UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelos para análise de textos: um comparativo do número de tópicos

Edvaldo Capobiango Coelho Filho

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP
SÃO CARLOS

Edvaldo Capobiango Coelho Filho

Modelos para análise de textos: um comparativo do número de tópicos

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Daiane Aparecida Zuanetti

USP – São Carlos
Outubro de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

C245m

Capobiango Coelho Filho, Edvaldo
Modelos para análise de textos: um comparativo do número de tópicos / Edvaldo Capobiango Coelho Filho; orientadora Daiane Aparecida Zuanetti. -- São Carlos, 2024.
65 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Inferência Bayesiana. 2. Latent Dirichlet allocation. 3. Métricas de desempenho. 4. Modelagem de tópicos. 5. Modelo de mistura. I. Aparecida Zuanetti, Daiane, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Edvaldo Capobiango Coelho Filho

Models for text analysis: a comparison of the number of topics

Dissertation submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Daiane Aparecida Zuanetti

USP – São Carlos
October 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Edvaldo Capobiango Coelho Filho, realizada em 27/08/2024.

Comissão Julgadora:

Profa. Dra. Daiane Aparecida Zuanetti (UFSCar)

Prof. Dr. Carlos Tadeu Pagani Zanini (UFRJ)

Prof. Dr. Erlandson Ferreira Saraiva (UFMS)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

Por não medirem esforços para que este momento fosse possível,
dedico este trabalho aos meus pais,
Edvaldo e Niara.

AGRADECIMENTOS

Agradeço aos meus pais, Edvaldo e Niara, por serem a minha fonte de inspiração e superação, meus maiores incentivadores, acreditarem na minha capacidade, em momentos que até mesmo eu duvidava, e por todo o amor de sempre, sem eles com certeza eu não teria conseguido chegar até aqui e concluir este ciclo.

Agradeço ao meu cachorro, Rex, por me descontrair mesmo nos piores momentos.

Agradeço à minha orientadora, Daiane, por todos os ensinamentos, conversas, conselhos, paciência, parceria e ajuda em todos os momentos, sua presença foi crucial.

Agradeço aos demais professores e funcionários do Departamento de Estatística da UFSCar e ICMC da USP.

Agradeço aos professores Carlos e Erlandson, por terem aceitado fazer parte da banca examinadora do meu trabalho, fazendo ótimas contribuições.

Agradeço aos meus amigos, por estarem comigo em diversos momentos, sejam eles de alegria ou tristeza, aguentarem meus dramas e não desistirem de mim.

Agradeço aos meus familiares, por fazerem parte do que sou hoje.

Agradeço a todos que, de alguma forma, contribuíram para a realização desta etapa.

“Seja a mudança que você quer ver no mundo.”
(Mahatma Gandhi)

RESUMO

COELHO FILHO, E. C. **Modelos para análise de textos: um comparativo do número de tópicos**. 2024. 65 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

A modelagem de textos tem ganhado bastante visibilidade e popularidade nos últimos anos devido a grande e, cada vez maior, quantidade de informações presentes no dia a dia, consumidas de diversas maneiras. Para a eficiência e aplicabilidade destes modelos, é de suma importância a etapa de pré-processamento dos dados, que ajuda na organização e tratamento dos textos. Um ramo dentro da análise de textos é o de modelagem de tópicos, cujas metodologias visam entender a estrutura de tópicos (assuntos) que formam um documento, segmentando vários documentos por seus tópicos dominantes e simplificando assim a exploração de grandes volumes de dados textuais com a redução de dimensionalidade ocasionada. Um dos métodos pioneiros neste contexto é o Modelo de Mistura (MM), este que parte-se do pressuposto de que cada documento será composto de palavras advindas de um único tópico. Diante dessa limitação, tem ganhado bastante visibilidade o modelo de *Latent Dirichlet Allocation* (LDA), por conta de sua maior flexibilidade, visto que permite que cada documento possa exibir vários tópicos. Em ambas as metodologias, a inferência é realizada, em geral, via abordagem Bayesiana. No entanto, uma das características do MM e LDA consiste na exigência de que o usuário defina de partida a quantidade de tópicos do modelo. Sendo assim, o uso de métricas de desempenho se faz necessário após a aplicação do método, visando a ajuda na definição e estimação do melhor número de tópicos a ser escolhido. Nesse trabalho, portanto, além de contrapor as metodologias de análises textuais, fazemos o comparativo entre as métricas que mensuram a qualidade dos modelos e são utilizadas para a escolha do número de tópicos. Para isso, aplicamos os modelos e as métricas de seleção em dois conjuntos de dados reais.

Palavras-chave: Inferência Bayesiana, *latent Dirichlet allocation*, métricas de desempenho, modelagem de tópicos, modelo de mistura.

ABSTRACT

COELHO FILHO, E. C. **Models for text analysis: a comparison of the number of topics**. 2024. 65 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Text modeling has gained significant visibility and popularity in recent years due to the large and ever-increasing amount of information present in daily life, consumed in various ways. For the efficiency and applicability of these models, the prior step of data preprocessing is of utmost importance, as it helps in the organization and treatment of texts. One branch within text analysis is topic modeling, whose methodologies aim to understand the topic structure that forms a document, segmenting multiple documents by their dominant topics (subjects) and thus simplifying the exploration of large volumes of textual data with the resulting dimensionality reduction. One of the pioneering methods in this context is the Mixture Model (MM), which assumes that each document will be composed of words from a single topic. Given this limitation, the technique of Latent Dirichlet Allocation (LDA) has gained considerable visibility due to its greater flexibility, as it allows each document to exhibit multiple topics. In both methodologies, model inference is generally given via a Bayesian approach. However, one of the characteristics of MM and LDA is the requirement that the user define the number of topics in the model from the outset. Therefore, the use of performance metrics becomes necessary after the application of the method, aiming to help in the definition and estimation of the best number of topics to be chosen. In this work, therefore, in addition to contrasting text analysis methodologies, we compare the metrics that measure the quality of the models and are used for choosing the number of topics. To do this, we apply the models and selection metrics to two sets of real data.

Keywords: Bayesian approach, latent Dirichlet allocation, performance metrics, topic modeling, mixture model.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Desde a Terceira Revolução Industrial, também conhecida como Revolução Tecnocientífica, ocorrida em meados do século XX, o mundo da tecnologia vem se desenvolvendo gradativamente, muito por conta da série de descobertas e evoluções ocorridas durante esse período histórico.

Diante disso, com o grande avanço da tecnologia, outros setores também apresentaram uma crescente, como o ramo das informações. Atrelado a isso, pode-se dizer que a enorme quantidade de dados disponibilizados, consumidos e gerados por todos, em diversos momentos, como notícias, redes sociais e serviços de *streaming*, são denominados informação. Estes são alguns dos motivos que justificam o fato de que vivemos na Era das Informações.

Neste contexto, faz-se cada vez mais necessária a utilização da estatística, ciência que, resumidamente, visa transformar dados em informações confiáveis para a tomada de decisões. Dentre os muitos segmentos em que a estatística pode ser aplicada, está a modelagem de textos, sendo que um ramo que tem tido destaque dentro desta categoria é o de modelagem de tópicos. A maioria das técnicas visa a simplificação da exploração de grandes volumes de dados textuais, pois sua finalidade é separar subconjuntos de textos por seus respectivos temas.

É importante destacar que quando estamos lidando com dados textuais, independentemente do método estatístico que será utilizado, uma etapa essencial é o pré-processamento de dados, que consiste no tratamento e organização das informações contidas nos textos.

Uma das metodologias pioneiras nas análises textuais foram os Modelos de Mistura (PEEL; MCLACHLAN, 2000) utilizados neste ramo com o intuito de agrupar conjuntos de documentos (textos) a partir dos seus tópicos (assuntos). Neste modelo, parte-se do pressuposto de que cada documento sempre será composto de palavras advindas de um único tópico.

Sobre os métodos de inferência tradicionalmente aplicados aos Modelos de Mistura para a análise de textos, um dos mais usuais é o Amostrador de Gibbs (CASELLA; GEORGE, 1992).

No entanto, devido a limitação dos Modelos de Mistura, em busca de uma heterogeneidade maior, Blei, Ng e Jordan (2003) propuseram a *Latent Dirichlet Allocation* (LDA), que também agrupa documentos, por seus tópicos dominantes. Porém, esta metodologia tem ganhado bastante visibilidade nos últimos anos devido a sua maior flexibilidade, visto que permite que cada documento possa exibir vários tópicos.

Em relação aos métodos de inferência do modelo LDA, os mais usuais são o método variacional, apresentado por Blei, Ng e Jordan (2003) e o Amostrador de Gibbs, exposto em Griffiths e Steyvers (2004).

Uma característica dos dois métodos sintetizados anteriormente consiste na exigência de que o usuário inicialize o algoritmo de estimação com um número de tópicos pré-definido. Porém, em determinadas ocasiões, o pesquisador pode não ter o conhecimento aprofundado sobre o conjunto de dados, acarretando assim em uma escolha inadequada desse valor solicitado e prejudicando a eficiência do método.

Com o passar do tempo, foram feitos estudos acerca de métricas que mensuram a qualidade dos modelos de tópicos e que podem, por consequência, ajudar na escolha e estimação do melhor número de tópicos para cada conjunto de dados analisados. Uma possibilidade, como proposto por Wallach *et al.* (2009), é usar a própria função de verossimilhança. Newman *et al.* (2007) utiliza a métrica de perplexidade. Com o intuito de avaliar a semântica dos tópicos, Mei, Shen e Zhai (2007) abordaram o *Pointwise Mutual Information* (PMI), enquanto sua variação, o *Normalized Pointwise Mutual Information* (NPMI) foi apresentado por Bouma (2009). Posteriormente, Mimno *et al.* (2011) propôs a *Log Conditional Probability* (LCP).

Em contrapartida, um modelo Bayesiano não-paramétrico que não exige a pré-especificação do número de tópicos é o *Hierarchical Dirichlet Process* (HDP), apresentado por Teh *et al.* (2004). Porém, por permitir um número máximo ilimitado de tópicos, é mais custoso computacionalmente para grandes conjuntos de dados em detrimento ao LDA e, em alguns casos, pode retornar um número de tópicos muito grande.

O principal objetivo deste trabalho, além de contrapor as metodologias de análises textuais (Modelos de Mistura e LDA), inclui também fazer um estudo comparativo entre as métricas que mensuram a qualidade dos modelos e são utilizadas para a escolha do número de tópicos. Assim, desejamos verificar as particularidades de cada medida e seus desempenhos na definição adequada do número de tópicos.

