0.1$ e $L1 = 0.1$), embora a escolha pelos outros modelos do topo da Tabela fosse igualmente justificada.

Valor da incoerência

Enquanto o trabalho que propôs a medida de estabilidade apresenta um nível para a instabilidade, não há um nível que defina um tópico como incoerente. Para estimar

Tabela 5.1: BBC Sports: comparação entre modelos.

Classificação	Modelo	Mal classificados
1	NMF-cd-frobenius-alpha(0.1)-l1(0.0)	39
2	NMF-cd-frobenius-alpha(0.1)-l1(0.1)	39
3	NMF-cd-frobenius-alpha(0.0)	39
4	NMF-cd-frobenius-alpha(0.5)-l1(0.0)	40
5	NMF-mu-frobenius-alpha(0.1)-l1(0.0)	48
6	NMF-mu-frobenius-alpha(0.1)-l1(0.1)	50
7	NMF-mu-frobenius-alpha(0.5)-l1(0.0)	50
8	NMF-mu-kullback-leibler-alpha(0.0)	74
9	NMF-mu-kullback-leibler-alpha(0.1)-l1(0.0)	74
10	NMF-mu-kullback-leibler-alpha(0.1)-l1(0.5)	75
11	NMF-mu-kullback-leibler-alpha(0.1)-l1(0.1)	75
12	NMF-mu-kullback-leibler-alpha(0.5)-l1(0.1)	76
13	NMF-mu-kullback-leibler-alpha(0.5)-l1(0.0)	76
14	NMF-mu-kullback-leibler-alpha(0.5)-l1(0.5)	83
15	NMF-cd-frobenius-alpha(0.1)-l1(0.5)	106
16	NMF-cd-frobenius-alpha(0.5)-l1(0.1)	109
17	NMF-mu-frobenius-alpha(0.1)-l1(0.5)	115
18	NMF-mu-frobenius-alpha(0.5)-l1(0.1)	119
19	NMF-mu-frobenius-alpha(0.5)-l1(0.5)	195
20	LDA	206
21	NMF-cd-frobenius-alpha(0.5)-l1(0.5)	302

Tabela 5.2: BBC Sports: divergência entre modelos.

Documento	Modelo 1	Modelo 3	Tópico verdadeiro
374	football	rugby	football
464	football	cricket	football
534	football	rugby	football
588	football	rugby	football
597	football	rugby	rugby
605	football	rugby	rugby
619	football	rugby	rugby
678	football	rugby	rugby
697	football	cricket	rugby
726	football	rugby	rugby

esse nível, foram sorteados aleatoriamente 1000 conjuntos de 10 palavras para formar tópicos sintéticos. É de se esperar que a maior parte destes sorteios resulte em tópicos incoerentes. O resultado é mostrado na Figura 5.1.

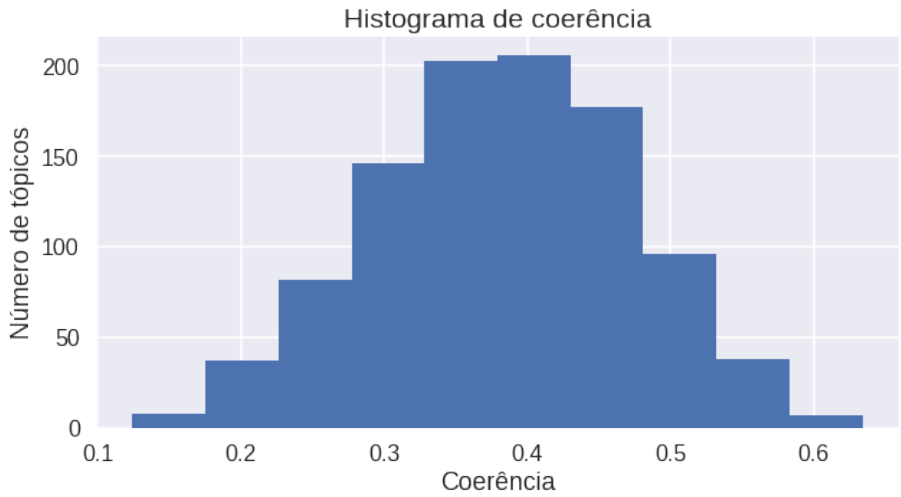


Figura 5.1: Análise de coerência para tópicos aleatórios.

Já para o resultado da extração com o número de tópicos anotados, apresentado na Tabela 5.3, percebe-se empiricamente que os tópicos obtidos parecem ser coerentes, com exceção do tópico de cricket, que apresenta palavras muito gerais, e do de rugby, pelo mesmo motivo. A medida de coerência para cada tópico é colocada ao lado de suas principais palavras e pode-se constatar que o mínimo de coerência está associado justamente ao tópico referente a cricket. Embora o tópico de rugby também pareça pouco coerente, o nível de coerência é mais alto.

Em cricket, o modelo de tópicos parece ter misturado aspectos do jogo, representados pelas palavras *runs*, *series*, *tour* e *test*, com os países mais relevantes para o esporte, representados nas outras palavras do tópico. O nível de coerência é mais baixo pela menor coocorrência entre estes dois diferentes conjuntos. As palavras que mais contribuíram para a baixa nota de coerência foram *runs* e *south*, o que foi evidenciado pelo fato da coerência subir para 0.60 quando estes termos foram retirados. Embora a primeira palavra coocorra frequentemente com *series* e *cricket*, por ser específica ao jogo, a frequência de coocorrência é baixa com os países. Já *south*, presente por representar o país África do Sul, coocorre frequentemente apenas com *africa*. Neste caso, o modelo de identificação de bigramas frequentes falhou ao não capturar *South Africa* como um só *token*, o que provavelmente aumentaria a coerência deste tópico.

Já em rugby, que apresenta exatamente a mesma característica de misturar aspectos do jogo com os países que disputam o esporte, não sofreu tanta influência desta mistura de termos e obteve a mais alta nota de coerência. O tópico de futebol,

Tabela 5.3: Resultado da extração para 5 tópicos.

Tópico	Principais palavras	Coerência
futebol	chelsea, united, arsenal, club, league, liverpool, football, cup, manager, mourinho	0.59
cricket	test, pakistan, cricket, india, series, australia, south, tour, africa, runs	0.53
tênis	open, seed, australian, roddick, final, beat, federer, set, win, hewitt	0.72
rugby	england, wales, ireland, france, nations, robinson, scotland, half, rugby, game	0.72
atletismo	olympic, race, athens, european, champion, indoor, holmes, athletics, kenteris, world	0.71

pelo contrário, é mais coerente do que os tópicos de rugby e cricket, mas possui nota de coerência igual a 0.59.

Esse resultado indica que uma nota de coerência acima de 0.6 implica em uma região de coerência. Já uma nota entre 0.5 e 0.6 indica uma região de confusão, contendo tópicos coerentes e incoerentes, enquanto uma nota abaixo de 0.5 implica em incoerência.

Escolha do número de tópicos

Para a escolha do número de tópicos, foi seguida a abordagem descrita no capítulo anterior: primeiro seleciona-se empiricamente um intervalo de tópicos com base no número de documentos. Como essa base apresenta um total de 737 documentos, 23 tópicos agrega na média cerca de 30 documentos por tópico e foi escolhido como limite superior.

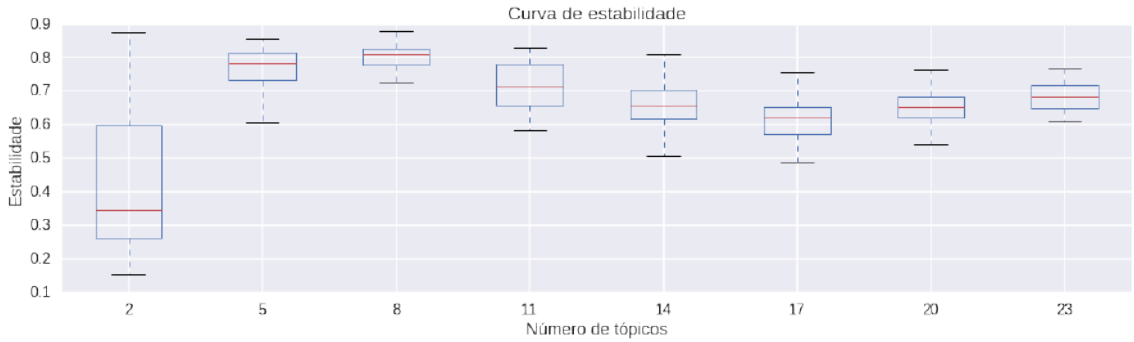


Figura 5.2: Primeiro intervalo de escolha: entre 2 e 23 tópicos, com intervalo de 3.

A Figura 5.2 mostra a variação da medida de estabilidade para o intervalo de 2 a 23 tópicos. Essa variação é feita usando um intervalo de 3 tópicos para economizar tempo de processamento e ainda assim tornar possível visualizar a tendência geral da medida de estabilidade. Visto que a escolha por 2 tópicos pode apresentar extrações instáveis, este número é automaticamente excluído do intervalo possível. Nota-se também uma tendência de queda a partir de 13 tópicos. Como a busca inicial é por uma extração mais geral, vamos optar por um menor número de tópicos, o que restringe o segundo intervalo de busca entre 3 e 13 tópicos.

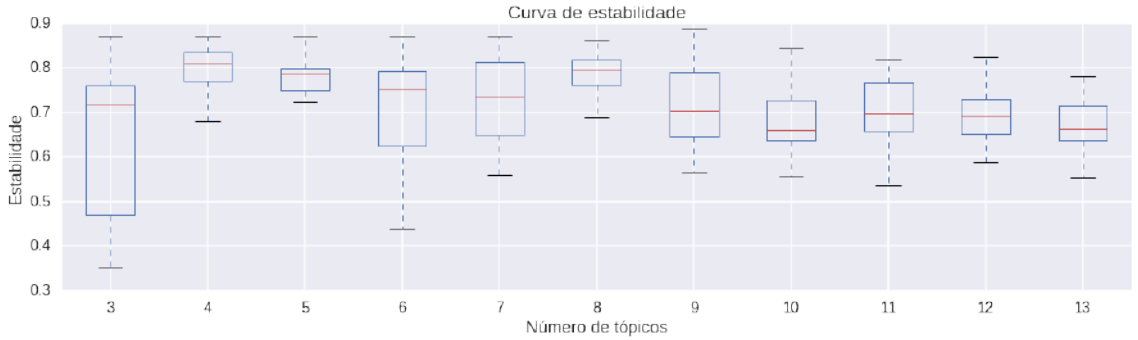


Figura 5.3: Segundo intervalo de escolha: entre 3 e 13 tópicos, com intervalo de 1.

No segundo intervalo, mostrado na Figura 5.3, a escolha por 3 tópicos sai da barra de erro e fica próxima ao nível de instabilidade, sendo descartada. Qualquer outra escolha parece razoável, com destaque para 4, 5 e 8 tópicos, pelas maiores medianas e menor intervalo interquartil.

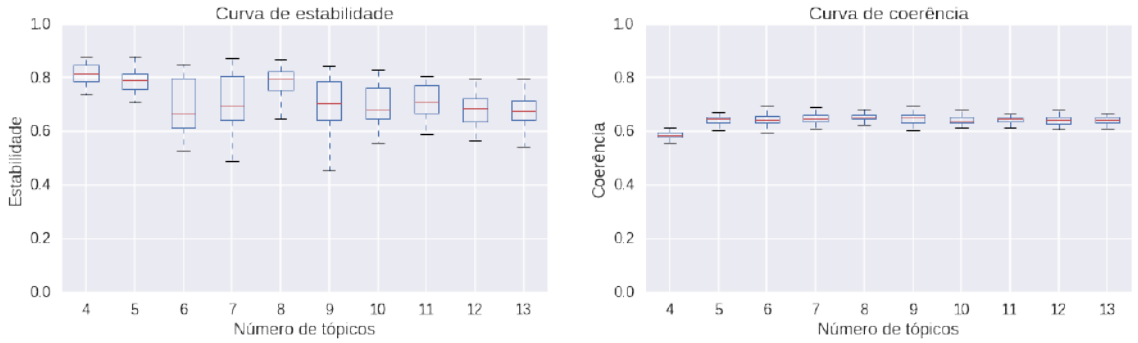


Figura 5.4: Último intervalo de escolha: análise de estabilidade e coerência para 4 a 13 tópicos.

Por fim, é feita a análise de coerência média para o intervalo de 4 a 13, como mostrado na Figura 5.4. Todas as escolhas parecem apresentar tópicos coerentes. No entanto, como não há graduação da escala desta medida, é difícil afirmar se uma escolha é melhor do que a outra.