Este trabalho está organizado da seguinte maneira. No Capítulo 2, apresentamos, além de técnicas de pré-processamento dos dados, informações sobre o Modelo de Mistura e o LDA, como seus conceitos básicos, funções de verossimilhança, estimações, o Amostrador de Gibbs e, também, as métricas que serão estudadas para a verificação da qualidade do modelo. O Capítulo 3 consiste nas aplicações das metodologias e métricas em dois diferentes bancos de dados reais. Por fim, no Capítulo 4 temos as conclusões e uma proposta para trabalhos futuros.

CAPÍTULO

2

ANÁLISE DE TEXTOS

Neste capítulo são apresentadas, na Seção 2.1, algumas técnicas de pré-processamento dos dados. Informações sobre o Modelo de Mistura e o LDA são abordadas nas Seções 2.2 e 2.3, respectivamente. Na Seção 2.4 discute-se a principal diferença entre as duas metodologias, enquanto as métricas de mensuração da qualidade dos modelos é exposta na Seção 2.5.

2.1 Pré-processamento dos dados

Uma etapa indispensável quando estamos trabalhando com dados textuais é o seu pré-processamento, pois além de otimizar tempo computacional, aumenta a eficiência das metodologias utilizadas para a modelagem, visto que os principais objetivos deste passo são a estruturação, organização e tratamento das informações.

Dentre as diversas técnicas que podem ser usadas para esta finalidade, estão:

- Remoção de *Stopwords*: As ditas *stopwords* são definidas como palavras que aparecem com uma grande frequência nos textos, mas em geral são consideradas irrelevantes na análise de tópicos, como artigos e pronomes, por exemplo. Sendo assim, recomenda-se a remoção das *stopwords* antes da aplicação dos algoritmos de aprendizado. Uma lista de *stopwords* em português pode ser vista em Github (2020).

- *Stemming*: Processo que visa reduzir as diferentes formas de uma palavra ao seu radical comum. Um exemplo do que ocorre nesta etapa são as palavras “corra”, “corrida”, “correram”, que poderiam ser representadas por “corr”. A vantagem desta técnica é que, ao aplicá-la, será menor a quantidade de palavras presentes no dicionário[1] que irá ser trabalhado.

[1] O conjunto de todas as palavras dispostas em cada um dos textos é denominado de dicionário.

- *Case Folding*: Etapa cujo intuito é a padronização das letras das palavras em maiúsculas ou minúsculas.

- *Bag-of-words*: Representação utilizada em certas aplicações envolvendo dados de texto que tem como finalidade contabilizar a presença/ausência ou frequência das palavras do dicionário em cada um dos textos, ignorando-se a ordem em que as palavras ocorrem nos documentos. Para exemplificar seu funcionamento na prática, suponha que o dicionário estudado seja composto pelas palavras dos dois textos hipotéticos:

  1) Zeca odeia assistir TV;
  2) Ana ama assistir filmes.

  Com o *bag-of-words* teremos os textos estruturadas conforme a Tabela 1:

Tabela 1 – Exemplo *bag-of-words*.

	Zeca	odeia	assistir	TV	Ana	ama	filmes
Texto 1	1	1	1	1	0	0	0
Texto 2	0	0	1	0	1	1	1

- Tokenização: Fase em que ocorre a separação de cada uma das palavras, delimitando seus inícios e fins, desconsiderando, por exemplo, os espaços em branco entre uma palavra e outra. Para exemplificar, supondo que o dicionário estudado fosse composto por apenas dois textos, “Zeca ama correr” e “Ana odeia nadar”, os respectivos *tokens* atribuídos seriam “Zeca”, “ama”, “correr”, “Ana”, “odeia” e “nadar”.

Além dos métodos citados, é de suma importância também, por não agregarem informação da semântica dos textos, a remoção de pontuações, numerais e acentos.

No *software* R, existem diversos pacotes que englobam as técnicas de pré-processamento dos textos, de acordo com as finalidades desejadas, como, por exemplo, o *tokenizers* (MULLEN *et al.*, 2018), utilizado para o processo de tokenização das palavras e o *tm* (FEINERER, 2013), cuja aplicação se dá para a remoção de números, pontuações, espaços, etc.

2.2 Modelos de Mistura (MM)

Nesta Seção são apresentadas informações acerca dos Modelos de Mistura. Na Subseção 2.2.1 são abordadas sua origem e propriedades. A função de verossimilhança e a estimação Bayesiana do modelo são mostradas, respectivamente, nas Subseções 2.2.2 e 2.2.3. Na Subseção 2.2.4 discutimos sobre o amostrador de Gibbs.

2.2.1 Conceitos básicos

Os Modelos de Mistura tem como finalidade modelar dados oriundos de distribuições mais complexas, em que uma população é proveniente de determinadas subpopulações (TITTERINGTON; SMITH; MAKOV, 1985).

No âmbito dos conjuntos de dados textuais, propostos inicialmente por Peel e McLachlan (2000), essa técnica busca simplificar a grande gama de textos em determinados tópicos, simplificando assim sua dimensionalidade inicial.

Os Modelos de Mistura conduzem a sua modelagem a nível de cada documento, ou seja, cada documento sempre será composto de palavras oriundas de um único tópico. Para entender melhor o funcionamento do Modelo de Mistura na prática, considere as seguintes frases hipotéticas (documentos):

1. O Botafogo é time de futebol mais tradicional do Brasil;
2. Neymar é considerado pela FIFA o melhor jogador de futebol do mundo;
3. A pandemia do coronavírus está próxima do fim;
4. Brasil registra a menor média móvel de mortes pelo novo coronavírus desde janeiro;
5. Com a liberação de torcidas nos estádios de futebol, aumentam-se os índices de contágio da variante delta do coronavírus na Inglaterra.

Com o auxílio do Modelo de Mistura identificamos a qual tópico cada documento (frase) será designada. Então, supondo que, no exemplo acima, consideramos 3 tópicos, I, que será representado em geral por documentos em que o assunto seja futebolístico, II, que terá em geral documentos referentes ao coronavírus e III, que englobará tanto futebol quanto coronavírus. Assim, as palavras que formam as frases 1 e 2 são sorteadas a partir de uma distribuição de probabilidade sobre o dicionário de palavras que caracteriza o tópico I. As frases 3 e 4 são sorteadas a partir da distribuição sobre as palavras caracterizadas pelo tópico II. Por fim, de maneira análoga, a frase 5 será sorteada a partir das palavras que caracterizam o tópico III.

2.2.2 Função de verossimilhança

Seja $\mathbf{S} = (S_1, ..., S_m)$ um conjunto de variáveis aleatórias discretas e independentes, tal que $S_j \sim Discreta(\boldsymbol{\theta})$, em que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, K é o número de diferentes tópicos e m o número de documentos analisados. S_j representa de qual tópico as palavras do documento j serão sorteadas, com $j = 1, \ldots, m$ e $S_j \in \{1, \ldots, K\}$. Sendo assim, definimos:

$$P(S_j = k) = \theta_k, \tag{2.1}$$

em que $0 \leq \theta_k \leq 1$ e $\sum_{k=1}^{K} \theta_k = 1$.

Seja V o número de palavras do dicionário, dispostas conforme a estrutura *bag-of-words* e w_{jv} o número de vezes que a palavra v-ésima palavra ocorre no documento j, para $j = 1,...,m$ e $v = 1,...,V$. Denote por $\mathbf{w}_j = (w_{j1},\ldots,w_{jV})$, o vetor composto pelas quantidades de vezes que cada uma das V palavras do vocabulário aconteceram no documento j, sendo $\sum_{v=1}^{V} w_{jv} = n_{d_j}$ o número de palavras no documento j. Sendo assim, a matriz $\mathbf{X} = (\mathbf{w}_1,\ldots,\mathbf{w}_m)^T$ representa o *corpus* de documentos que formam a base de dados.

Neste contexto, temos que $\mathbf{w}_j = (w_{j1},\ldots,w_{jV})|S_j = k \sim Multinomial(n_{d_j}; \boldsymbol{\phi}_k = (\phi_{k1},\ldots,\phi_{kV}))$, em que ϕ_{kv} representa a probabilidade de ocorrência da v-ésima palavra no k-ésimo tópico e $\boldsymbol{\phi} = (\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_K)$). Com isso, a função de verossimilhança aumentada (devido à inclusão das variáveis latentes S) do modelo pode ser vista como

$$
\begin{aligned}
L(\boldsymbol{\theta},\boldsymbol{\phi}|\mathbf{S},\mathbf{X}) &= f_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x}) \\
&= f_{\mathbf{S}}(\mathbf{s}) f_{\mathbf{X}|\mathbf{S}}(\mathbf{x}|\mathbf{s}) \\
&= \Big(\prod_{j=1}^{m} P(S_j)\Big)\Big(\prod_{j=1}^{m} f(\mathbf{w}_j|S_j)\Big) \\
&= \prod_{k=1}^{K}\Bigg(\prod_{j=1}^{m}\theta_k^{I(S_j=k)}\prod_{j=1}^{m}\Bigg(\Bigg(\frac{n_{d_j}!}{w_{j1}!\ldots w_{jV}!}\Bigg)\phi_{k1}^{w_{j1}}\ldots\phi_{kv}^{w_{jV}}\Bigg)^{I(S_j=k)}\Bigg) \\
&= \prod_{k=1}^{K}\Bigg(\theta_k^{\sum_{j=1}^{m} I(S_j=k)}\prod_{j=1}^{m}\Bigg(\Bigg(\frac{n_{d_j}!}{w_{j1}!\ldots w_{jV}!}\Bigg)^{I(S_j=k)}\phi_{k1}^{w_{j1}I(S_j=k)}\ldots\phi_{kV}^{w_{jV}I(S_j=k)}\Bigg)\Bigg) \\
&= \prod_{k=1}^{K}\Bigg(\theta_k^{m_k}\phi_{k1}^{\sum_{j=1}^{m} w_{j1}I(S_j=k)}\ldots\phi_{kV}^{\sum_{j=1}^{m} w_{jV}I(S_j=k)}\prod_{j=1}^{m}\Bigg(\frac{n_{d_j}!}{w_{j1}!\ldots w_{jV}!}\Bigg)^{I(S_j=k)}\Bigg) \\
&\propto \prod_{k=1}^{K}\Bigg(\theta_k^{m_k}\phi_{k1}^{\sum_{j=1}^{m} w_{j1}I(S_j=k)}\ldots\phi_{kV}^{\sum_{j=1}^{m} w_{jV}I(S_j=k)}\Bigg),
\end{aligned}
$$

em que $m_k = \sum_{j=1}^{m} I(S_j = k)$ é o numero de documentos alocados no tópico k e $\sum_{j=1}^{m} w_{jv} I(S_j = k)$ é o número de vezes que a palavra v aparece em documentos do tópico k, em que $v \in \{1,\ldots,V\}$.

2.2.3 Estimação Bayesiana

Esse modelo geralmente é estimado através da abordagem Bayesiana, por considerar variáveis não observadas, $\mathbf{S}$, que dificultam a estimação e convergência de algoritmos através de outras abordagens. Uma alternativa seria a estimação frequentista, que utiliza o algoritmo de expectativa-maximização (EM), porém por se tratar de um modelo com muitos parâmetros e variáveis não observadas, além da matriz de documentos ser muito esparsa, provavelmente o algoritmo não seria eficiente por questões de convergência.

A distribuição *Dirichlet* é conjugada da distribuição $Multinomial$. Dessa maneira, combinando a verossimilhança com as seguintes distribuições *a priori*

1. $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \sim Dirichlet(\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K))$, em que $0 \leq \theta_k \leq 1$ e $\sum_{k=1}^{K} \theta_k = 1$;

2. $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kV}) \sim Dirichlet(\boldsymbol{\beta}_k = (\beta_{k1}, ..., \beta_{kV}))$, em que $0 \leq \phi_{kv} \leq 1$ e $\sum_{v=1}^{V} \phi_{kv} = 1$,

sendo $\alpha_k > 0$ e $\beta_{kv} > 0$ hiper-parâmetros conhecidos e a distribuição *Dirichlet* é melhor descrita no Apêndice A, podemos escrever a distribuição *a posteriori* conjunta do modelo como

$$\begin{aligned}
\pi(\boldsymbol{\theta}, \boldsymbol{\phi} | \mathbf{S}, \mathbf{X}) &\propto L(\boldsymbol{\theta}, \boldsymbol{\phi} | \mathbf{S}, \mathbf{X}) \pi(\boldsymbol{\theta}) \prod_{k=1}^{K} \pi(\boldsymbol{\phi}_k) \\
&\propto \prod_{k=1}^{K} \left(\theta_k^{m_k} \phi_{k1}^{\sum_{j=1}^{m} w_{j1} I(S_j = k)} \ldots \phi_{kV}^{\sum_{j=1}^{m} w_{jV} I(S_j = k)} \right) \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \right) \left(\prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{kv}^{\beta_{kv} - 1} \right) \\
&= \prod_{k=1}^{K} \left(\theta_k^{m_k} \phi_{k1}^{\sum_{j=1}^{m} w_{j1} I(S_j = k)} \ldots \phi_{kV}^{\sum_{j=1}^{m} w_{jV} I(S_j = k)} \theta_k^{\alpha_k - 1} \left(\prod_{v=1}^{V} \phi_{kv}^{\beta_{kv} - 1} \right) \right).
\end{aligned}$$

Desta forma, as distribuições *a posteriori* condicionais podem ser descritas como

1.

$$\begin{aligned}
\pi(\boldsymbol{\theta} | \cdots) &\propto \prod_{k=1}^{K} (\theta_k^{m_k} \theta_k^{\alpha_k - 1}) \\
&= \prod_{k=1}^{K} \theta_k^{m_k + \alpha_k - 1} \\
&\equiv Dirichlet(m_1 + \alpha_1, ..., m_k + \alpha_k),
\end{aligned}$$

em que $\cdots$ inclui todas as variáveis e parâmetros exceto o(s) parâmetro(s) em questão;

2.

$$\begin{aligned}
\pi(\boldsymbol{\phi}_k | \cdots) &\propto \left(\phi_{k1}^{\sum_{j=1}^{m} w_{j1} I(S_j = k)} ... \phi_{kV}^{\sum_{j=1}^{m} w_{jV} I(S_j = k)} \left(\prod_{v=1}^{V} \phi_{kv}^{\beta_{kv} - 1} \right) \right) \\
&= \left(\phi_{k1}^{\sum_{j=1}^{m} w_{j1} I(S_j = k)} ... \phi_{kV}^{\sum_{j=1}^{m} w_{jV} I(S_j = k)} \phi_{k1}^{\beta_{k1} - 1} ... \phi_{kV}^{\beta_{kV} - 1} \right) \\
&= \left(\phi_{k1}^{\sum_{j=1}^{m} w_{j1} I(S_j = k) + \beta_{k1} - 1} ... \phi_{kV}^{\sum_{j=1}^{m} w_{jV} I(S_j = k) + \beta_{kV} - 1} \right) \\
&= \prod_{v=1}^{V} \phi_{kv}^{\sum_{j=1}^{m} w_{jv} I(S_j = k) + \beta_{kv} - 1} \\
&\equiv Dirichlet(\sum_{j=1}^{m} w_{j1} I(S_j = k) + \beta_{k1}, ..., \sum_{j=1}^{m} w_{jV} I(S_j = k) + \beta_{kV}).
\end{aligned}$$

A probabilidade *a posteriori* de $S_j = k$ pode ser representada por

$$
\begin{aligned}
P(S_j = k|\mathbf{w}_j, \boldsymbol{\theta}, \boldsymbol{\phi}) &= \frac{f(S_j = k, \mathbf{w}_j|\boldsymbol{\theta}, \boldsymbol{\phi})}{f(\mathbf{w}_j|\boldsymbol{\theta}, \boldsymbol{\phi})} \\
&= \frac{f(S_j = k, \mathbf{w}_j|\boldsymbol{\theta}, \boldsymbol{\phi})}{\sum_{t=1}^{K} f(S_j = t, \mathbf{w}_j|\boldsymbol{\theta}, \boldsymbol{\phi})} \\
&= \frac{P(S_j = k|\boldsymbol{\theta}) f(\mathbf{w}_j|S_j = k, \boldsymbol{\phi})}{\sum_{t=1}^{K} P(S_j = t|\boldsymbol{\theta}) f(\mathbf{w}_j|S_j = t, \boldsymbol{\phi})} \\
&= \frac{P(S_j = k|\boldsymbol{\theta}) f(\mathbf{w}_j|\boldsymbol{\phi}_k)}{\sum_{t=1}^{K} P(S_j = t|\boldsymbol{\theta}) f(\mathbf{w}_j|\boldsymbol{\phi}_t)} \\
&= \frac{\theta_k f(\mathbf{w}_j|\boldsymbol{\phi}_k)}{\sum_{t=1}^{K} \theta_t f(\mathbf{w}_j|\boldsymbol{\phi}_t)},
\end{aligned}
$$

para $k = 1, ..., K$, em que $f(\mathbf{w}_j|\boldsymbol{\phi}_k)$ é a função massa de probabilidade da $Multinomial(n_{d_j}, \boldsymbol{\phi}_k)$ calculada em $\mathbf{w}_j$.

Visto que as distribuições *a posteriori* condicionais de cada um dos parâmetros são representadas por distribuições *Dirichlet* e *discreta*, é possível, a partir de métodos de inferência, obter simulações das mesmas com a finalidade de conseguir amostras da distribuição conjunta.

2.2.4 Amostrador de Gibbs

O amostrador de Gibbs é relatado como um dos principais métodos de inferência do Modelo de Mistura, por ser considerado de fácil implementação e, em geral, obter resultados satisfatórios em relação a outros métodos, principalmente em termos computacionais. O método é considerado um algoritmo de Monte Carlo em Cadeia de Markov (MCMC) (GRIFFITHS; STEYVERS, 2004), estes que foram desenvolvidos com o intuito de simular amostras de distribuições conjuntas a partir de suas condicionais completas.

Para o Modelo de Mistura, a amostragem visada pelo Amostrador de Gibbs é feita, primeiramente, com foco na obtenção das variáveis não observadas em $\mathbf{S}$, para que, posteriormente, simulemos valores adequados de $\boldsymbol{\theta}$ e $\boldsymbol{\phi}$. Isso ocorre através da utilização das distribuições *a posteriori* condicionais utilizando os dados observáveis $\mathbf{X}$. Sendo assim, o algoritmo do Amostrador de Gibbs pode ser dado por:

1. Defina o número total de iterações, simbolizado por Q, os valores dos hiper-parâmetros em $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ e especifique o número de tópicos K;

2. Para $j = 1, ..., m$, inicialize, por exemplo, de modo aleatório as variáveis latentes S_j entre os tópicos de 1 até K, considerando que, primeiramente, $P(S_j = k) = \frac{1}{K}$, em que $k = 1, ..., K$;

3. Para a q-ésima iteração, em que $q = 1, ..., Q$, considerando $k = 1, ..., K$:

   3.1. Gere $\boldsymbol{\phi}_k^{(q)} \sim Dirichlet(\sum_{j=1}^{m} w_{j1} I(S_j = k) + \beta_{k1}, ..., \sum_{j=1}^{m} w_{jV} I(S_j = k) + \beta_{kV})$;

3.2. Gere $\boldsymbol{\theta}^{(q)} \sim Dirichlet(m_1+\alpha_1,...,m_K+\alpha_K)$;

3.3. Para $j = 1,...,m$ atualize o valor de S_j a partir de uma distribuição condicional completa *a posteriori*:

$$P(S_j^{(q)} = k|\mathbf{w}_j, \boldsymbol{\theta}^{(q)}, \boldsymbol{\phi}^{(q)}) = \frac{P(S_j^{(q)}=k|\boldsymbol{\theta}^{(q)}) f(\mathbf{w}_j|S_j^{(q)}=k, \boldsymbol{\phi}^{(q)})}{\sum_{t=1}^{K} P(S_j^{(q)}=t|\boldsymbol{\theta}^{(q)}) f(\mathbf{w}_j|S_j^{(q)}=t, \boldsymbol{\phi}^{(q)})}, k = 1,...,K$$

em que $f(\mathbf{w}_j|S_j = k, \boldsymbol{\phi}^{(q)})$ é a função massa de probabilidade da $Multinomial(n_{d_j}, \boldsymbol{\phi}_k^{(q)})$ calculada em $\mathbf{w}_j$ e $P(S_j = k|\boldsymbol{\theta}^{(q)}) = \theta_k^{(q)}$.