Outro aspecto que acaba sendo ignorado é a presença de tópicos incoerentes nas extrações, uma vez que um tópico com um maior nível de coerência pode cancelar a presença de um tópico incoerente ao aumentar o valor médio. Assim, além de ver a média, deve-se analisar também o valor mínimo de coerência dessas extrações, o que é mostrado na Figura 5.5.

Nesta Figura, as escalas foram fixadas entre os valores mínimos e máximos de coerência. Com isso é possível ver que o nível médio de coerência é próximo entre as diferentes escolhas de números de tópicos. Já para a coerência mínima, as notas ficam entre 0.4, nível que representa tópicos incoerentes, e 0.6, que representa tópicos coerentes. Assim, essa curva mostra que a escolha por 4 tópicos possui uma mediana próxima ao nível de incoerência.

Como ambas as características são desejáveis, um valor médio mais alto e a presença apenas de tópicos semanticamente coerentes, a medida seletora de tópicos é o balanceamento entre o valor médio e mínimo de coerência, mostrado na Figura 5.6. Nesta curva, a escolha por 5 tópicos destaca-se como a maior mediana para uma extração mais geral e portanto deve ser o número de tópicos escolhido, condizendo com a anotação da base.

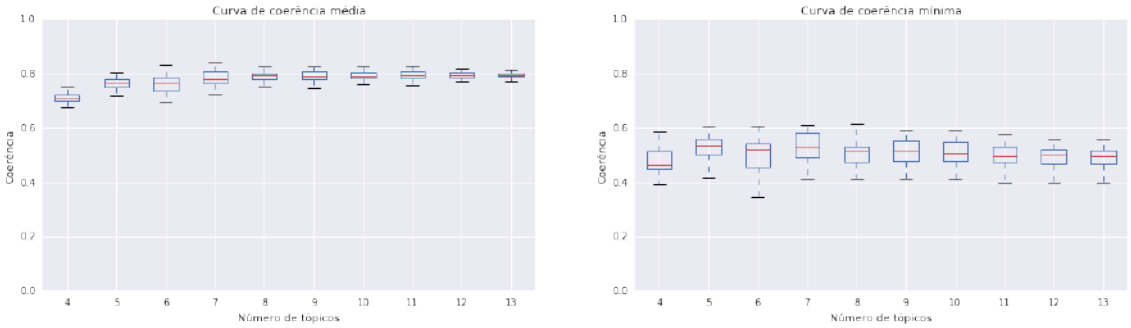


Figura 5.5: Análise de coerência média e mínima para 4 a 13 tópicos.

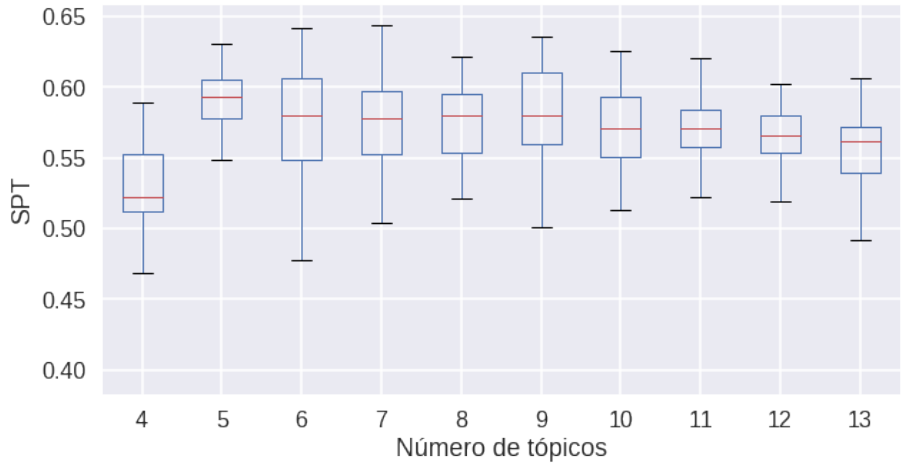


Figura 5.6: Análise de SPT para 4 a 13 tópicos.

Análise do resultado

Os tópicos, representados pelas 10 principais palavras, e seus respectivos pesos podem ser vistos na Tabela 5.4. Apesar destas palavras remeterem aos tópicos anotados, é necessária uma análise complementar da associação de cada documento-tópico para assim avaliar a qualidade da extração.

Esta análise é feita a partir da Figura 5.7, onde os documentos são colocados em cada célula a partir da anotação da base e do tópico determinado pelo modelo, que por sua vez é obtido pelo maior peso da associação documento-tópico. Nela, é possível ver que o tópico 1 é composto em sua maioria por notícias de futebol, mas conta também com notícias de rugby e tênis. De fato, a associação futebol-rugby

Tabela 5.4: Tópicos encontrados.

Tópico	Principais palavras	Peso
1	chelsea, united, arsenal, club, league, liverpool, football, cup, manager, mourinho	0.26
2	test, pakistan, cricket, india, series, australia, south, tour, africa, runs	0.19
3	open, seed, australian, roddick, final, beat, federer, set, win, hewitt	0.18
4	england, wales, ireland, france, nations, robinson, scotland, half, rugby, game	0.21
5	olympic, race, athens, european, champion, indoor, holmes, athletics, kenteris, world	0.18

parece ser a maior causa de confusão por parte do modelo. No tópico 1 são 11 documentos mal classificados, enquanto o tópico 4 conta com mais 10. Somados, os 21 documentos mal classificados correspondem a mais da metade dos casos em que o modelo erra.

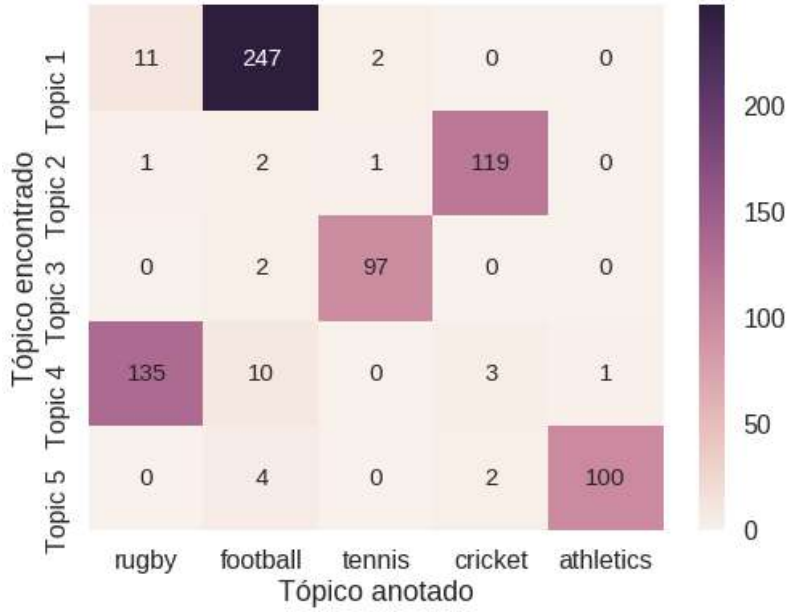


Figura 5.7: BBC Sports: matriz de confusão para 5 tópicos.

Por fim, é possível ver a separação entre os documentos na Figura 5.8. Os tópicos apresentam uma fronteira clara. Há alguns casos de documentos que ultrapassam a fronteira e são mal classificados, principalmente na região central da Figura, na área de fronteira entre futebol e rugby.

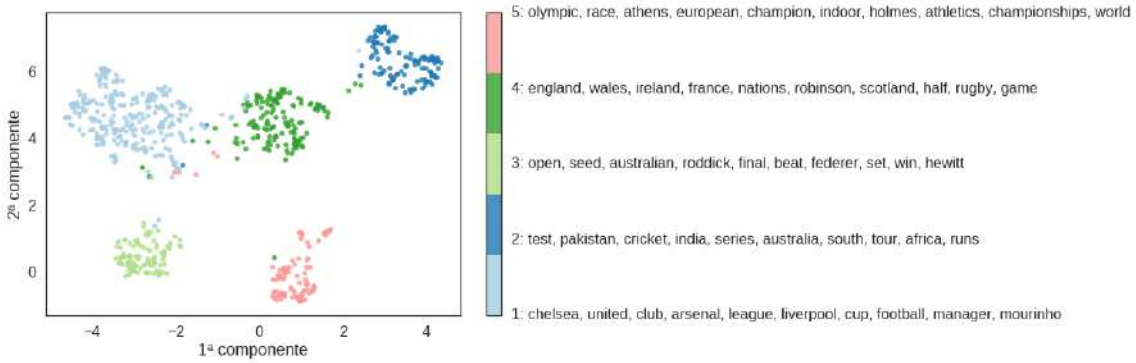


Figura 5.8: BBC Sports: UMAP.

Zoom in

Existe uma tendência de queda no nível de coerência com o aumento do número de tópicos, o que sugere que os tópicos ficam muito específicos. De fato, a Figura 5.10 evidencia o baixo número de documentos associados a alguns tópicos, enquanto a Figura 5.9 mostra os tópicos encontrados e as fronteiras entre eles. Alguns tópicos, como o 6 ou 9, mostram-se incoerentes em uma primeira análise.

O significado desses tópicos pode ser inferido apenas com uma análise posterior à extração. Por exemplo, a avaliação das notícias do tópico 6 mostra que este tópico agregou as notícias que relatam lances do jogo, tanto para futebol quanto para rugby. Isso é evidenciado pelas principais palavras, como *goal*, *ball*, *shot* e *penalty*.

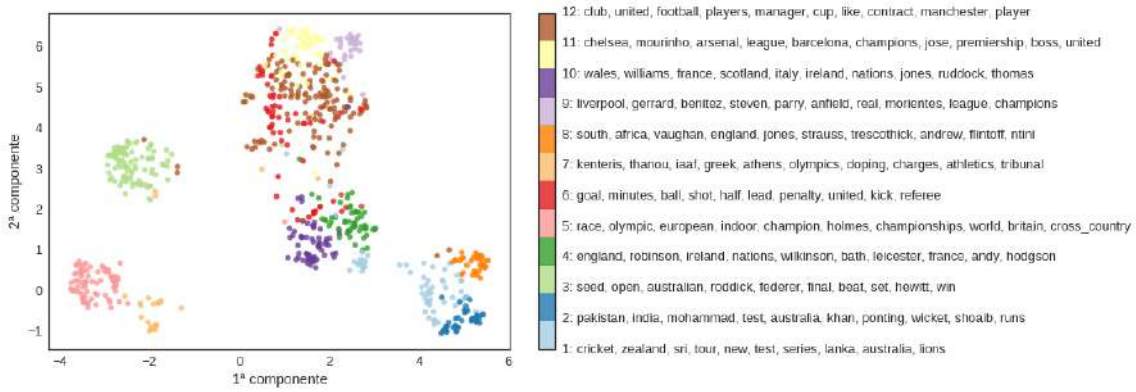


Figura 5.9: BBC Sports: UMAP para 12 tópicos.

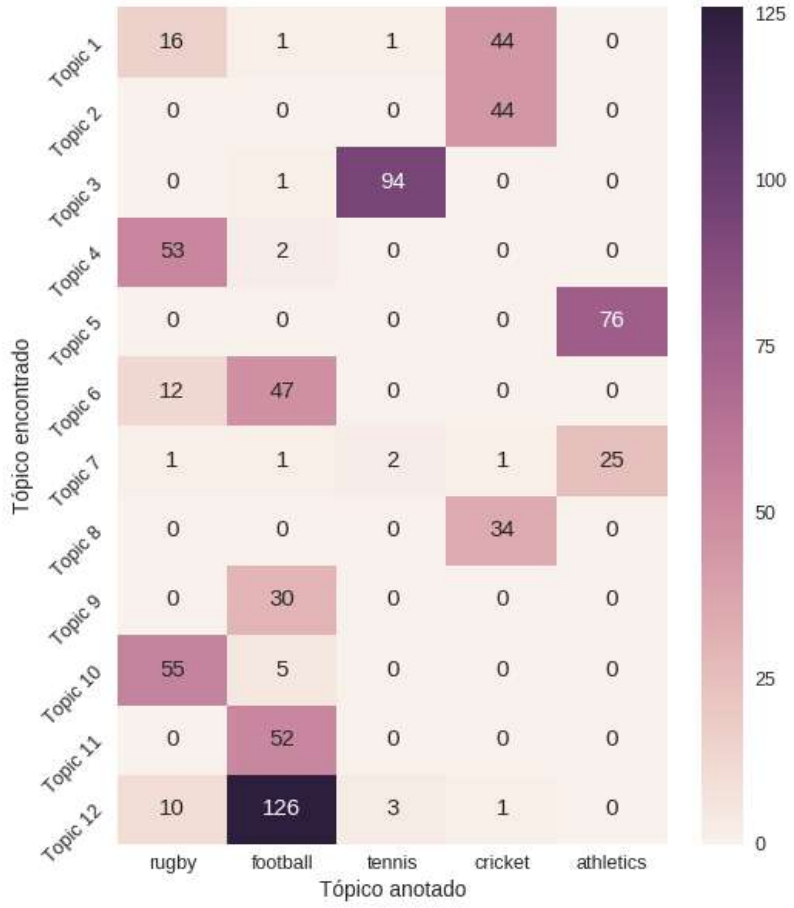


Figura 5.10: BBC Sports: confusão entre tópicos para 12 tópicos.

5.2 Base de dados não-anotada

Para esta base, é feita uma pesquisa pela palavra "brazil"para delimitar o contexto mais geral e todas as notícias entre 1º de janeiro de 2018 e 1º de março de 2019 são recuperadas, totalizando 1696 documentos. Os duplicados, um total de 80 documentos, são identificados, seguindo o método do capítulo anterior, e removidos.

Como o número de tópicos desta base não é conhecido *a priori*, alguns dos resultados obtidos para a base anotada serão aplicados sem uma análise quantitativa.

5.2.1 Análise de qualidade de dados

Os 1616 documentos são analisados pela sua qualidade seguindo as métricas definidas. A base de dados conta com 99% de acurácia e completude, onde a segmentação da nota por fonte da notícia é mostrada na Figura 5.11 e a segmentação por atributo é vista na Figura 5.12.

Como pode-se perceber, apenas uma das fontes contribui para a queda da nota de acurácia. Mas, pelo fato de contar apenas com 5 documentos, a nota global

Tabela 5.5: Análise de qualidade para fonte.

Documento	Acurácia	Completude
1	1	0.61
2	1	0
3	1	0
4	1	0
5	0.54	0

de acurácia praticamente não sofre impacto. Quanto à completude, duas fontes contribuem para abaixar a nota global: as fontes 2 e 3. Como a Figura 5.12 mostra, o problema está majoritariamente relacionado com a data de publicação.

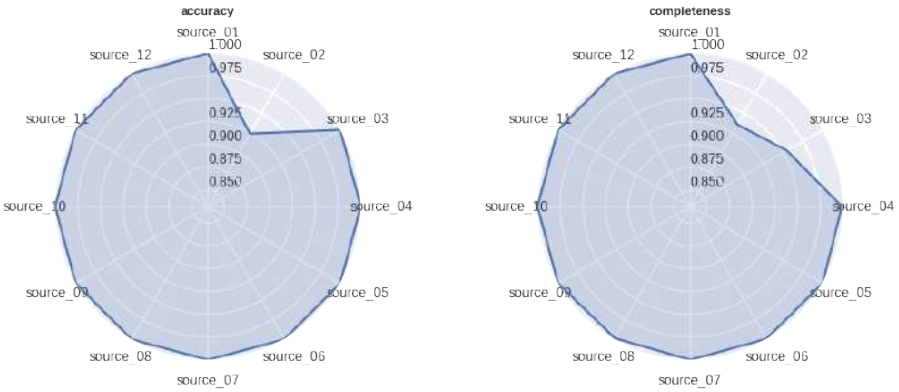


Figura 5.11: Análise da qualidade de dados das fontes.

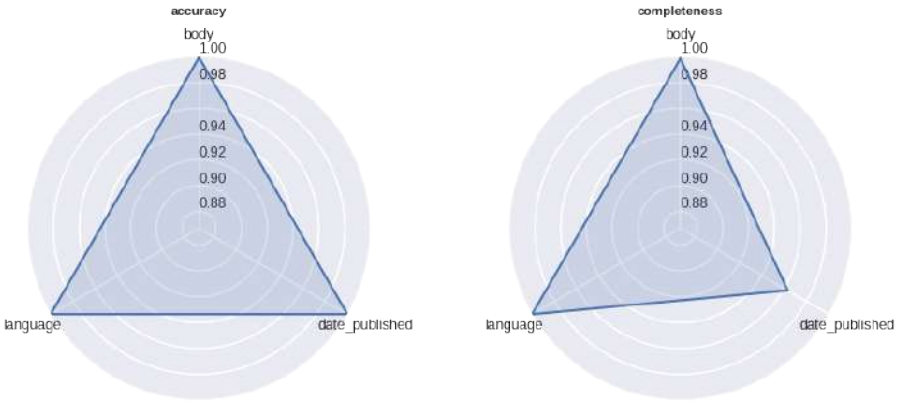


Figura 5.12: Análise da qualidade de dados dos atributos.

Uma inspeção mais detalhada da fonte 2 mostra as notas para cada um dos documentos publicados por ela na Tabela 5.5. Os documentos 1 e 5 apresentam algum problema de qualidade.

Para o primeiro documento, suas notas de qualidade por atributo são vistas na Tabela 5.6. A baixa nota de completude deste documento é inteiramente devida ao baixo número de palavras no corpo da notícia, que conta com o seguinte texto:

Tabela 5.6: Análise de qualidade de dados para documento 1 da fonte 2.

Atributo	Completude	Acurácia	Peso
idioma	1	1	5
data	1	1	1
corpo	0.15	1	5

Tabela 5.7: Análise de qualidade de dados para documento 5 da fonte 2.

Atributo	Completude	Acurácia	Peso
idioma	1	1	5
data	1	1	1
corpo	1	0	5

Danny Glover wants to tell you why democracy is under threat in Brazil. pic.twitter.com/MMUzKhWdQJ

Já para o outro documento, o problema apresentado é a baixa acurácia. Olhando para a segmentação das suas notas por atributo, listadas na Tabela 5.7, pode-se notar que o problema é devido ao corpo da notícia.

O motivo pelo qual a nota de acurácia deste documento é igual a zero deve-se à presença de trechos de código no corpo, como evidenciado pelo início do texto:

.controlsoutline:none.controls .controls-buttondisplay:block;z-index:9.controls .controls-buttonleft:0.controls.controls-nextright:0.controls.controls-next

Por fim, é feita a eliminação das notícias que não atendam aos critérios mínimos de qualidade, definidos empiricamente pela completude do corpo menor do que 1 e pela acurácia do corpo menor ou igual a 0.95. Estes critérios eliminam 8 documentos do conjunto, restando 1608 notícias para análise.

5.2.2 Pré-processamento

O pré-processamento padrão foi adotado. Para a transformação dos documentos em vetores, foram eliminados os termos que aparecem 4 vezes ou menos e em 80% dos documentos ou mais, antes de serem transformados utilizando tf-idf. A matriz resultante de documentos e termos conta com 9579 termos.

5.2.3 Escolha do número de tópicos

Para a escolha do número de tópicos, primeiro se escolheu um intervalo maior de tópicos, para analisar o perfil geral da estabilidade no *corpus*. Baseado no número de notícias, o intervalo escolhido ficou entre 2 e 47 tópicos, com um intervalo de 5

em 5. O aspecto geral é de uma extração estável, visto que o menor nível detectado foi pouco menor que 0.5, como pode-se ver na Figura 5.13.

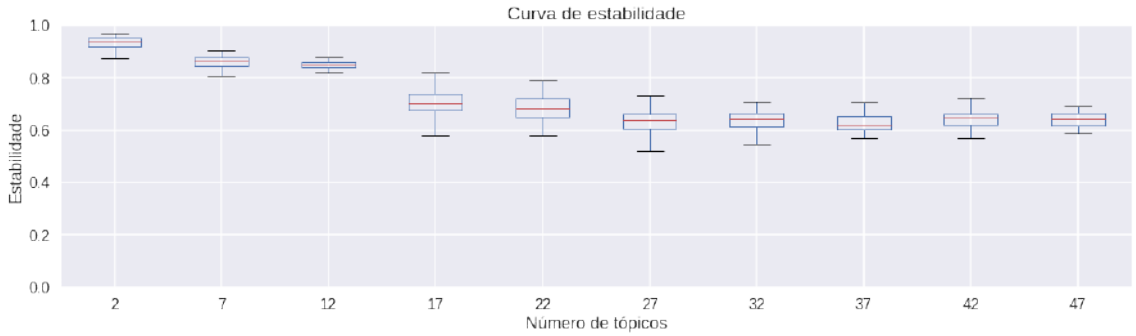


Figura 5.13: Análise de estabilidade.

Assim, pode-se prosseguir para a análise da coerência. Para tal, vamos nos concentrar no intevalo de 2 a 20, uma vez que esta é a região com o menor número de tópicos e apresenta um nível de estabilidade superior. A medida seletora de tópicos é vista na Figura 5.14 e indica que a escolha deve ser por 6 tópicos.

Nota-se uma tendência de queda de coerência com o aumento no número de tópicos, o que indica uma deterioricação da qualidade da extração com o maior detalhamento dos dados.

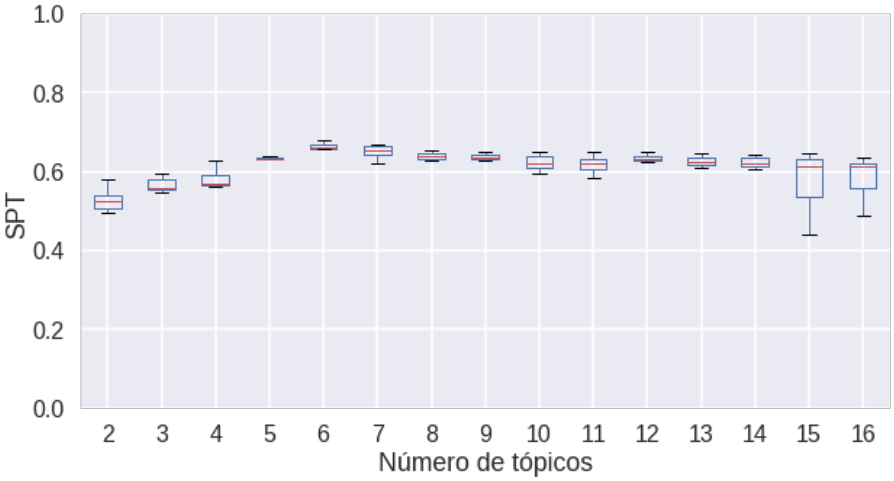


Figura 5.14: Análise de coerência.

5.2.4 Análise do resultado

Os tópicos encontrados, e os respectivos pesos, são apresentados na Tabela 5.8. Pode-se notar que os tópicos são abrangentes: enquanto o primeiro e o último falam especificamente de cultura no Rio de Janeiro, o terceiro abrange a política brasileira, com destaque para a disputa presidencial de 2018 e o caso da prisão do ex-presidente Lula. Já o segundo trata do mercado de petróleo, enquanto o quarto é do tema

Tabela 5.8: Tópicos encontrados.

Tópico	Principais palavras	Peso
1	rio, janeiro, carnival, city, samba, brazilian, contributing, reporter, festival, music	0.26
2	percent, oil, company, production, year, petrobras, energy, market, growth, crude	0.17
3	lula, bolsonaro, president, court, silva, candidate, party, haddad, corruption, political	0.17
4	cup, neymar, world, game, team, match, coutinho, liverpool, goal, minute	0.16
5	trade, trump, u.s., tariffs, china, steel, united, states, countries, chinese	0.13
6	ipanema, rua, entrance, copacabana, agave, free, blue, tel, bar, canastra	0.11

esporte, com destaque para a Copa do Mundo de 2018, Neymar e a transferência do Philippe Coutinho. Por fim, o quinto tópico abrange a relação econômica Estados Unidos-China.

Na Figura 5.15 pode-se observar como os documentos estão distribuídos ao longo dos tópicos. De início, fica claro que o tópico de esportes destaca-se dos demais e que grande parte do noticiário internacional capturado nesta base destaca o carnaval carioca. O tópico de cultura ocupa a parte de baixo do gráfico, com duas seções delimitadas: uma referente especificamente ao carnaval, e outra referente a outras atividades culturais, com destaque para Ipanema. Por fim, os tópicos de economia e política ocupam o canto esquerdo do gráfico.