Com o intuito de obter estimativas para os parâmetros θ_k e ϕ_k, ao finalizar as Q iterações, inicialmente desconsideraremos as B primeiras, quando o algoritmo ainda não mostra convergência. Então, guardaremos os valores amostrados nas iterações a cada determinado *"espaçamento"*, que serão definidos a fim de garantir amostras independentes ou com baixa correlação entre si. Com isso, obteremos uma cadeia de tamanho final $Q_F = int(\frac{Q-B}{espaçamento})$, em que *int* representa a parte inteira da divisão. A partir dessa cadeia, conseguimos representar as estimativas pontuais dos parâmetros por

1.
$$\hat{\phi_{kv}} = \frac{1}{Q_F}\sum_{q=1}^{Q_F} \phi_{kv}^{(q)};$$

2.
$$\hat{\theta_k} = \frac{1}{Q_F}\sum_{q=1}^{Q_F} \theta_k^{(q)}.$$

A probabilidade *a posteriori* do documento j pertencer ao tópico k será dada pela quantidade de vezes em que ele esteve alocado a k, N_{jk}, dentro da cadeia final de tamanho Q_F, ou seja, $P(S_j = k) = \frac{N_{jk}}{Q_F}$.

Apesar de que não serão utilizadas neste trabalho, para o leitor que se interessar, existem algumas formas para a definição da convergência do Amostrador de Gibbs, como através do gráfico de traços ou de métricas específicas para esta finalidade, como a estatística criada por Brooks e Gelman (1998) ou a de Cowles e Carlin (1996).

2.3 *Latent Dirichlet Allocation* (LDA)

Nesta Seção são apresentadas informações acerca do LDA. Na Subseção 2.3.1 são abordadas sua origem e propriedades. A função de verossimilhança e a estimação Bayesiana do modelo são mostradas, respectivamente, nas Subseções 2.3.2 e 2.3.3. Na Subseção 2.3.4 discutimos sobre o Amostrador de Gibbs.

2.3.1 Conceitos básicos

O modelo LDA surge como um dos principais modelos probabilísticos do tipo generativo[2] para coleção de dados discretos com o formato de *corpus* de documentos (BLEI; NG; JORDAN, 2003). O LDA descreve como são gerados esses documentos, em que as variáveis observáveis são os termos (palavras) presentes nos mesmos, enquanto as não observáveis representam de que tópico cada uma delas é proveniente, com os parâmetros dados *a priori* no modelo (FALEIROS; LOPES *et al.*, 2016). Sendo assim, o objetivo do LDA é encontrar os tópicos aos quais cada documento pertence, através das palavras que ele contém.

Os documentos mapeiam a distribuição de probabilidade dos termos nos tópicos latentes e, a distribuição destes é descrita pela distribuição de probabilidade de Dirichlet, que pode ser vista no Apêndice A.

O LDA faz a sua modelagem a nível de cada palavra de cada documento, ou seja, cada documento pode exibir vários tópicos. Para entender melhor o funcionamento do LDA na prática, utilizaremos as mesmas frases hipotéticas do exemplo do Modelo de Mistura:

1. O Botafogo é time de futebol mais tradicional do Brasil;
2. Neymar é considerado pela FIFA o melhor jogador de futebol do mundo;
3. A pandemia do coronavírus está próxima do fim;
4. Brasil registra a menor média móvel de mortes pelo novo coronavírus desde janeiro;
5. Com a liberação de torcidas nos estádios de futebol, aumentam-se os índices de contágio da variante delta do coronavírus na Inglaterra.

Com o auxílio do LDA identificamos a qual conjunto de tópicos cada frase será designada. Então, supondo que, no exemplo acima, consideramos 2 tópicos, I, que será representado em geral por documentos em que o assunto seja futebolístico e II, que terá em geral documentos referentes ao coronavírus. Assim, as palavras das frases 1 e 2 tem maior probabilidade de serem alocadas para o tópico I, enquanto as palavras das frases 3 e 4 são mais propícias a serem alocadas para o tópico II. Por fim, o tópico I descreve 40% das palavras da frase 5, enquanto o tópico II descreve 60% das palavras.

Os documentos podem se referir a diversos contextos, como no caso da frase 5, em que ela poderia pertencer tanto ao tópico I, quanto ao II, pois há as palavras “torcidas”, “estádios” e “futebol”, que possivelmente teriam uma alta probabilidade de pertencerem ao tópico I e, também, as palavras “índices”, “contágio”, “variante”, “delta” e “coronavírus”, que devem estar entre as com maior probabilidade de pertencerem ao tópico II. Note que para este conjunto de documentos, em teoria, o LDA permitiria 1 tópico a menos que o Modelo de Mistura.

[2] Um modelo é dito generativo quando aleatoriamente gera dados a partir de variáveis latentes (não observáveis).

2.3.2 Função de verossimilhança

Seja $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_m)$, com $\mathbf{Z}_j = (Z_{j1}, ..., Z_{jn_{dj}})$, um conjunto de variáveis aleatórias discretas e independentes, tal que $Z_{ji} \sim Discreta(\boldsymbol{\theta}_j)$, em que $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jK})$, K representa o número de diferentes tópicos, m o número de diferentes documentos e n_{d_j} o número de palavras dentro do documento j. Dessa forma, Z_{ji} representa o tópico associado a i-ésima palavra do j-ésimo documento, simbolizada por Y_{ji}, com $j = 1, \ldots, m$, $i = 1, \ldots, n_{d_j}$ e $Z_{ji} \in \{1, \ldots, K\}$. Sendo assim, definimos:

$$P(Z_{ji} = k) = \theta_{jk}, \tag{2.2}$$

em que $0 \leq \theta_{jk} \leq 1$ e $\sum_{k=1}^{K} \theta_{jk} = 1$.

Seja V o número de palavras do dicionário, dispostas conforme a estrutura *bag-of-words*, temos que $\mathbf{w}_j = (w_{j1}, \ldots, w_{jV})$, em que $\sum_{v=1}^{V} w_{jv} = n_{d_j}$. Dessa forma, $\mathbf{w}_j$ é o vetor que representa a quantidade de vezes que cada uma das V palavras do vocabulário aconteceram no documento j e n_{d_j} é a quantidade de palavras presentes no documento j. A estrutura *bag-of-words* para representar os documentos, apesar de não ser diretamente utilizada na definição do modelo LDA que analisa palavra a palavra de cada documento e não as palavras do vocabulário, é uma maneira eficiente de estruturar as informações e conhecer quais e quantas vezes as palavras acontecem em cada documento para uma análise posterior individual (palavra a palavra).

Sendo assim, originamos a matriz de documentos $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_m)^T$, em que $\mathbf{Y}_m = (Y_{m1}, \ldots, Y_{mn_{dm}})$ é o vetor que representa as n_j palavras presentes no documento m. Neste contexto, temos que $Y_{ji}|Z_{ji} = k \sim Discreta(\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kV}))$. Com isso, a função de verossimilhança aumentada do modelo pode ser vista como

$$\begin{aligned}
L(\boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{Z}, \mathbf{Y}) &= f_{\mathbf{Z},\mathbf{Y}}(\mathbf{z}, \mathbf{y}) \\
&= f_{\mathbf{Z}}(\mathbf{z}) f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}, \mathbf{z}) \\
&= \Big(\prod_{j=1}^{m}\prod_{i=1}^{n_{d_j}} P(Z_{ji})\Big)\Big(\prod_{j=1}^{m}\prod_{i=1}^{n_{d_j}} f(y_{ji}|Z_{ji})\Big) \\
&= \prod_{j=1}^{m}\prod_{i=1}^{n_{d_j}}\prod_{k=1}^{K}\left[\theta_{jk}^{I(Z_{ji}=k)}\prod_{v=1}^{V}\left(\phi_{kv}^{I(Y_{ji}=v)I(Z_{ji}=k)}\right)\right] \\
&= \left(\prod_{k=1}^{K}\prod_{j=1}^{m}\theta_{jk}^{\sum_{i=1}^{n_{d_j}} I(Z_{ji}=k)}\right)\left(\prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{kv}^{\sum_{j=1}^{m}\sum_{i=1}^{n_{d_j}} I(Y_{ji}=v)I(Z_{ji}=k)}\right) \\
&\propto \left(\prod_{k=1}^{K}\prod_{j=1}^{m}\theta_{jk}^{n_{jk}}\right)\left(\prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{kv}^{m_{kv}}\right),
\end{aligned}$$

em que n_{jk} é o número palavras do j-ésimo documento associadas ao tópico k e m_{kv} é o numero de vezes que a palavra v do vocabulário acontece no tópico k.

2.3.3 Estimação Bayesiana

Esse modelo geralmente é estimado através da abordagem Bayesiana pelos mesmos motivos descritos na Subseção 2.2.3. Dessa maneira, combinando a verossimilhança com as seguintes distribuições *a priori*

1. $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jK}) \sim Dirichlet(\boldsymbol{\alpha}_j = (\alpha_{j1}, \cdots, \alpha_{jK}))$, em que $0 \leq \theta_{jk} \leq 1$ e $\sum_{k=1}^{K} \theta_{jk} = 1$;
2. $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kV}) \sim Dirichlet(\boldsymbol{\beta}_k = (\beta_{k1}, ..., \beta_{kV}))$, em que $0 \leq \phi_{kv} \leq 1$ e $\sum_{v=1}^{V} \phi_{kv} = 1$,

e $\alpha_{jk} > 0$ e $\beta_{kv} > 0$ são hiper-parâmetros conhecidos, podemos escrever a distribuição *a posteriori* conjunta do modelo como

$$\begin{aligned}
\pi(\boldsymbol{\theta}, \boldsymbol{\phi} | \mathbf{Z}, \mathbf{Y}) &\propto L(\boldsymbol{\theta}, \boldsymbol{\phi} | \mathbf{Z}, \mathbf{Y}) \prod_{j=1}^{m} \pi(\boldsymbol{\theta}_j) \prod_{k=1}^{K} \pi(\boldsymbol{\phi}_k) \\
&\propto \left(\prod_{k=1}^{K} \prod_{j=1}^{m} \theta_{jk}^{n_{jk}} \right) \left(\prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{kv}^{m_{kv}} \right) \left(\prod_{k=1}^{K} \prod_{j=1}^{m} \theta_{jk}^{\alpha_{jk}-1} \right) \left(\prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{kv}^{\beta_{kv}-1} \right) \\
&= \prod_{k=1}^{K} \left[\left(\prod_{j=1}^{m} \theta_{jk}^{n_{jk}} \right) \left(\prod_{v=1}^{V} \phi_{kv}^{m_{kv}} \right) \left(\prod_{j=1}^{m} \theta_{jk}^{\alpha_{jk}-1} \right) \left(\prod_{v=1}^{V} \phi_{kv}^{\beta_{kv}-1} \right) \right] \\
&= \prod_{k=1}^{K} \left[\prod_{j=1}^{m} \left(\theta_{jk}^{n_{jk}} \theta_{jk}^{\alpha_{jk}-1} \right) \prod_{v=1}^{V} \left(\phi_{kv}^{m_{kv}} \phi_{kv}^{\beta_{kv}-1} \right) \right] \\
&= \prod_{k=1}^{K} \left[\left(\prod_{j=1}^{m} \theta_{jk}^{n_{jk}+\alpha_{jk}-1} \right) \left(\prod_{v=1}^{V} \phi_{kv}^{m_{kv}+\beta_{kv}-1} \right) \right].
\end{aligned}$$