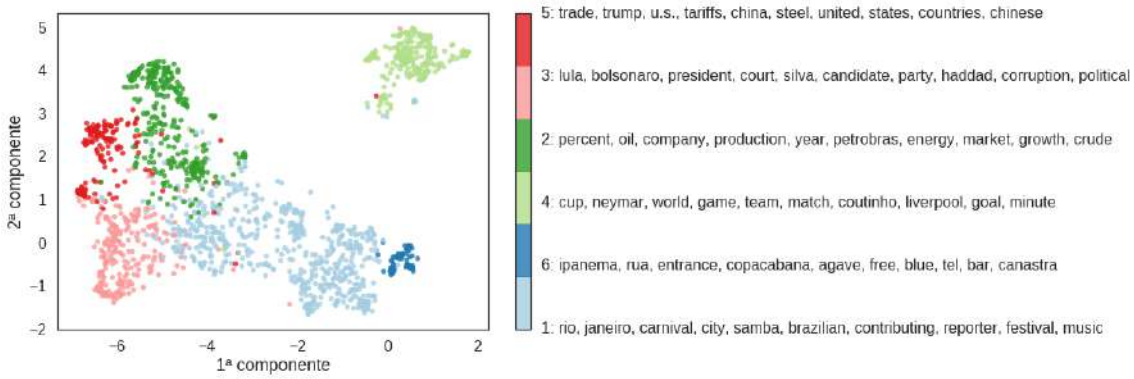


Figura 5.15: Tópicos encontrados (UMAP).

5.2.5 *Zoom* na base

Abaixo, na Tabela 5.9 pode-se ver os tópicos encontrados, quando aumentamos para 15 o número de tópicos, enquanto a Figura 5.16 mostra a separação entre eles.

Tabela 5.9: Tópicos encontrados.

Tópico	Principais palavras	Peso
1	company, according, brazilian, alves, lise, senior, government, vale, president, temer	0.11
2	rio, festival, music, album, film, jazz, janeiro, saturday, event, brazilian	0.09
3	cup, neymar, world, team, game, match, belgium, coach, goal, league	0.08
4	rio, city, beach, neighborhood, tijuca, copacabana, located, janeiro, barra, restaurants	0.08
5	oil, production, company, crude, energy, offshore, petrobras, rigzone, year, drilling	0.07
6	police, rio, state, janeiro, security, franco, violence, military$_p$*olice*, *intervention*, *public*$_s$*ecutiry*	0.07
7	trade, trump, tariffs, u.s., china, steel, united, chinese, states, imports	0.06
8	percent, increase, ibge, inflation, year, index, square$_m$*eter*, *prices*, *registered*, *real*$_e$*state*	0.06
9	bolsonaro, haddad, candidate, jair, party, president, campaign, candidates, election, percent	0.06
10	lula, court, silva, president, corruption, decision, judge, conviction, party, supreme	0.06
11	ipanema, rua, entrance, agave, copacabana, blue, free, tel, canastra, bar	0.06
12	carnival, parade, samba, grupo, especial, school, samba$_s$*chool*, *blocos*, *bloco*, *rio*	0.06
13	maduro, venezuela, guaido, venezuelan, border, government, opposition, venezuelans, colombia, aid	0.04
14	coutinho, liverppol, barcelona, club, philippe, summer, klopp, transfer, player, deal	0.04
15	flamengo, vasco, game, minute, botafogo, match, fluminense, taça, brasileirão, rubro	0.04

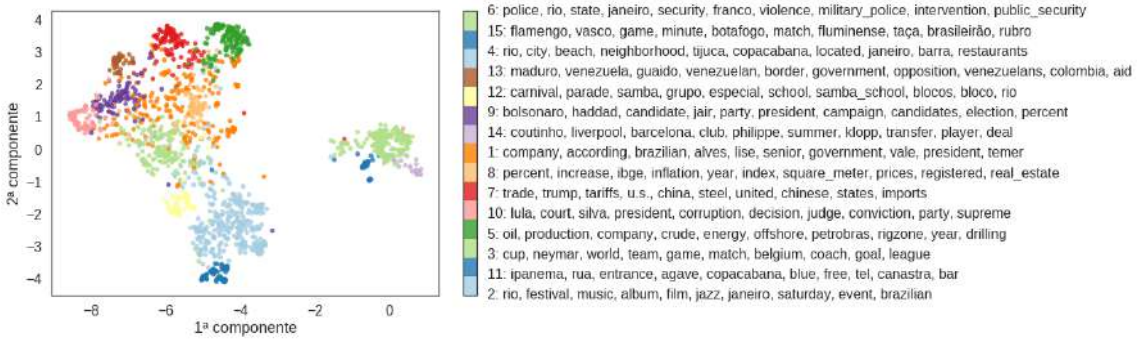


Figura 5.16: Tópicos encontrados em uma extração mais detalhada (UMAP).

A seção de esporte, representada por um único tópico na extração anterior, é agora representada por 3 tópicos, presentes no canto superior direito da Figura 5.16: um tópico sobre o futebol brasileiro, com destaque para o Brasileirão e a Taça Rio; um tópico sobre a transferência do jogador Philippe Coutinho, do Liverpool para o Barcelona e um tópico sobre a Copa do Mundo 2018.

A seção de cultura, representada por dois tópicos na extração anterior, dividiu-se em mais duas: uma que fala apenas sobre carnaval (em amarelo na Figura), uma que fala sobre os bares de Ipanema, e duas que falam sobre festivais no Rio de Janeiro e atividades culturais.

Os tópicos de petróleo e economia Estados Unidos-China continuam presentes, representados no canto superior esquerdo da Figura pelas cores verde e vermelho, respectivamente. Um pouco mais abaixo está representado em laranja o tópico 7, que trata da economia brasileira e refere-se às notícias sobre estudos de preço e inflação do IBGE.

Mais abaixo está o tópico 1, representado em verde claro, que trata da intervenção federal feita no Rio de Janeiro para combater a violência urbana. É interessante notar a presença das palavras "lise"e "alves"neste tópico, nome da correspondente

internacional.

O canto inferior esquerdo divide-se entre os tópicos 12, 8 e 11. O primeiro fala sobre a crise na Venezuela e as negociações de Brasil e Colômbia. O segundo fala sobre a campanha presidencial de 2018, que elegeu Jair Bolsonaro como presidente do Brasil. E, por fim, o último refere-se ao julgamento do ex-presidente Lula.

Qualitativamente, é possível notar que não houve perda de coerência, ao contrário do que indicava a medida seletora. Apesar de haver tópicos mais específicos, o surgimento deles ajudou a evidenciar assuntos que antes estavam misturados, como os novos tópicos de política e esporte.

5.2.6 Filtragem de documentos com problemas de qualidade de dados

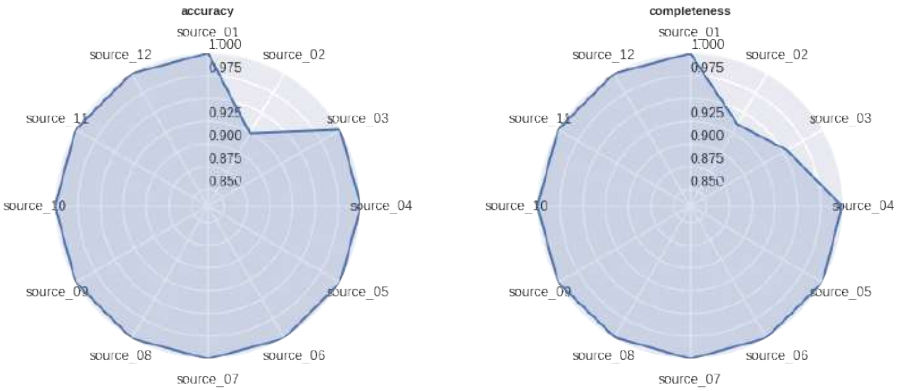
Apenas 7 documentos, dos 1616 presentes na base, não atenderam os critérios definidos de completude e acurácia. Pela baixa representatividade, a remoção desses documentos não impactou o resultado da extração. O nível de estabilidade e coerência permaneceram iguais, assim como os tópicos encontrados após a escolha por 6 tópicos.

5.3 Base de dados não-anotada: análise da influência de ruído

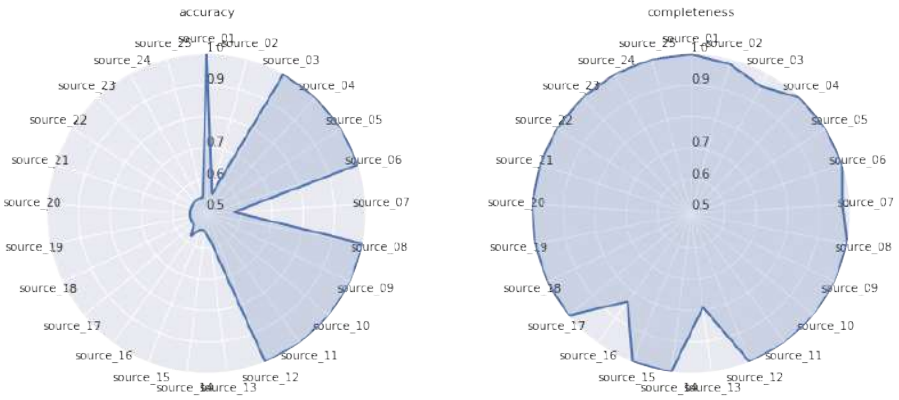
Como a base anterior não sofreu influência significativa de ruído, foram selecionados manualmente documentos ruidosos para introduzir na base. Esses documentos são notícias escritas em língua portuguesa, filtradas pela mesma palavra, "brazil", e pelo mesmo período de tempo, entre 1º de janeiro de 2018 e 1º de março de 2019. Após essa etapa, selecionam-se 100, 300 e 600 documentos, escolhidos de maneira aleatória, para introduzir na base de dados original. O impacto destes documentos é avaliado a seguir.

5.3.1 Impacto na qualidade de dados

A introdução de documentos escritos em língua portuguesa irá afetar a acurácia do *corpus*, impactando principalmente o idioma dos documentos, que receberá a nota zero de acurácia. A introdução das fontes de documentos em língua portuguesa pode ser vista na Figura 5.17, onde fica evidente a inferioridade de acurácia dessas fontes. Também é possível perceber que algumas dessas fontes irão impactar na completude da base ruidosa.



(a) Base de dados original



(b) Base de dados com introdução de notícias escritas em portugês

Figura 5.17: Impacto do ruído na qualidade de dados das fontes.

Ruído	Documentos introduzidos	Acurácia	Completude
Nenhum	0	99%	99%
Baixo	100	97%	99%
Médio	300	93%	98%
Alto	600	88%	97%

Tabela 5.10: Notas de qualidade para os diferentes níveis de ruído.

De uma maneira global, as notas de qualidade obtidas para os três diferentes níveis de ruído podem ser encontradas na Tabela 5.10. A um nível macro pode-se perceber que a completude foi afetada pela introdução desses documentos, embora não da mesma forma que a acurácia.

Para entender a causa dessa queda, esta nota pode ser decomposta pelos atributos das notícias (Figura 5.18). Vê-se que a queda da nota de completude é devida à introdução de documentos que não possuem data de publicação completas e textos abaixo do limite inferior de caracteres definido. Já a acurácia sofre impacto exclusivamente do idioma das notícias.

5.3.2 Impacto na estabilidade do *corpus*

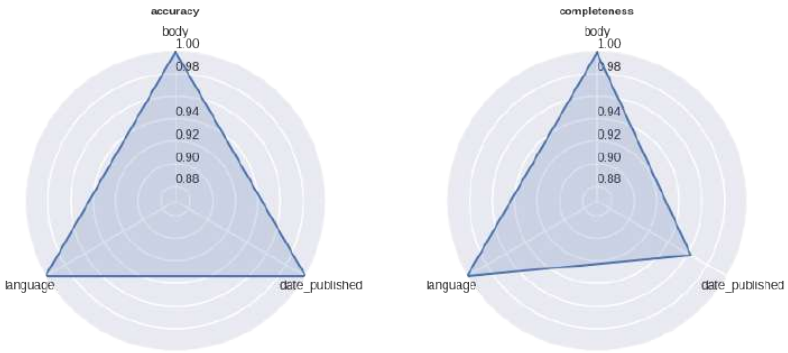
A Figura 5.19 mostra a evolução da medida de estabilidade com o aumento do número de tópicos. É possível ver que não existe queda significativa no nível médio de estabilidade, e que a introdução do ruído não foi capaz de tornar o *corpus* instável.

Observa-se também que a queda de patamar de estabilidade, localizada entre 12 e 17 tópicos na extração original, e entre 7 e 12 para o restante das bases. Além disso, enquanto a extração original apresenta uma estabilização do nível médio, as bases com ruído apresentam uma tendência de queda com o aumento do número de tópicos, de maneira leve com 100 documentos, e de forma mais acentuada com 300 e 600 documentos ruidosos.