Desta forma, as distribuições *a posteriori* condicionais podem ser descritas como

1.
$$\begin{aligned}
\pi(\boldsymbol{\theta}_j | \cdots) &\propto \prod_{k=1}^{K} \theta_{jk}^{n_{jk}+\alpha_{jk}-1} \\
&\equiv Dirichlet(n_{j1} + \alpha_{j1}, ..., n_{jK} + \alpha_{jK}),
\end{aligned}$$

em que $\cdots$ inclui todas as variáveis e parâmetros exceto o(s) parâmetro(s) em questão;

2.
$$\begin{aligned}
\pi(\boldsymbol{\phi}_k | \cdots) &\propto \prod_{v=1}^{V} \phi_{kv}^{m_{kv}+\beta_{kv}-1} \\
&\equiv Dirichlet(m_{k1} + \beta_{k1}, ..., m_{kV} + \beta_{kV}).
\end{aligned}$$

2.3.4 Amostrador de Gibbs

Assim como para o Modelo de Mistura e pelos mesmos motivos já descritos no primeiro parágrafo da Subseção 2.2.4, o Amostrador de Gibbs também é um dos principais métodos de inferência do modelo LDA.

Para uma específica configuração das variáveis Zs, as probabilidades da v-ésima palavra segundo o tópico k (ϕ_{kv}) e de sortear uma palavra segundo o tópico k para o documento j (θ_{jk}) podem ser estimadas como a sua média *a posteriori*, dadas por

1.
$$\hat{\phi}_{kv} = \frac{m_{kv} + \beta_{kv}}{n_k + \sum_{v=1}^{V} \beta_{kv}};$$

2.
$$\hat{\theta}_{jk} = \frac{n_{jk} + \alpha_{jk}}{n_{d_j} \sum_{k=1}^{K} \alpha_{jk}}.$$

As variáveis não observadas Zs, por sua vez, podem ser preditas através da sua probabilidade *a posteriori* condicional, dada por

$$\begin{aligned} P(Z_{ji} = k | Y_{ji} = v, \mathbf{Y}_{-ji}, \mathbf{Z}_{-ji}) &\propto Pr(Y_{ji} = v | Z_{ji} = k, \mathbf{Y}_{-ji}, \mathbf{Z}_{-ji}) Pr(Z_{ji} = k | \mathbf{Y}_{-ji}, \mathbf{Z}_{-ji}) \\ &= \phi_{kv}^{-ji} \theta_{jk}^{-ji}, \end{aligned}$$

em que $-ji$ representa que a i-ésima palavra do j-ésimo documento foi excluída. Essa probabilidade pode ser estimada como

$$\hat{P}(Z_{ji} = k | Y_{ji} = v, \mathbf{Y}_{-ji}, \mathbf{Z}_{-ji}) \propto \frac{m_{kv}^{-ji} + \beta_{kv}}{n_k^{-ji} + \sum_{v=1}^{V} \beta_{kv}} \frac{n_{jk}^{-ji} + \alpha_{jk}}{n_{d_j}^{-ji} \sum_{k=1}^{K} \alpha_{jk}},$$

em que n_k é o número de palavras atribuídas ao tópico k. Além disso, a primeira expressão nos mostra a probabilidade da palavra i no tópico k, enquanto a segunda corresponde a propensão do tópico k para o documento j.

Para o leitor que se interessar, mais detalhes podem ser consultados em Griffiths e Steyvers (2004) e Phan, Nguyen e Horiguchi (2008).

No *software* R, a estimação do modelo LDA via amostrador de Gibbs pode ser realizada via pacote *topicmodels* (HORNIK; GRÜN, 2011). É importante ressaltar que existem outros métodos de inferência que podem ser usados para estimar os parâmetros do modelo LDA, como o método variacional, cujos detalhes, para o leitor que se interessar, podem ser consultados em Teh, Newman e Welling (2006).

2.4 Diferença entre MM e LDA

A principal diferença entre as duas metodologias abordadas anteriormente se dá em torno das variáveis Z e S, que no Modelo de Mistura ela é a nível de cada texto (documento), enquanto no LDA ela é para cada palavra de cada documento.

Essa diferença implica que no Modelo de Mistura cada documento sempre será composto de palavras advindas de um único tópico, em contrapartida, no LDA há uma maior flexibilidade, pois permite que cada documento possa exibir vários tópicos, visto que a variável não observada, Z_{ji} varia conforme cada uma das palavras que ele contém.

Em termos de dimensão paramétrica e número de variáveis não observadas, o MM possui $K+VK$ parâmetros a serem estimados e m variáveis não observadas a serem preditas. O LDA, por sua vez, possui $mK+VK$ parâmetros e $\sum_{j=1}^{m} n_{d_j}$ variáveis não observadas. Dessa maneira, o número de informações desconhecidas no LDA é maior do que no MM para um mesmo K, mas, em contrapartida, acreditamos que o número de tópicos K no LDA seja menor do que no MM.

2.5 Métricas de mensuração da qualidade dos modelos

Uma das críticas em relação ao MM e LDA se dá em torno da obrigatoriedade, desde o início, da definição do número de tópicos pelo usuário. Desta forma, algumas adversidades podem ser ocasionadas, como a escolha errada do número de tópicos pela falta de conhecimento do usuário sobre o seu conjunto de dados, acarretando assim, possivelmente, em resultados pouco efetivos do método.

Neste contexto, é de extrema importância a escolha correta do número de tópicos que será modelado. Uma forma é utilizar métricas que mensuram a qualidade de modelos com diferentes valores K. Nesse sentido, apresentaremos a seguir algumas medidas que buscam verificar a capacidade do modelo em caracterizar os textos em determinadas quantidades de tópicos.

Uma das maneiras de avaliar a eficiência do modelo é através de sua própria verossimilhança (WALLACH *et al.*, 2009). Quanto maior o valor do logaritmo da verossimilhança, melhor é o modelo. A expressão da função de log-verossimilhança a ser calculada é a mesma escrita na definição dos modelos, anteriormente.

Outra opção é a medida de perplexidade, frequentemente utilizada para a avaliação de modelos textuais (NEWMAN *et al.*, 2007). A mesma é inspirada na entropia, sendo sua exponencial. Quanto menor o valor da perplexidade, melhor é o modelo. Adaptada para o MM e LDA, a perplexidade é dada por

$$perplexidade = \exp\left(-\frac{\log L(\boldsymbol{\theta},\boldsymbol{\phi}|\mathbf{S},\mathbf{X})}{\sum_{j=1}^{m} n_{d_j}}\right).$$

Através da log-verossimilhança e da perplexidade, o esperado é que um modelo com um maior número de tópicos seja sempre melhor que um modelo com menor número de tópicos, levando ao super-ajuste dos dados. No entanto, assim como em outras análises de agrupamento, podemos escolher como ideal o número de tópicos que estabiliza o valor dessas métricas, ou seja, o número de tópicos a partir do qual o modelo não é mais tão melhorado.

No entanto, apesar de comumente utilizada, Chang *et al.* (2009) mostrou através de um experimento que, geralmente, a perplexidade e o julgamento humano não estão correlacionados, pois a métrica é útil para avaliar o modelo preditivo, mas não aborda os objetivos mais explicativos da modelagem de tópicos, que são os conteúdos dos textos. Sendo assim, o estudo do autor levantou indícios de que a perplexidade pode nos levar a tópicos com baixa importância semântica, ou seja, pouco agregadores.

Diante disso, outras alternativas são as métricas que visam avaliar a interpretabilidade dos tópicos e verificar a coerência dos mesmos. Mei, Shen e Zhai (2007) abordaram o *Pointwise Mutual Information* (PMI), que visa, justamente, avaliar a interpretabilidade e a associação das palavras nos tópicos. Para isto, este método automático faz a verificação através de pares de palavras que formam o tópico e analisa se a probabilidade conjunta de ocorrência do par é igual ao produto das probabilidades marginais de cada palavra. Quanto maior o seu valor, melhor o modelo.

Então, ordenando as palavras da mais provável até a menos dentro de cada tópico, temos o vetor $\mathbf{w}_k = (w_{k1}, \cdots, w_{kL})$, em que w_{kL} é a L-ésima palavra mais provável do tópico k. Logo, L delimita a quantidade de palavras verificadas por tópico. De acordo com Syed e Spruit (2017), as 10 palavras com maior probabilidade de pertencerem a um determinado tópico, geralmente são suficientes para interpretá-lo. Quanto maior o valor da métrica, melhor tende a ser o modelo. Com isso, a métrica adaptada (pois, na literatura não é apresentada como uma média tal como é proposto aqui e não considera apenas as L palavras mais prováveis de cada tópico) pode ser expressa por

$$PMI^* = \frac{\sum_{k=1}^{K}\sum_{r=1}^{L-1}\sum_{s=2}^{L}\log\frac{p(w_{kr},w_{ks})}{p(w_{kr})p(w_{ks})}}{K\frac{L!}{2!(L-2)!}}, \forall r < s,$$

sendo que $p(w_{kr})$ é a probabilidade de ocorrer a palavra w_{kr} dentro de um documento. Isto é,

$$p(w_{kr}) = \frac{\sum_{j=1}^{m} I(w_{kr}|j)}{m},$$

em que a indicadora $I(w_{kr}|j)$ assume o valor 1 se a palavra w_{kr} está presente no documento j e o valor 0, se ela não está. De maneira análoga, $p(w_{kr}, w_{ks})$ é a probabilidade de ocorrerem ambas as palavras, w_{kr} e w_{ks}, dentro de um mesmo documento. Ou seja,

$$p(w_{kr}, w_{ks}) = \frac{\sum_{j=1}^{m} I(w_{kr}, w_{ks}|j)}{m},$$

em que a indicadora $I(w_{kr}, w_{ks}|j)$ assume o valor 1 se as palavras w_{kr} e w_{ks} estão presentes no documento j e o valor 0, se elas não estão.