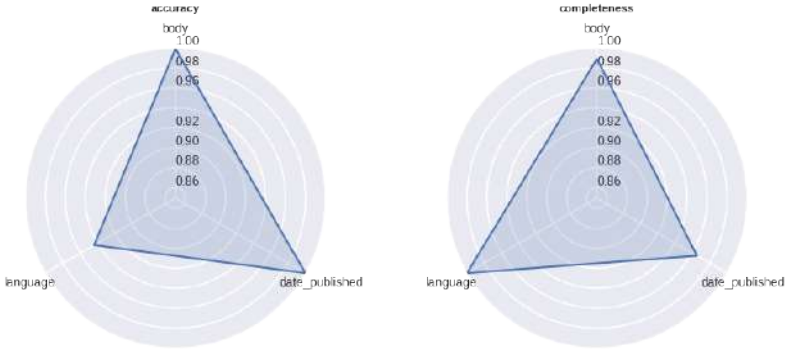
5.3.3 Impacto na seleção do número de tópicos

Na Figura 5.20 pode-se ver como a medida seletora de tópicos comporta-se com os diferentes níveis de ruído. De uma maneira geral, pode-se observar que o nível médio de coerência é maior para a base de dados livre de documentos com problemas de qualidade, ficando acima de 0.6 para mais do que 5 tópicos.

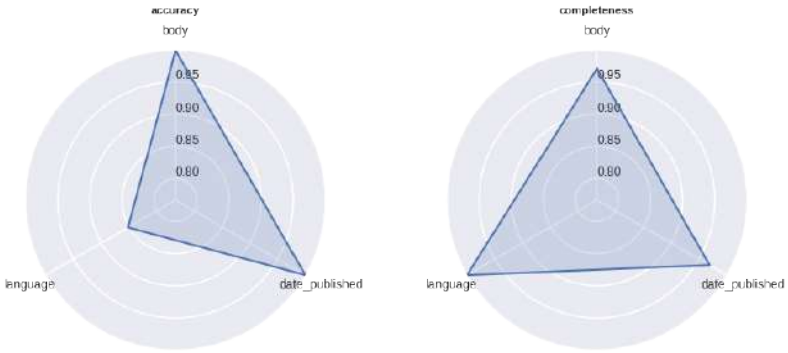
Para o menor nível de ruído há uma queda no nível de coerência a partir de 10 tópicos, enquanto a escolha final é por 8 tópicos. Já para o nível médio, todas as escolhas mostram-se inferiores à base original, enquanto a escolha final é por 3 tópicos. Por fim, a extração com maior nível de ruído apresenta um nível médio de coerência inferior aos dados originais, mas superior ao nível médio de ruído. A



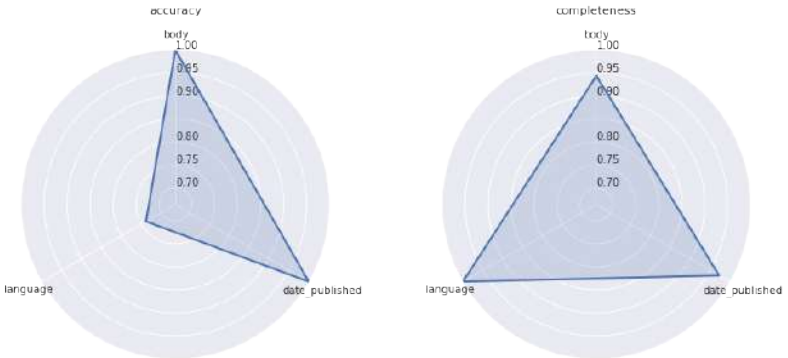
(a) Extração original



(b) 100 documentos

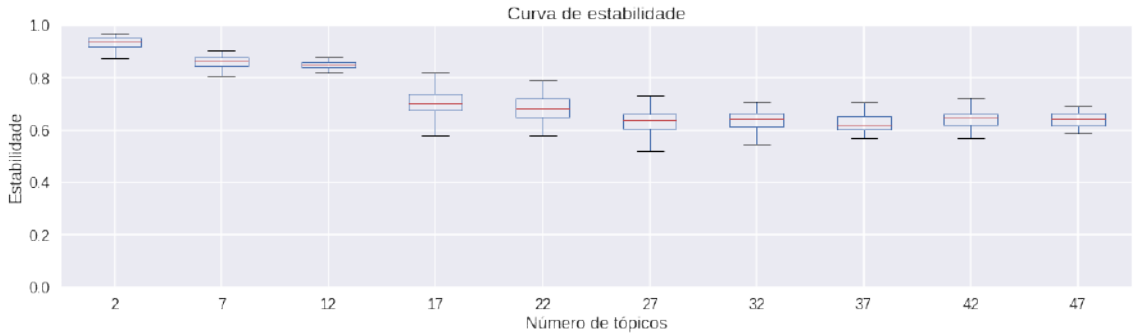


(c) 300 documentos

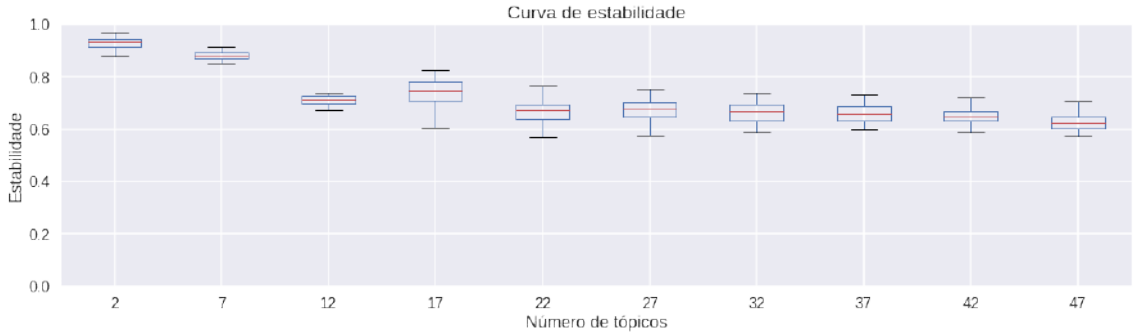


(d) 600 documentos

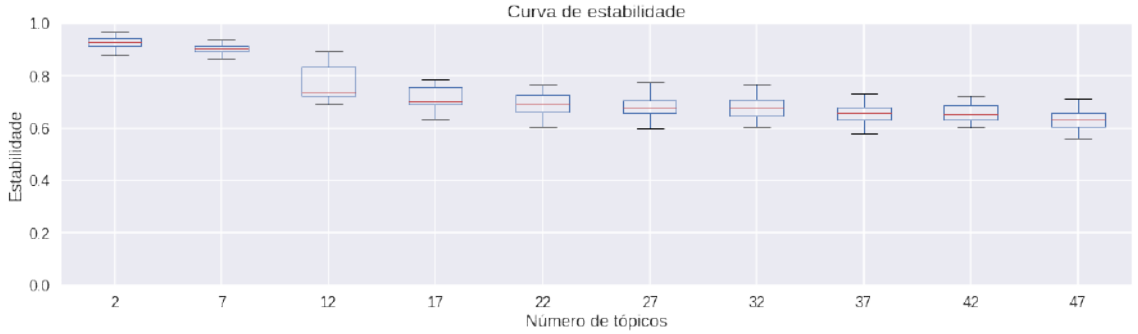
Figura 5.18: Impacto do ruído na qualidade de dados dos atributos.



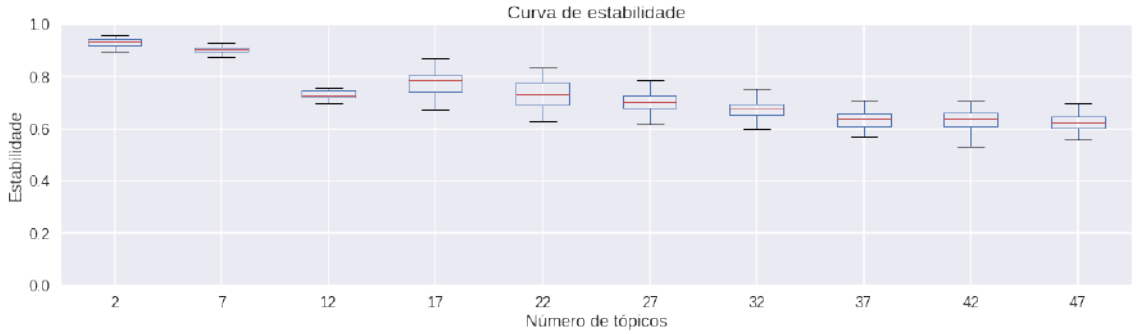
(a) Extração original



(b) 100 documentos



(c) 300 documentos



(d) 600 documentos

Figura 5.19: Impacto da introdução de ruído na estabilidade do *corpus*.

escolha pelo número final de tópicos neste último nível de ruído mostra-se mais equilibrada, pouco distinguindo-se entre 7 e 13, ainda que o maior nível médio foi obtido para 8 tópicos.

5.3.4 Impacto no resultado

A Figura 5.21 mostra o resultado obtido para todas as extração de 6 tópicos, escolha sugerida pelo SPT para a base de dados original.

O modelo de tópicos encontrou os mesmos tópicos, independentemente da quantidade de documentos com problemas de qualidade tenham sido introduzidos na base de dados original. De uma maneira geral, todos os documentos em língua portuguesa foram agrupados em dois tópicos: um deles exclusivamente composto por *stop words* da língua portuguesa, que agregou a maior parte das notícias publicadas em português; e o outro, que agregou as notícias que são compostas apenas por uma foto e duas sentenças de descrição.

As notícais deste segundo grupo geralmente não atendem ao critério mínimo de completude e, na etapa de qualidade de dados, caso não fossem eliminadas pelo critério de acurácia, seriam eliminadas pelo critério de completude. Este segundo tópico encontra-se separado do primeiro, visto que nem sempre as *stop words* da língua portuguesa estão presentes.

Os dois tópicos surgiram em detrimento do tópico sobre petróleo, incorporado pelo tópico da relação econômica China-Estados Unidos, e o tópico sobre carnaval, que foi incorporado ao tópico da disputa presidencial de 2018.

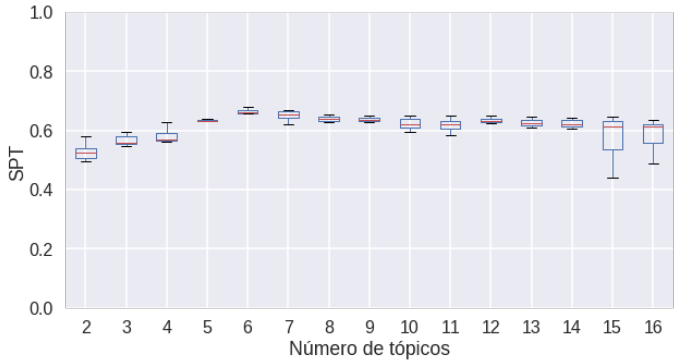
Qualitativamente, e desconsiderando os tópicos que agregaram os documentos em língua portuguesa, a coerência da extração de tópicos realizada após a introdução dos documentos com problemas de qualidade é inferior, com tópicos muito gerais e que misturam assuntos pouco afins, como a disputa presidencial e o carnaval.

Ao aumentar para 8 o número de tópicos extraídos, os dois tópicos da extração original que deram lugar aos tópicos de língua portuguesa voltam a aparecer, como mostrado na Figura 5.22.

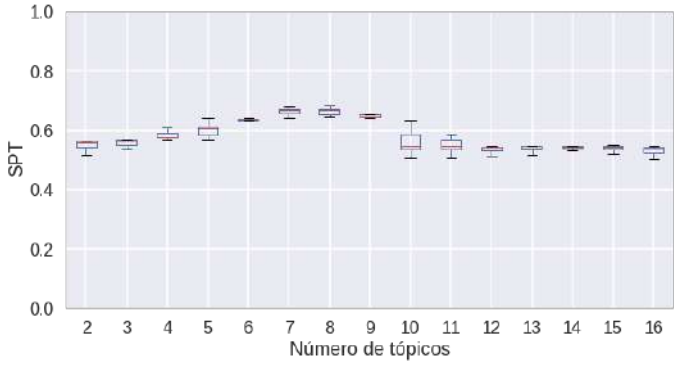
Por fim, a Figura 5.23 mostra e extração para 15 tópicos. O padrão dos casos anteriores se repete, com o isolatamento das notícias em língua portuguesa em dois tópicos, independente do número de notícias com problemas de qualidade.

5.4 Discussão

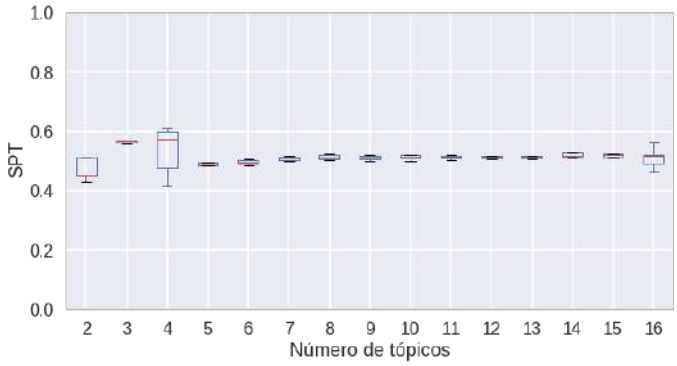
A medida de estabilidade mostrou que a base de dados anotada é instável apenas para as escolhas por 2 e 3 tópicos, enquanto a base não-anotada não apresentou sinais de instabilidade, nem mesmo com a introdução de diferentes níveis de ruído.



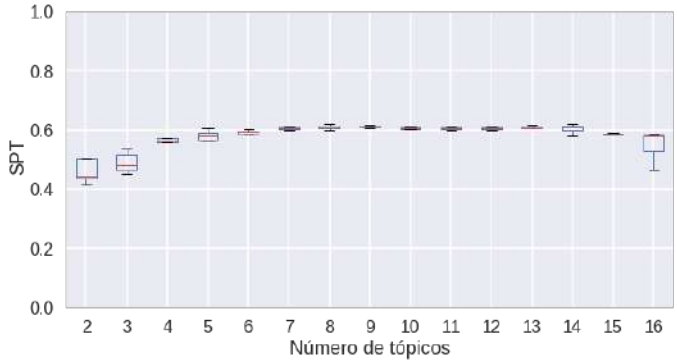
(a) Extração original



(b) 100 documentos

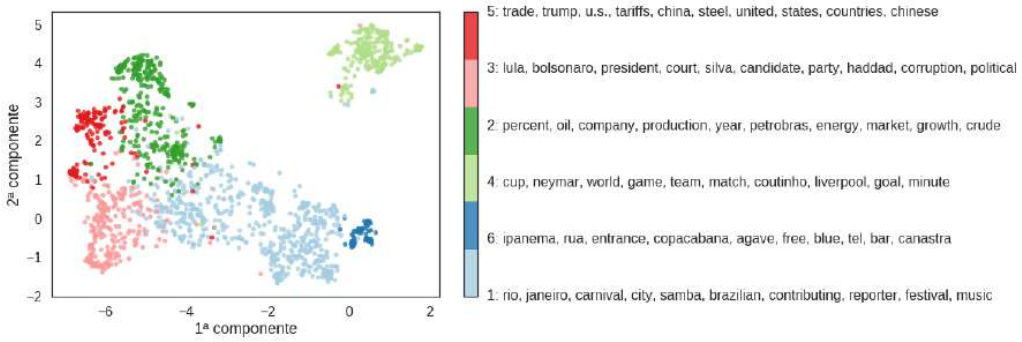


(c) 300 documentos

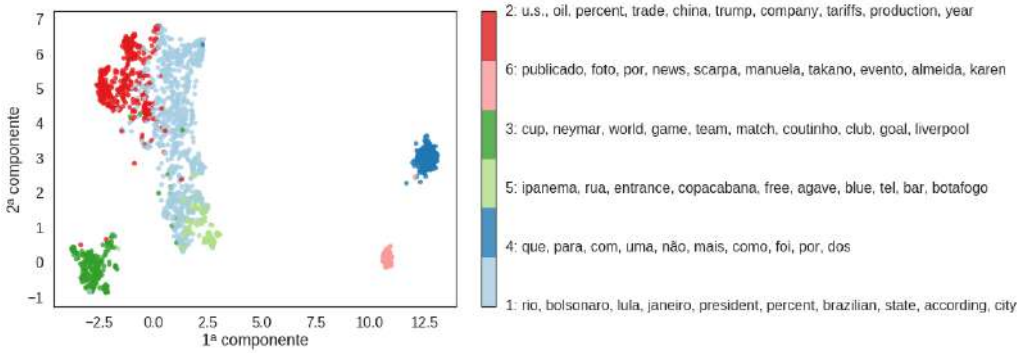


(d) 600 documentos

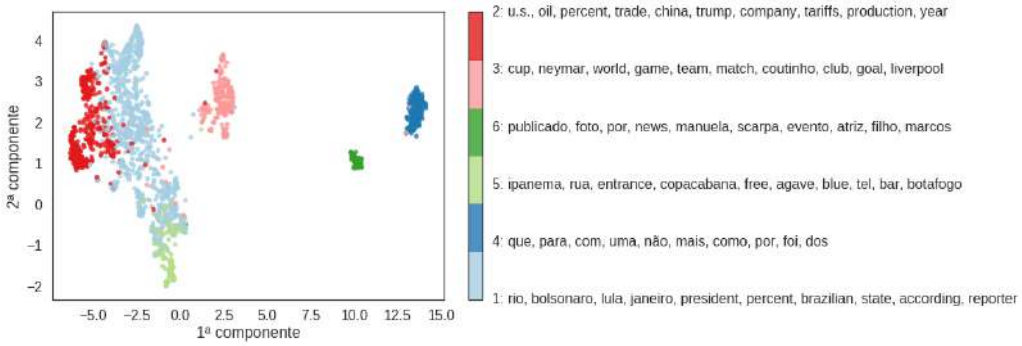
Figura 5.20: Impacto da introdução de ruído na seleção do número de tópicos.



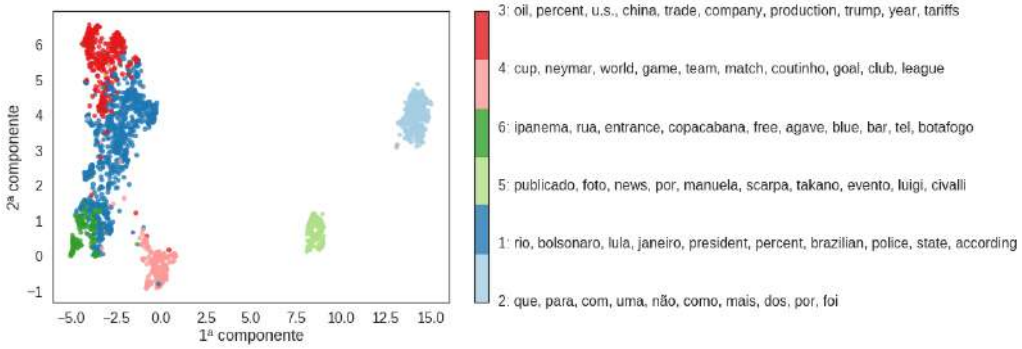
(a) Extração original



(b) 100 documentos

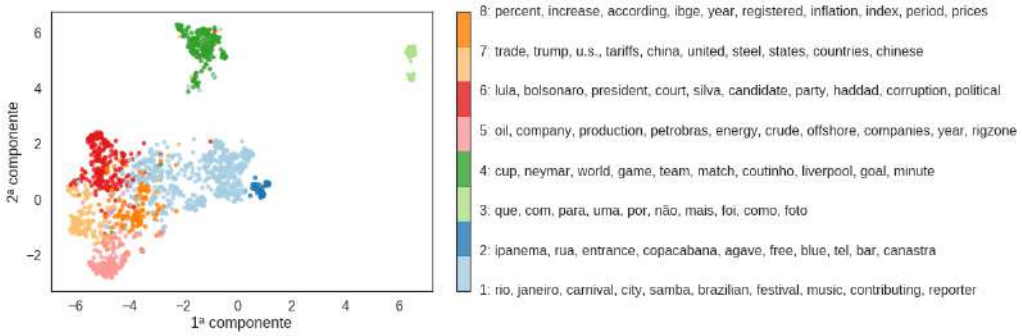


(c) 300 documentos

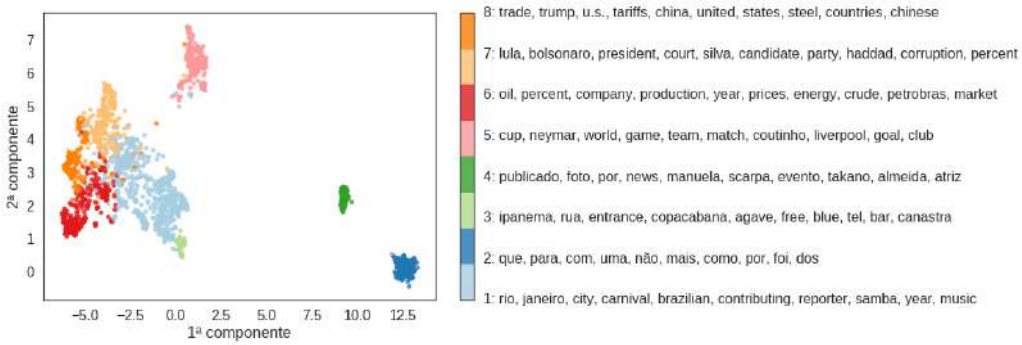


(d) 600 documentos

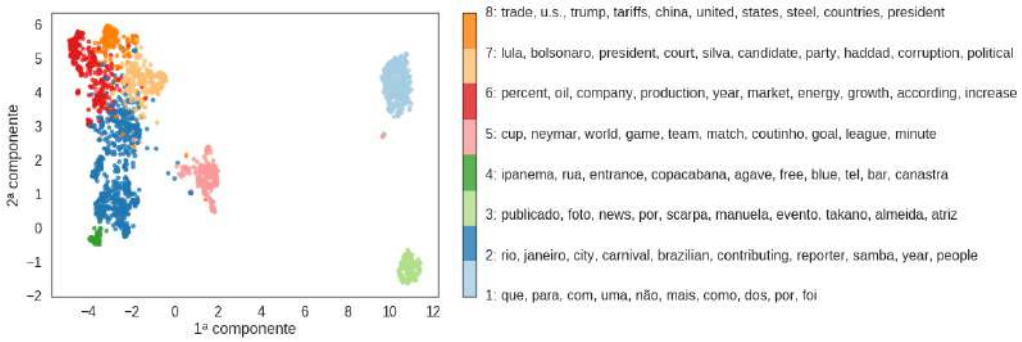
Figura 5.21: Impacto da introdução de ruído no resultado da extração de 6 tópicos.



(a) 100 documentos

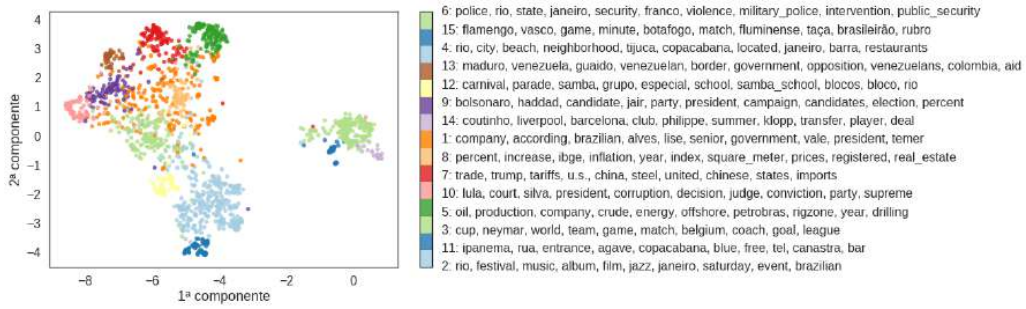


(b) 300 documentos

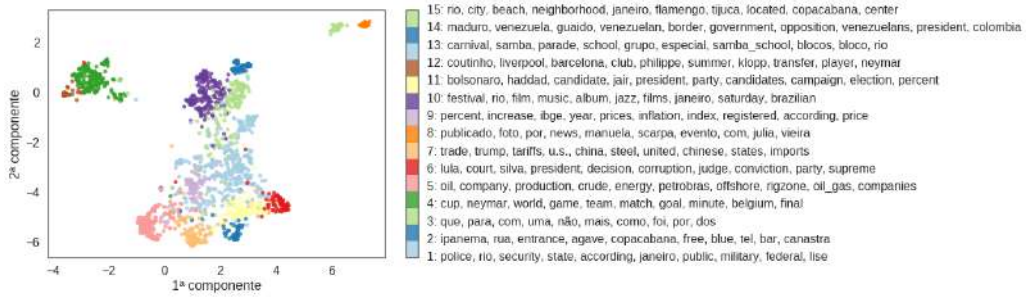


(c) 600 documentos

Figura 5.22: Resultado da extração de 8 tópicos.