Para o PMI e as demais métricas que serão expostas posteriormente, no denominador há uma multiplicação de K com uma combinação de L tomados 2 a 2. O K é utilizado para que seja feita uma média, dado que para cada tópico há uma determinada distribuição de palavras mais prováveis, enquanto a combinação de L tomados 2 a 2 é realizada por conta de que não faz diferença a ordem da dupla de palavras para o cálculo das probabilidades, ou seja, $p(w_{kr}, w_{ks}) = p(w_{ks}, w_{kr})$, por exemplo.

Além disso, ainda para o PMI e as demais métricas que serão expostas posteriormente, quanto maiores os seus valores, melhores tendem as serem os modelos. Isso é dado já que quanto maiores as probabilidades das palavras, principalmente conjuntamente, maiores são as suas associações e, consequentemente, melhores são os tópicos, visto que isso significa que eles agruparam palavras que possuem alto grau de sintonia.

Uma alternativa e variação do PMI, mudando a sua escala, é o *Normalized Pointwise Mutual Information* (NPMI), proposto por Bouma (2009), que normaliza o valor obtido para o intervalo [-1, 1] e, quanto mais próximo de 1, melhor o modelo. A equação da métrica adaptada é dada por

$$NPMI^* = \frac{\sum_{k=1}^{K}\sum_{r=1}^{L-1}\sum_{s=2}^{L}\frac{\log\frac{p(w_{kr},w_{ks})}{p(w_{kr})p(w_{ks})}}{-\log p(w_{kr},w_{ks})}}{K\frac{L!}{2!(L-2)!}}, \forall r < s.$$

Uma medida semelhante ao PMI foi proposta por Mimno *et al.* (2011), a *Log Conditional Probability* (LCP), que quanto maior o seu valor melhor tende a ser o modelo. Sua fórmula adaptada, pode ser representada por

$$LCP^* = \frac{\sum_{k=1}^{K}\sum_{r=1}^{L-1}\sum_{s=2}^{L}\log\frac{p(w_{kr},w_{ks})}{p(w_{ks})}}{K\frac{L!}{2!(L-2)!}}, \forall r < s.$$

Vale ressaltar que apesar de que neste trabalho iremos focar nas métricas descritas nesta subseção, existem mais medidas presentes na literatura.

O LDA foi processado utilizando a biblioteca *topicmodels* no R, enquanto o algoritmo de estimação do MM foi implementado no mesmo *software*. Os principais códigos estão disponíveis no Apêndice B.

CAPÍTULO

3

RESULTADOS COMPARATIVOS DAS APLICAÇÕES

Neste capítulo serão apresentadas duas aplicações das metodologias em bancos de dados reais. Na Seção 3.1 a abordagem será feita em um banco de dados do ramo esportivo, enquanto na Seção 3.2 os documentos englobam artigos de notícias de uma agência americana.

3.1 Banco de dados *BBC Sport*

O banco de dados *"BBC Sport"* está disponível em Kaggle (2020). O conjunto de dados, que dispõe de 737 documentos, com 4613 palavras (termos) no total, é composto por notícias esportivas separadas e classificadas previamente por especialistas em 5 esportes: atletismo, críquete, futebol, rúgbi e tênis, publicadas entre 2004 e 2005. Os documentos possuem, em média, 166 termos, com o mínimo de 50 e o máximo de 888 palavras, ou seja, se tratam de textos de médio comprimento.

Este conjunto de documentos é atrativo para análise pelo fato de que como os documentos já estão classificados por esportes, temos o número "ideal" de tópicos pré-estabelecido, sendo interessante, além de verificar o comportamento das diferentes métricas, fazer o comparativo de qual método, MM ou LDA, consegue alocar melhor os documentos em seus respectivos temas originais.

Apesar do trabalho em questão ser escrito em português e os dados esterem dispostos na língua inglesa, preferimos deixá-lo em seu idioma original, a fim de evitar erros de tradução que possam impactar nos resultados finais.

Inicialmente, os dados foram dispostos no formato de *bag-of-words* e feito o seu pré-processamento. Depois, para o prosseguimento das análises, variamos a quantidade de tópicos estabelecida previamente em $K = 1, 2, 3, ..., 18, 19, 20, 25, 30, 35, 40, 45, 50, 75, 100, 125,$

150, 175, 200, 250, 300, 400 e 500 para que fossem feitos os ajustes dos modelos de LDA e MM e, posteriormente, o cálculo das métricas que auxiliam na escolha do melhor número de tópicos. Foram considerados $\boldsymbol{\alpha} = (1, ..., 1)$ e $\boldsymbol{\beta} = (1, ..., 1)$ como valores para os hiper-parâmetros. Utilizamos para a inferência dos modelos o amostrador de Gibbs, considerando 1000 iterações para cada valor de K no LDA e 500 no MM. Apesar de ser, a princípio, um número pequeno de iterações MCMC, como se trata da simulação de distribuições bem conhecidas e simples em cada passo, a convergência dos algoritmos foi observada e os valores simulados em uma iteração pouco diferenciavam dos valores simulados na iteração seguinte, a partir da vigésima iteração.

3.1.1 Resultados com número de tópicos pré-estabelecido (K=5)

Como dito anteriormente, no banco de dados *BBC Sport* há uma alocação prévia dos documentos em assuntos, separando-os por esportes. Os 101 primeiros textos se referem ao atletismo, os 124 posteriores se enquadram no críquete, em seguida, os próximos 265 se dizem respeito ao futebol, enquanto os 147 subsequentes correspondem ao rugby e, por fim, os últimos 100 se encaixam no tênis.

Devido a isso, antes de verificarmos o melhor número de tópicos através das métricas analisadas, analisaremos o comportamento das metodologias, MM e LDA, considerando a separação pré-estabelecida pelo autor do banco de dados, isto é, compararemos os métodos considerando o número de tópicos igual a 5.

Primeiramente, avaliaremos as métricas que mensuram a qualidade dos modelos.

Tabela 2 – *BBC Sport* - Métricas de desempenho para os modelos com $K = 5$.

	Log Verossimilhança	**Perplexidade**	**PMI**	**NPMI**	**LCP**
MM	-878487.2	1287.1	0.295	0.120	-1.138
LDA	-813616.9	758.6	0.328	0.143	-1.027

Na Tabela 2 trouxemos, para o MM e LDA, os respectivos valores de cada uma das métricas analisadas: log verossimilhança, perplexidade, PMI, NPMI e LCP. Lembrando que, com exceção da perplexidade, para a interpretação das demais medidas, quanto maior o seu valor, melhor tende a ser o modelo. Diante disso, vemos que o LDA apresenta, em todas as cinco métricas, valores melhores em comparação ao MM.

Agora, analisaremos, separando o conjunto de textos pelos cinco esportes definidos previamente, para qual tópico cada documento foi designado, com o intuito de verificar se os tópicos gerados pelo MM e LDA conseguiram separar os textos de maneira semelhante a alocação prévia feita pelo autor. Além disso, avaliaremos também as 10 palavras mais prováveis por tópicos em cada uma das metodologias, palavras estas que foram as utilizadas para calcular as métricas PMI, NPMI e LCP.

Tabela 3 – *BBC Sport* - Percentual dos tópicos por esporte via MM.

	Atletismo	**Críquete**	**Futebol**	**Rugby**	**Tênis**
Tópico 1	-	0.016	0.868	-	-
Tópico 2	0.990	0.218	0.034	0.184	0.090
Tópico 3	-	-	0.079	0.340	-
Tópico 4	0.010	-	0.019	0.476	0.910
Tópico 5	-	0.766	-	0.019	-
Total	1	1	1	1	1

A Tabela 3 traz, para o MM, a porcentagem de documentos alocados em cada um dos cinco tópicos dentro de cada classe “verdadeira”. Dessa maneira, vemos que os documentos dos esportes atletismo, futebol e tênis foram designados, majoritariamente, mais de 85%, para os tópicos 2, 1 e 4, respectivamente. Isto é, por exemplo, dos 101 textos de atletismo, 100 foram alocados ao tópico 2, representando 99%. Em relação ao críquete, 76,6% dos textos foram atribuídos ao tópico 5 e 21,8% ao tópico 2. Por fim, os tópicos definidos pelo MM não distinguiram claramente os textos associados ao rugby, tendo em vista que três tópicos demonstraram percentuais significativos dos documentos do esporte, sendo eles o tópico 4 (47,6%), 3 (34,0%) e 2 (18,4%). Pela metodologia, alguns textos do rugby se confundiram com textos do tênis e do atletismo principalmente.

Tabela 4 – *BBC Sport* - 10 palavras mais prováveis por tópico via MM.

Tópico 1	**Tópico 2**	**Tópico 3**	**Tópico 4**	**Tópico 5**
game	world	ireland	play	test
player	olymp	minut	game	play
play	win	england	win	england
club	year	side	england	pakistan
chelsea	test	ball	first	seri
arsen	athlet	win	against	first
league	time	game	open	ball
win	race	try	match	run
footbal	indoor	penalti	world	team
unit	team	itali	two	two

Na Tabela 4 podemos visualizar as 10 palavras mais prováveis por tópico no MM. Conforme visto anteriormente, o tópico 1 é o que alocou a maioria dos documentos de futebol, o que é explicado ao vermos que entre as suas palavras mais prováveis estão "player", "chelsea", "arsen"e "footbal", termos estes comumente utilizados no ramo futebolístico. Além disso, o tópico 2, que englobou praticamente todos os documentos de atletismo, apresenta dentre as 10 palavras mais propícias, "athlet"e "race". Em contrapartida, os documentos de tênis, designados majoritariamente no tópico 4, retorna a palavra "open"como uma das top 10, que deve fazer referência a um dos principais torneios do esporte, o US Open. O tópico 5 retorna "test"como sua principal palavra e é o mais frequente no críquete, o que deve acontecer por conta de no esporte haver um formato de partida chamada de teste, que acontece durante 5 dias. Para finalizar, o

tópico 3, apesar de não ser o mais frequente no rugby, apresenta a palavra "ireland"como a mais propícia, o que pode ser justificado por conta da Irlanda ser o país líder no ranking mundial de rugby.

Tabela 5 – *BBC Sport* - Percentual dos tópicos por esporte via LDA.