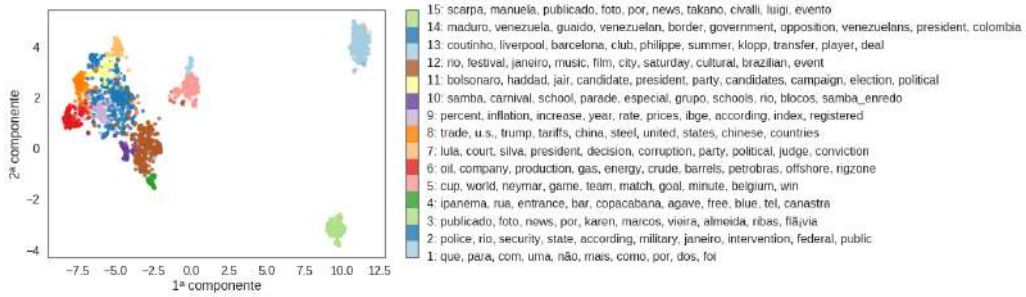


(a) Extração original



(b) 100 documentos

(c) 300 documentos



(d) 600 documentos

Figura 5.23: Impacto da introdução de ruído no resultado da extração de 15 tópicos.

Para selecionar o número de tópicos, optou-se por analisar o SPT. Para a base de dados anotada, esta medida indicou o mesmo número de tópicos da anotação, e mostrou uma tendência de queda com o aumento do número de tópicos.

Uma análise qualitativa dos tópicos encontrados para a extração com 12 tópicos mostrou que a qualidade do resultado encontrado foi de fato inferior devido à presença de tópicos muito específicos.

Já para a base não-anotada, o SPT não apresentou uma tendência uniforme para os diferentes níveis de ruído. De uma maneira geral, para a extração original e as bases com problemas de qualidade, a medida estabilizou-se a partir de um determinado número de tópicos, sem indicar uma escolha clara por um número final. Isso sugere inicialmente que as escolhas por estes diferentes números de tópicos são igualmente coerentes do ponto de vista semântico.

Qualitativamente, isto é confirmado pela inspeção dos tópicos para a extração com 15 tópicos. Os assuntos presentes da extração por 6 tópicos dão origem a tópicos mais detalhadas na extração por 15 tópicos, e ambas as extrações possuem tópicos coerentes para um avaliador humano, apesar do diferente nível de detalhamento.

Observou-se ainda que o nível de SPT foi degradado com a introdução de 100 e 300 documentos com problemas de qualidade, mas voltou a subir com a introdução de 600 documentos. No entanto, a qualidade da extração com o número de tópicos igual a 6 foi claramente inferior para as bases com problemas de qualidade, devido à presença dos dois tópicos com palavras em português.

Ainda assim, a coerência não foi capaz de capturar esta perda de qualidade porque a medida utilizada neste trabalho é calculada a partir da coocorrência entre termos da própria base de dados, o que é modificado com a introdução de mais documentos de língua portuguesa. Caso a coerência adotada neste trabalho utilizasse a estatística de um *corpus* externo, a tendência é que a nota de coerência fosse menor.

Por fim, o modelo de tópicos pareceu isolar as notícias com problemas de qualidade do restante dos tópicos encontrados, independentemente da quantidade de documentos ruidosos. Ainda assim, a extração com análise de qualidade de dados e posterior remoção dos documentos ruidosos apresentou o melhor resultado, após uma avaliação qualitativa.

Capítulo 6

Conclusões

Os resultados mostraram que os critérios de qualidade de dados adotados neste trabalho contribuíram com a maior qualidade da extração de tópicos, quando a extração de tópicos feita na base de dados com filtragem de qualidade apresentou nível médio de coerência superior à extração da base sem remoção dos documentos ruidosos.

Embora o modelo de tópicos tenha apresentado robustez ao ruído introduzido nos documentos, separando os documentos ruidosos dos demais, este resultado não necessariamente se repete para outros tipos de problemas de qualidade. Dessa forma, a qualidade de dados mostra-se como um fator determinante para garantir o nível de qualidade dos documentos, e assim não diminuir o nível de coerência dos tópicos encontrados.

No entanto, a definição de dimensões de qualidade por si só não garante um bom projeto de extração de tópicos, visto que ainda é preciso ainda escolher o número de tópicos. Para esta escolha, foram utilizadas como guias as medidas de estabilidade e SPT, cuja definição é baseada na medida de coerência. Estas medidas eliminaram escolhas ruins de número de tópicos, ajudando assim a diminuir o intervalo possível de escolhas. Levando em conta o resultado obtido na base anotada, o uso conjugado de ambas as medidas escolheram com sucesso o número correto de tópicos.

Ainda assim, não foi encontrada uma maneira de conjugar estabilidade, coerência e o número de tópicos para encontrar extrações mais generalistas ou detalhistas. Ainda, a medida de coerência utilizada não foi efetiva na base de dados com a presença de documentos em outro idioma.

6.1 Trabalhos futuros

A medida de estabilidade utilizada neste trabalho leva em consideração apenas as associações de palavras e tópicos. Uma possível melhoria consiste em considerar as associações entre documentos e tópicos.

Já para a medida de coerência, uma extensão do trabalho é o uso de uma base de dados externa ao *corpus* sob análise para o cálculo das notas de coerência. Dessa forma, espera-se que a SPT confira notas de coerência mais baixas para tópicos em outro idioma, por exemplo.

É importante ressaltar também que os resultados obtidos nas referências bibliográficas consultadas para este trabalho foram obtidos para a língua inglesa, e não generalizam automaticamente para a língua portuguesa.

Ainda, ambas as medidas não possuem uma calibração, o que não permite dizer, apenas pela diferença numérica, se uma extração é muito mais estável ou coerente do que a outra. Esta calibração é um importante desdobramento deste trabalho.

Quanto à análise da qualidade dos dados, supôs-se durante este trabalho que a janela de tempo contemplada na extração não era suficiente para variações significativas no dicionário, como mudanças de entidades ou criação de neologismos. Para extrações que envolvam janelas temporais mais longas, é necessário estender este trabalho para incorporar o aspecto temporal da língua.

No tocante ao modelo de tópicos, a característica do número de tópicos de funcionar como o grau de detalhamento da base de dados sugere o uso de modelos hierárquicos para a extração dos grupos. Dessa forma, seria possível detalhar apenas um grupo específico, de forma a controlar os grupos de forma mais efetiva.

Referências Bibliográficas

[1] "Número de usuários na web".
Acessado em fevereiro de 2019. http://www.internetlivestats.com/internet-users/#trend.

[2] "The digital universe in 2020".
Acessado em fevereiro de 2019. http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

[3] KAWAMOTO, K. *Digital journalism: Emerging media and the changing horizons of journalism*. Rowman & Littlefield Publishers, 2003.

[4] "Jornal do Brasil".
Acessado em fevereiro de 2019. www.jb.com.br.

[5] "Americans' online news use is closing in on TV news use".
Acessado em fevereiro de 2019. http://www.pewresearch.org/fact-tank/2017/09/07/americans-online-news-use-vs-tv-news-use/, .

[6] "Audiência dos telejornais em queda."
Acessado em fevereiro de 2019. http://www.economist.com/node/13642689.

[7] "Pesquisa sobre o número de usuários online e seus hábitos."
Acessado em fevereiro de 2019. http://stateofthemedia.org/.

[8] ALLCOTT, H., GENTZKOW, M. "Social media and fake news in the 2016 election", *Journal of Economic Perspectives*, v. 31, n. 2, pp. 211–36, 2017.

[9] SHAO, C., CIAMPAGLIA, G. L., VAROL, O., et al. "The spread of fake news by social bots", *arXiv preprint arXiv:1707.07592*, pp. 96–104, 2017.

[10] CONROY, N. J., RUBIN, V. L., CHEN, Y. "Automatic deception detection: Methods for finding fake news". In: *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, p. 82. American Society for Information Science, 2015.

[11] LAZER, D. M., BAUM, M. A., BENKLER, Y., et al. "The science of fake news", *Science*, v. 359, n. 6380, pp. 1094–1096, 2018.

[12] STEINKRAUS, D. W. "Method and apparatus for concept searching using a Boolean or keyword search engine". mar. 26 2002. US Patent 6,363,373.

[13] BLEI, D. M. "Introduction to Probabilistic Topic Models", *Communications of the ACM*, 2011. Disponível em: <http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf>.

[14] "Scopus".
Acessado em fevereiro de 2019. http://www.scopus.com.

[15] CHO, J., GARCIA-MOLINA, H., PAGE, L. "Efficient crawling through URL ordering", *Computer Networks and ISDN Systems*, v. 30, n. 1-7, pp. 161–172, 1998.

[16] HEYDON, A., NAJORK, M. "Mercator: A scalable, extensible web crawler", *World Wide Web*, v. 2, n. 4, pp. 219–229, 1999.

[17] SHKAPENYUK, V., SUEL, T. "Design and implementation of a high-performance distributed web crawler". In: *Data Engineering, 2002. Proceedings. 18th International Conference on*, pp. 357–368. IEEE, 2002.

[18] NAJORK, M., HEYDON, A. "High-performance web crawling". In: *Handbook of massive data sets*, Springer, pp. 25–45, 2002.

[19] TANEV, H., PISKORSKI, J., ATKINSON, M. "Real-time news event extraction for global crisis monitoring". In: *Natural Language and Information Systems*, Springer, pp. 207–218, 2008.

[20] SOUSA, J. P. "Uma história breve do jornalismo no Ocidente", .

[21] GILLIES, J., CAILLIAU, R. *How the Web Was Born: The Story of the World Wide Web*. OXFORD, 2000.

[22] "Empresários bancam campanha contra o PT pelo WhatsApp".
Acessado em fevereiro de 2019. https://www1.folha.uol.com.br/poder/2018/10/empres

[23] "Growth in mobile news use driven by older adults".
Acessado em fevereiro de 2019. http://www.pewresearch.org/fact-tank/2017/06/12/growth-in-mobile-news-use-driven-by-older-adults/, .

[24] BOND, R. M., FARISS, C. J., JONES, J. J., et al. "A 61-million-person experiment in social influence and political mobilization", *Nature*, v. 489, n. 7415, pp. 295, 2012.

[25] GOODE, L. "Social news, citizen journalism and democracy", *New media & society*, v. 11, n. 8, pp. 1287–1305, 2009.

[26] "CompusServe".
Acessado em fevereiro de 2019. `http://webcenters.netscape.compuserve.com/menu/`.

[27] BALDESSAR, M. J. "Mundo digital: Jornal do Brasil na Internet no tempo do PC 386", *7º encontro da ALCAR–Associação Brasileira de Pesquisadores de História da Mídia*, 2009.

[28] "O GLOBO NA REDE".
Acessado em fevereiro de 2019. `http://memoria.oglobo.globo.com/linha-do-tempo/o-globo-na-rede-9200005`.

[29] PRIMO, A., TRÄSEL, M. "Webjornalismo participativo e a produção aberta de notícias", *Revista Contracampo*, , n. 14, pp. 37–53, 2006.

[30] PINHEIRO, G. "O cidadão-repórter e o papel do jornalista profissional através do jornalismo participativo". In: *Intercom–Sociedade Brasileira de Estudos Interdisciplinares da Comunicação, XIV Congresso de Ciências da Comunicação na Região Sudeste [http://www. intercom. org. br/papers/regionais/sudeste2009/resumos/R14-0289-1. pdf, acedido em 16/09/2013]*, 2009.

[31] "A HISTORY OF JOURNALISM ON THE INTERNET: A state of the art and some methodological trends".
Acessado em fevereiro de 2019.
`http://revistainternacionaldehistoriadelacomunicacion.org/n%C3%BAmeros-anteriores/item/download/63_3c074aafce5451dc31a846a0146f2b20`.

[32] HENZINGER, M., CHANG, B.-W., MILCH, B., et al. "Query-free news search", *World Wide Web*, v. 8, n. 2, pp. 101–126, 2005.

[33] RADEV, D. R., OTTERBACHER, J., WINKEL, A., et al. "NewsInEssence: summarizing online news topics", 2005.

[34] BALAHUR, A., STEINBERGER, R., KABADJOV, M., et al. "Sentiment analysis in the news", *arXiv preprint arXiv:1309.6202*, 2013.

[35] BLEI, D. M. "Probabilistic topic models", *Communications of the ACM*, v. 55, n. 4, pp. 77–84, 2012.

[36] BLEI, D. M., NG, A. Y., JORDAN, M. I. "Latent dirichlet allocation", *the Journal of machine Learning research*, v. 3, pp. 993–1022, 2003.

[37] STEYVERS, M., GRIFFITHS, T. "Probabilistic topic models", *Handbook of latent semantic analysis*, v. 427, n. 7, pp. 424–440, 2007.

[38] GREENE, D., CUNNINGHAM, P. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering". In: *Proc. 23rd International Conference on Machine learning (ICML'06)*, pp. 377–384. ACM Press, 2006.

[39] AGARWAL, S., GODBOLE, S., PUNJANI, D., et al. "How much noise is too much: A study in automatic text classification". In: *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pp. 3–12. IEEE, 2007.