	Atletismo	**Críquete**	**Futebol**	**Rugby**	**Tênis**
Tópico 1	0.238	0.008	0.008	0.857	0.010
Tópico 2	-	0.871	-	-	-
Tópico 3	-	-	0.430	0.007	-
Tópico 4	0.703	-	0.034	0.014	0.900
Tópico 5	0.059	0.121	0.528	0.122	0.090
Total	1	1	1	1	1

A Tabela 5 mostra, para o LDA, a porcentagem de documentos alocados em cada um dos cinco tópicos dentro de cada classe “verdadeira”. Como no LDA, as palavras de um mesmo documento pode ser alocadas em diferentes tópicos, escolhemos o tópico com maior probabilidade dentro de cada documento para fazer essa análise. Dessa maneira, vemos que os documentos dos esportes críquete, rugby e tênis foram designados, em sua grande maioria, mais de 85%, para os tópicos 2, 1 e 4, respectivamente. Isto é, por exemplo, dos 100 textos de tênis, 90 foram alocados ao tópico 4, representando 90%. Em relação ao atletismo, 70,3% dos textos foram atribuídos ao tópico 4 e 23,8% ao tópico 1. Por fim, dos cinco tópicos definidos pelo LDA, dois deles contemplam textos de futebol, considerando que os tópicos 3 e 5 apresentaram percentuais significativos dos documentos do esporte, 43,0% e 52,8%, respectivamente. Observe que o tópico 4 apresenta grande parte dos textos sobre atletismo e tênis, evidenciando que esses esportes foram confundidos quando consideramos $K = 5$.

Tabela 6 – *BBC Sport* - 10 palavras mais prováveis por tópico via LDA.

Tópico 1	**Tópico 2**	**Tópico 3**	**Tópico 4**	**Tópico 5**
england	test	game	win	player
against	england	minut	world	club
game	first	goal	year	play
ireland	four	chelsea	set	think
wale	match	unit	champion	want
nation	cricket	arsen	final	team
six	play	back	first	go
injuri	tour	befor	second	footbal
coach	run	chanc	open	time
rugbi	fastest	score	play	told

Na Tabela 6 podemos visualizar as 10 palavras mais prováveis por tópico no LDA. Conforme visto anteriormente, os tópicos 2 e 4, que alocaram a maioria dos documentos de críquete e tênis, respectivamente, apresentam, assim como no MM, palavras como "test"e "open"dentre as 10 mais prováveis. Além disso, no tópico 2, a palavra "cricket"também está presente. O tópico 1, onde estão designados a maior parte dos textos de rugby, além da palavra

"ireland", como no MM, retorna também dentre as 10 mais propícias, o termo "rugbi". Por fim, de fato e de acordo com o que levantamos ao analisar a Tabela 5, os tópicos 3 e 5 aparentam mesmo se referir ao futebol, visto que trazem as palavras "goal", "chelsea"e "arsen"dentre as 10 no tópico 3, enquanto no 5, são retornadas palavras como "player"e "footbal".

Apesar de ambas as metodologias apresentarem resultados satisfatórios quando consideramos 5 tópicos, e sendo o MM o modelo que melhor separou os textos dos 5 esportes, pois não confundiu muito os textos do atletismo e tênis como o LDA, faz-se necessária a análise para diferentes números de tópicos através das métricas que auxiliam na escolha do seu número ideal.

3.1.2 Comparativo das métricas e métodos

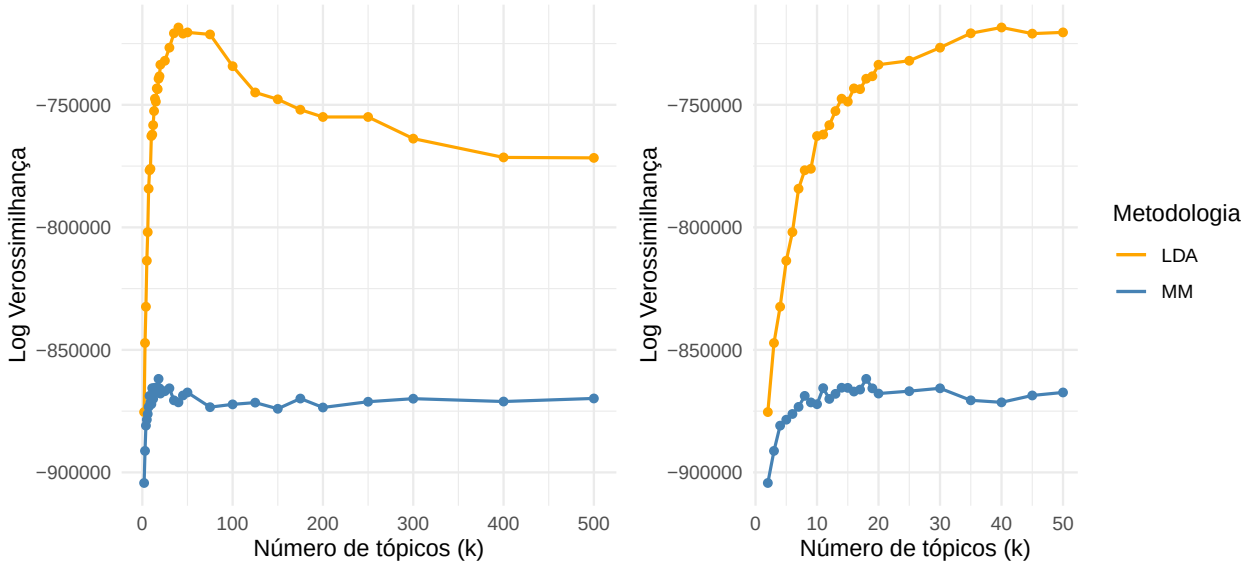

Figura 1 – *BBC Sport* - Log Verossimilhança por número de tópicos.

A Figura 1 apresenta a log verossimilhança dos modelos ajustados considerando os diferentes números de tópicos e as duas metodologias. Na esquerda estão todos os valores de K estudados, enquanto na direita estão os valores das métricas com mais detalhes para K de 1 a 50. Lembrando que quanto maior o valor da métrica, mais adequado tende a ser o modelo, vemos que o LDA apresenta valores maiores para todos os números de tópicos quando comparado com o MM. No entanto, o pico do LDA, se dá no $K = 40$, enquanto o ápice do MM é o $K = 18$. Para o MM, apesar do pico ser em $K = 18$, observamos que desde o valor de $K = 8$, a log verossimilhança já está razoavelmente estabilizada, fato que não acontece com o LDA.

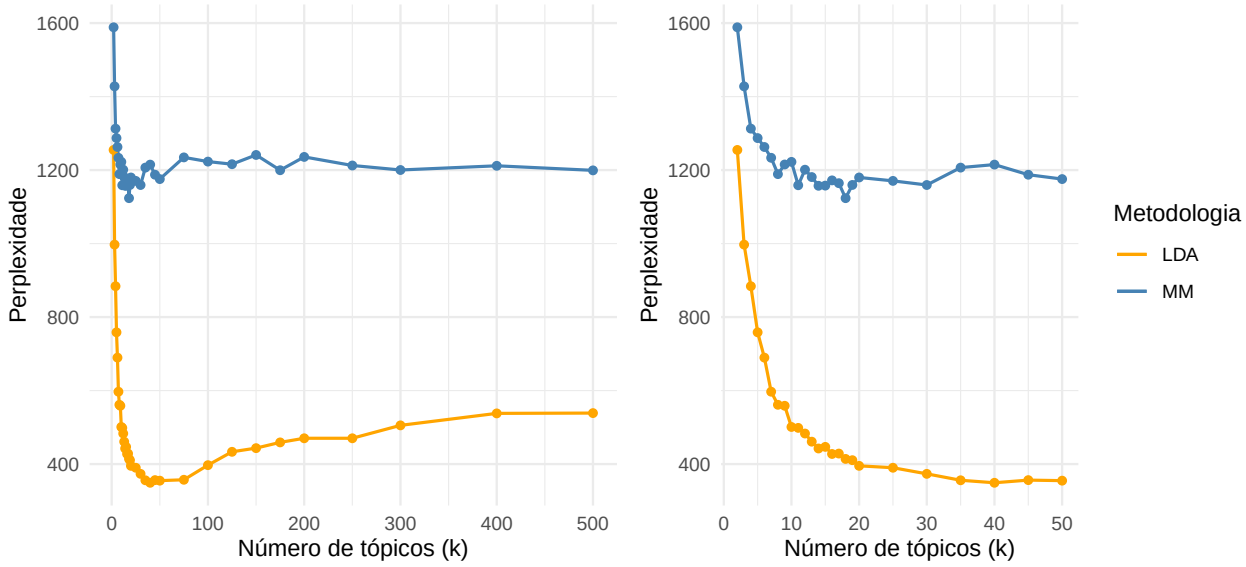

Figura 2 – *BBC Sport* - Perplexidade por número de tópicos.

A Figura 2 ilustra a perplexidade dos modelos ajustados considerando os diferentes números de tópicos e as duas metodologias. Novamente, na esquerda estão todos os valores de K estudados, enquanto na direita estão os valores apenas para K de 1 a 50. Como a perplexidade é uma transformação 1 para 1 da log verossimilhança, apesar da escala e de ser contrária em relação a interpretação, já que quanto menor o valor da métrica, mais adequado tende a ser o modelo, as conclusões se mantém de maneira análoga ao que foi dito na métrica anterior.

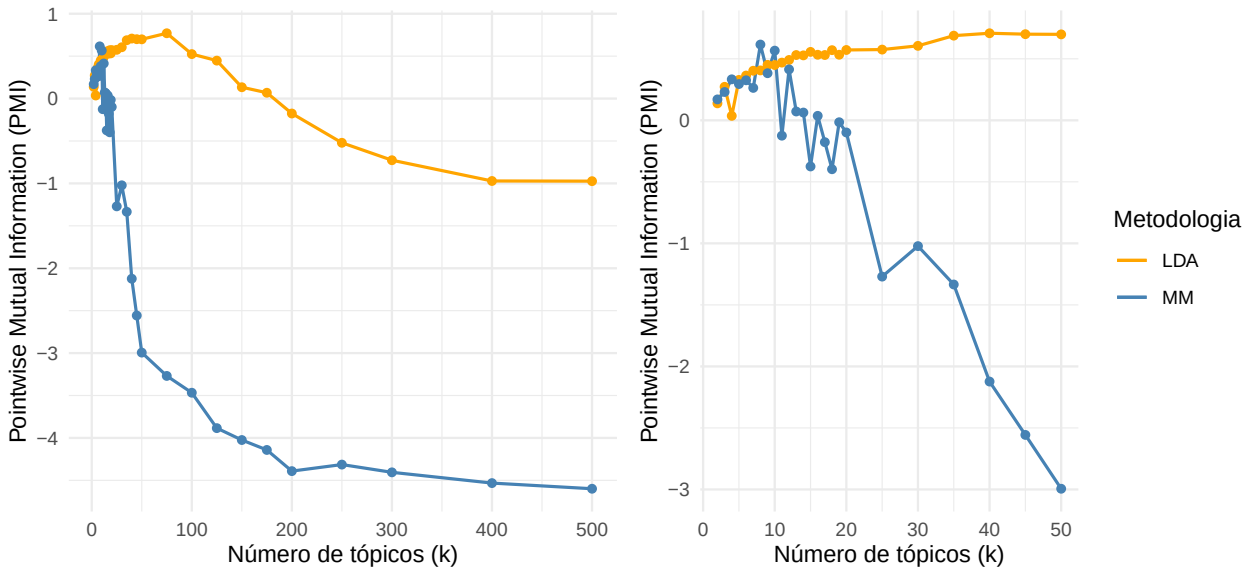

Figura 3 – *BBC Sport* - PMI por número de tópicos.