[40] SCHOLKOPF, B., SMOLA, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[41] MCCALLUM, A., NIGAM, K., OTHERS. "A comparison of event models for naive bayes text classification". In: *AAAI-98 workshop on learning for text categorization*, v. 752, pp. 41–48. Citeseer, 1998.

[42] SUBRAMANIAM, L. V., ROY, S., FARUQUIE, T. A., et al. "A survey of types of text noise and techniques to handle noisy text". In: *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, pp. 115–122. ACM, 2009.

[43] VINCIARELLI, A. "Noisy text categorization", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 27, n. 12, pp. 1882–1895, 2005.

[44] FAIER, J. M. *ANALISE DE COMPONENTES INDEPENDENTES PARA A MONITORAC AO DA QUALIDADE DE DADOS EM SÉRIES TEMPORAIS*. Tese de Doutorado, Universidade Federal do Rio de Janeiro, 2011.

[45] GUO, A., LIU, X., SUN, T. "Research on Key Problems of Data Quality in Large Industrial Data Environment". In: *Proceedings of the 3rd International Conference on Robotics, Control and Automation*, pp. 245–248. ACM, 2018.

[46] PROVOST, F., FAWCETT, T. "Data science and its relationship to big data and data-driven decision making", *Big data*, v. 1, n. 1, pp. 51–59, 2013.

[47] "Poor-Quality Data Imposes Costs and Risks on Businesses, Says New Forbes Insights Report".
Acessado em fevereiro de 2019. `https://www.forbes.com/sites/forbespr/2017/05/31/poor-quality-data-imposes-costs-and-risks-on-businesses-says-new-forbes-in#3af64b1f452b`.

[48] ENGLISH, L. P. "Total quality data management (TQdM)". In: *Information and database quality*, Springer, pp. 85–109, 2002.

[49] "ISO 8000".
Acessado em fevereiro de 2019. `https://www.iso.org/standard/50798.html`.

[50] BATINI, C., SCANNAPIECO, M. *Data and information quality: dimensions, principles and techniques*. Springer, 2016.

[51] GRAEL, F. F. *ATLASOM: Processamento de Texto para a Gerência de uma Colaboraćão Científica de Grande Porte*. Tese de Mestrado, UFRJ, 2013.

[52] WEBSTER, J. J., KIT, C. "Tokenization as the initial phase in NLP". In: *Proceedings of the 14th conference on Computational linguistics-Volume 4*, pp. 1106–1110. Association for Computational Linguistics, 1992.

[53] MIKOLOV, T., SUTSKEVER, I., CHEN, K., et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[54] COHEN, M. B., ELDER, S., MUSCO, C., et al. "Dimensionality reduction for k-means clustering and low rank approximation". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 163–172. ACM, 2015.

[55] SCHOFIELD, A., MIMNO, D. "Comparing apples to apple: The effects of stemmers on topic models", *Transactions of the Association for Computational Linguistics*, v. 4, pp. 287–300, 2016.

[56] TURNEY, P. D., PANTEL, P. "From frequency to meaning: Vector space models of semantics", *Journal of artificial intelligence research*, v. 37, pp. 141–188, 2010.

[57] MIKOLOV, T., CHEN, K., CORRADO, G., et al. "Efficient estimation of word representations in vector space", *arXiv preprint arXiv:1301.3781*, 2013.

[58] AIZAWA, A. "An information-theoretic perspective of tf–idf measures", *Information Processing & Management*, v. 39, n. 1, pp. 45–65, 2003.

[59] ZHANG, Y., JIN, R., ZHOU, Z.-H. "Understanding bag-of-words model: a statistical framework", *International Journal of Machine Learning and Cybernetics*, v. 1, n. 1-4, pp. 43–52, 2010.

[60] SINGHAL, A., BUCKLEY, C., MITRA, M. "Pivoted document length normalization". In: *ACM SIGIR Forum*, v. 51, pp. 176–184. ACM, 2017.

[61] HE, Q., CHANG, K., LIM, E.-P., et al. "Keep it simple with time: A reexamination of probabilistic topic detection models", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 32, n. 10, pp. 1795–1808, 2010.

[62] CICHOCKI, A., ANH-HUY, P. "Fast local algorithms for large scale nonnegative matrix and tensor factorizations", *IEICE transactions on fundamentals of electronics, communications and computer sciences*, v. 92, n. 3, pp. 708–721, 2009.

[63] WANG, X., GRIMSON, E. "Spatial latent dirichlet allocation". In: *Advances in neural information processing systems*, pp. 1577–1584, 2008.

[64] PEROTTE, A. J., WOOD, F., ELHADAD, N., et al. "Hierarchically supervised latent Dirichlet allocation". In: *Advances in Neural Information Processing Systems*, pp. 2609–2617, 2011.

[65] HOFMANN, T. "Unsupervised learning by probabilistic latent semantic analysis", *Machine learning*, v. 42, n. 1-2, pp. 177–196, 2001.

[66] CICHOCKI, A., ZDUNEK, R., PHAN, A. H., et al. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[67] LEE, D. D., SEUNG, H. S. "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems*, pp. 556–562, 2001.

[68] GAUSSIER, E., GOUTTE, C. "Relation between PLSA and NMF and implications". In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 601–602. ACM, 2005.

[69] DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., et al. "Indexing by latent semantic analysis", *Journal of the American society for information science*, v. 41, n. 6, pp. 391–407, 1990.

[70] GOLUB, G. H., REINSCH, C. "Singular value decomposition and least squares solutions". In: *Linear Algebra*, Springer, pp. 134–151, 1971.

[71] BERRY, M. W., FIERRO, R. D. "Low-rank Orthogonal Decompositions for Information Retrieval Applications", *Numerical linear algebra with applications*, v. 3, n. 4, pp. 301–327, 1996.

[72] DEERWESTER, S. "Improving information retrieval with latent semantic indexing", 1988.

[73] PAPADIMITRIOU, C. H., TAMAKI, H., RAGHAVAN, P., et al. "Latent semantic indexing: A probabilistic analysis". In: *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pp. 159–168. ACM, 1998.

[74] HOFMANN, T. "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57. ACM, 1999.

[75] TEH, Y. W., NEWMAN, D., WELLING, M. "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation". In: *Advances in neural information processing systems*, pp. 1353–1360, 2006.

[76] YUAN, B., GAO, X., NIU, Z., et al. "Discovering Latent Topics by Gaussian Latent Dirichlet Allocation and Spectral Clustering", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, v. 15, n. 1, pp. 25, 2019.

[77] ZHENG, L., CAIMING, Z., CAIXIAN, C. "MMDF-LDA: an improved multi-modal latent Dirichlet allocation model for social image annotation", *Expert Systems with Applications*, v. 104, pp. 168–184, 2018.

[78] VALLE, D., ALBUQUERQUE, P., ZHAO, Q., et al. "Extending the Latent Dirichlet Allocation model to presence/absence data: A case study on North American breeding birds and biogeographical shifts expected from climate change", *Global change biology*, v. 24, n. 11, pp. 5560–5572, 2018.

[79] PAATERO, P., TAPPER, U. "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values", *Environmetrics*, v. 5, n. 2, pp. 111–126, 1994.

[80] XU, W., LIU, X., GONG, Y. "Document clustering based on non-negative matrix factorization". In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 267–273. ACM, 2003.

[81] ARORA, S., GE, R., MOITRA, A. "Learning topic models–going beyond SVD". In: *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pp. 1–10. IEEE, 2012.

[82] VACA, C. K., MANTRACH, A., JAIMES, A., et al. "A time-based collective factorization for topic discovery and monitoring in news". In: *Proceedings of the 23rd international conference on World wide web*, pp. 527–538. ACM, 2014.

[83] STEYVERS, M., GRIFFITHS, T. *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2007.

[84] O'CALLAGHAN, D., GREENE, D., CARTHY, J., et al. "An analysis of the coherence of descriptors in topic modeling", *Expert Systems with Applications*, v. 42, n. 13, pp. 5645–5657, 2015.

[85] GREENE, D., O'CALLAGHAN, D., CUNNINGHAM, P. "How Many Topics? Stability Analysis for Topic Models", *CoRR*, v. abs/1404.4606, 2014. Disponível em: <`http://arxiv.org/abs/1404.4606`>.

[86] BRUNET, J.-P., TAMAYO, P., GOLUB, T. R., et al. "Metagenes and molecular pattern discovery using matrix factorization", *Proceedings of the national academy of sciences*, v. 101, n. 12, pp. 4164–4169, 2004.

[87] FARRIS, J. S. "On the cophenetic correlation coefficient", *Systematic Zoology*, v. 18, n. 3, pp. 279–285, 1969.

[88] MIMNO, D., WALLACH, H. M., TALLEY, E., et al. "Optimizing semantic coherence in topic models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262–272. Association for Computational Linguistics, 2011.

[89] "Wikipedia".
Acessado em fevereiro de 2019. `https://www.wikipedia.org/`.

[90] RÖDER, M., BOTH, A., HINNEBURG, A. "Exploring the space of topic coherence measures". In: *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408. ACM, 2015.

[91] MILLIGAN, G. W., COOPER, M. C. "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, v. 50, n. 2, pp. 159–179, 1985.

[92] SALVADOR, S., CHAN, P. "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms". In: *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pp. 576–584. IEEE, 2004.

[93] ROUSSEEUW, P. J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *Journal of computational and applied mathematics*, v. 20, pp. 53–65, 1987.

[94] DAVIES, D. L., BOULDIN, D. W. "A cluster separation measure", *IEEE transactions on pattern analysis and machine intelligence*, , n. 2, pp. 224–227, 1979.

[95] SUGAR, C. A., JAMES, G. M. "Finding the number of clusters in a dataset: An information-theoretic approach", *Journal of the American Statistical Association*, v. 98, n. 463, pp. 750–763, 2003.

[96] CELEUX, G., SOROMENHO, G. "An entropy criterion for assessing the number of clusters in a mixture model", *Journal of classification*, v. 13, n. 2, pp. 195–212, 1996.

[97] TIBSHIRANI, R., WALTHER, G., HASTIE, T. "Estimating the number of clusters in a data set via the gap statistic", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, v. 63, n. 2, pp. 411–423, 2001.

[98] LIU, A., SCHISTERMAN, E. F. "Principal component analysis", *Encyclopedia of Biopharmaceutical Statistics. New York: Marcel Dekker*, 2004.

[99] NEWMAN, D., BALDWIN, T., CAVEDON, L., et al. "Visualizing search results and document collections using topic maps", *Web Semantics: Science, Services and Agents on the World Wide Web*, v. 8, n. 2-3, pp. 169–175, 2010.

[100] MAATEN, L. V. D., HINTON, G. "Visualizing data using t-SNE", *Journal of machine learning research*, v. 9, n. Nov, pp. 2579–2605, 2008.

[101] MCINNES, L., HEALY, J., MELVILLE, J. "Umap: Uniform manifold approximation and projection for dimension reduction", *arXiv preprint arXiv:1802.03426*, 2018.

[102] BOYD-BARRETT, O., RANTANEN, T. *The globalization of news*. Sage, 1998.

[103] YAN, X., GUO, J., LAN, Y., et al. "A biterm topic model for short texts". In: *Proceedings of the 22nd international conference on World Wide Web*, pp. 1445–1456. ACM, 2013.

[104] FUKUNAGE, K., NARENDRA, P. M. "A branch and bound algorithm for computing k-nearest neighbors", *IEEE transactions on computers*, , n. 7, pp. 750–753, 1975.

[105] "Palavras reservadas - JavaScript".
Acessado em fevereiro de 2019. `https://www.w3schools.com/js/js_reserved.asp`.

[106] "Palavras reservadas - CSS".
Acessado em fevereiro de 2019. `https://www.w3schools.com/cssref/`.

[107] "Compact Language Detector 2". `https://github.com/CLD2Owners/cld2#internals`. Accessed: 2018-05-08.

[108] HANSBERRY, D. R., AGARWAL, N., SHAH, R., et al. "Analysis of the readability of patient education materials from surgical subspecialties", *The Laryngoscope*, v. 124, n. 2, pp. 405–412, 2014.

[109] HONG, L., DAVISON, B. D. "Empirical study of topic modeling in twitter". In: *Proceedings of the first workshop on social media analytics*, pp. 80–88. ACM, 2010.

[110] FERREIRA, F. G. *IDENTIFICAC AO DE EVENTOS BASEADA NA COMBINAC AO DE DETECTORES DE ALTAS ENERGIAS COM DIFERENTES TECNOLOGIAS E SEGMENTAC OES*. Tese de Doutorado, Universidade Federal do Rio de Janeiro, 2012.