A Figura 3 traz o PMI dos modelos ajustados considerando os diferentes números de tópicos e as duas metodologias. Lembrando que quanto maior o valor da métrica, mais adequado tende a ser o modelo, em geral o LDA apresenta valores maiores da medida quando comparado

com o tópico equivalente no MM. Uma das exceções se dá no maior valor do PMI no MM, que é com $K = 8$. Por outro lado, o melhor valor para o LDA, é visto apenas no $K = 75$.

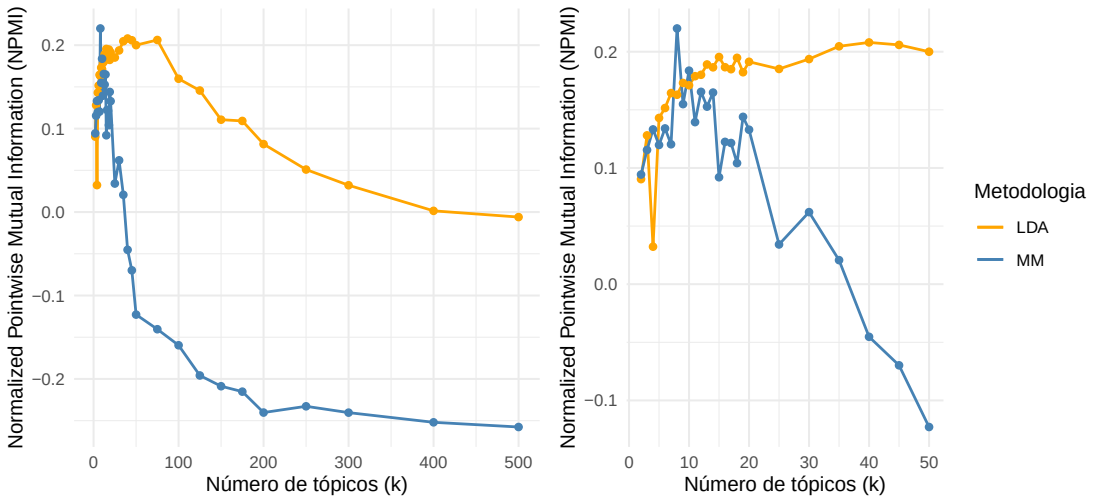

Figura 4 – *BBC Sport* - NPMI por número de tópicos.

A Figura 4 mostra o NPMI dos modelos ajustados. Lembrando que quanto mais próximo de 1 o valor da métrica, maior tende a ser a associação entre as palavras dos tópicos e mais adequado o modelo ajustado. O maior valor da métrica é visto no MM, com $K = 8$, igual ao PMI. Novamente, o LDA apresenta o melhor valor no $K = 40$.

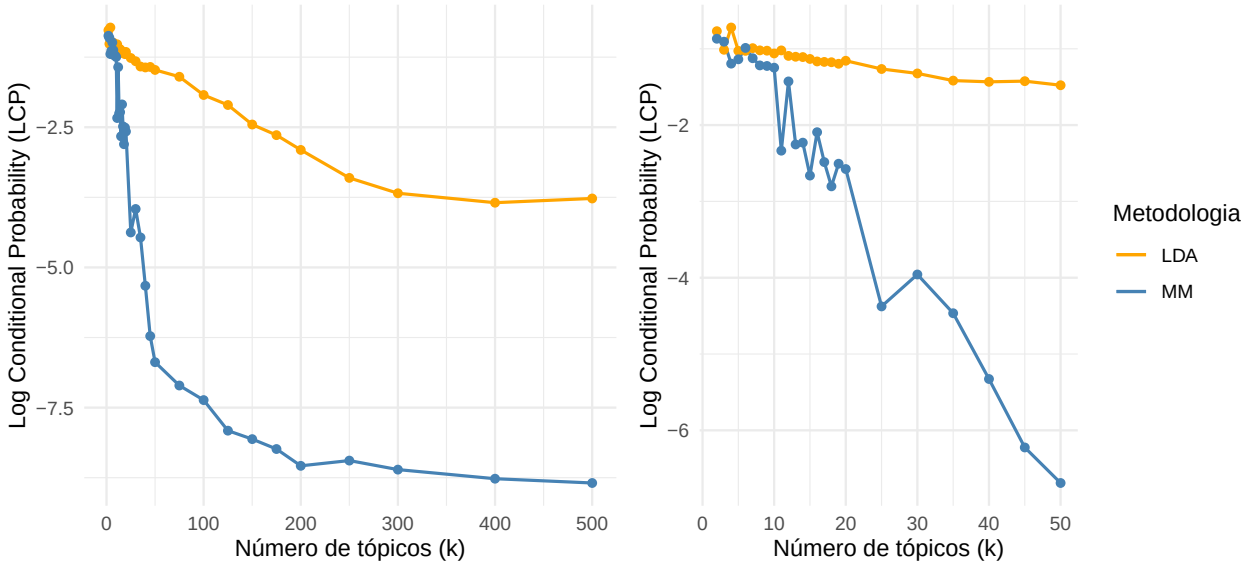

Figura 5 – *BBC Sport* - LCP por número de tópicos.

A Figura 5 exibe o LCP dos modelos ajustados. Lembrando que quanto maior o valor da métrica, mais adequado tende a ser o modelo, em geral o LDA apresenta valores maiores da medida quando comparado com o tópico equivalente no MM. O melhor valor da medida para o MM se dá no $K = 2$, enquanto para o LDA é retornado no $K = 4$.

Sob o conjunto de dados analisados, a métrica LCP se mostrou a mais restritiva escolhendo um número de tópicos menor que as outras métricas tanto para o MM quanto o LDA. No entanto, o número de tópicos escolhido foi menor que o número de grupos identificados por especialistas. Como esperado, as métricas PMI e NPMI apresentaram comportamento similar e o mesmo comentário vale para a log verossimilhança e a perplexidade. Para o MM, essas quatro métricas escolheram número de tópicos parecidos (ou iguais), mas para o LDA, a log verossimilhança e a perplexidade foram mais parsimoniosas.

Com o intuito de verificarmos se nos modelos sugeridos pelas métricas, principalmente para os com valores de K maiores, há uma concentração em determinados tópicos dos documentos alocados, observamos que isso não se estabelece, visto que a distribuição dos textos está pulverizada em vários tópicos.

3.1.3 Considerações

Ao compararmos os modelos considerando 5 tópicos, a partir da pré definição do autor de que existem cinco esportes diferentes, verificamos que as métricas utilizadas para a seleção de modelos mostram, em geral, que o LDA parece ser mais adequado ao conjunto de dados em questão do que o MM. Porém, ao analisarmos as palavras mais frequentes e a alocação dos documentos nos tópicos pelos cinco esportes, vemos que os dois métodos se assemelham, visto que ambos conseguem identificar bem os documentos de 4 esportes, mas apresentam um esporte que parece ser contemplado por mais de um tópico (rugby para o MM e futebol para o LDA). Se considerarmos, a taxa de classificação errada, o LDA se mostra pior vendo que um dos seus tópicos concentrou a maior parte dos documentos de atletismo e tênis.

No entanto, ao avaliarmos as métricas para os diferentes valores de Ks, nenhuma nos retornou o $K = 5$ como o ideal. Além disso, vemos que apesar de o LDA, em geral, apresentar valores melhores das medidas nos Ks equivalentes no MM, ele normalmente nos retorna como número ideal, valores maiores de tópicos quando comparado com o MM, o que dá origem a modelos mais complexos e com muitos parâmetros, sendo de mais difícil interpretabilidade e com uma segmentação muito mais detalhada dos documentos.

3.2 Banco de dados *Associated Press*

O banco de dados *"Associated Press"* está disponível no *software* R, dentro do pacote *topicmodels*, que será utilizado para fazer a aplicação do LDA. O conjunto de dados é composto por artigos de notícias de uma agência americana, sendo que a maioria deles foi publicado nos anos 90. O mesmo dispõe de 2246 documentos, com 10473 termos no total.

Estes dados são tradicionais em estudos envolvendo modelagem de textos, com vários resultados disponíveis na literatura, entre os quais podemos citar os de Hou (2017) e Bhattacharya e Sil (2017).

Assim como no banco de dados estudado anteriormente, apesar do trabalho em questão ser escrito em português e os dados estarem dispostos na língua inglesa, preferimos deixá-lo em seu idioma original, a fim de evitar erros de tradução que possam impactar nos resultados finais. Diferentemente do *BBC Sport*, como os documentos deste banco não possuem uma classificação prévia, faremos apenas a análise das métricas e comparativo das metodologias.

Inicialmente, aplicamos algumas técnicas de pré-processamento nos dados. Foi feita a tokenização das palavras e remoção das *stopwords*, que acarretou na redução da quantidade inicial de palavras do dicionário, ficando com 10134. Com isso, os documentos ficaram, em média, com 168 termos, com o mínimo de 2 e o máximo de 554 palavras, ou seja, se tratam de textos de médio comprimento. Foram considerados $\boldsymbol{\alpha} = (1, ..., 1)$ e $\boldsymbol{\beta} = (1, ..., 1)$ como valores para os hiper-parâmetros. Depois, variamos os mesmos números de tópicos e consideramos as mesmas quantidades de iterações do banco de dados anterior, ou seja, $K = 1, 2, 3, ..., 18, 19, 20, 25, 30, 35, 40, 45, 50, 75, 100, 125, 150, 175, 200, 250, 300, 400$ e 500 para cada metodologia e utilizamos o amostrador de Gibbs para a inferência dos modelos, com 1000 iterações para cada valor de K no LDA e 500 no MM.

3.2.1 Comparativo das métricas e métodos

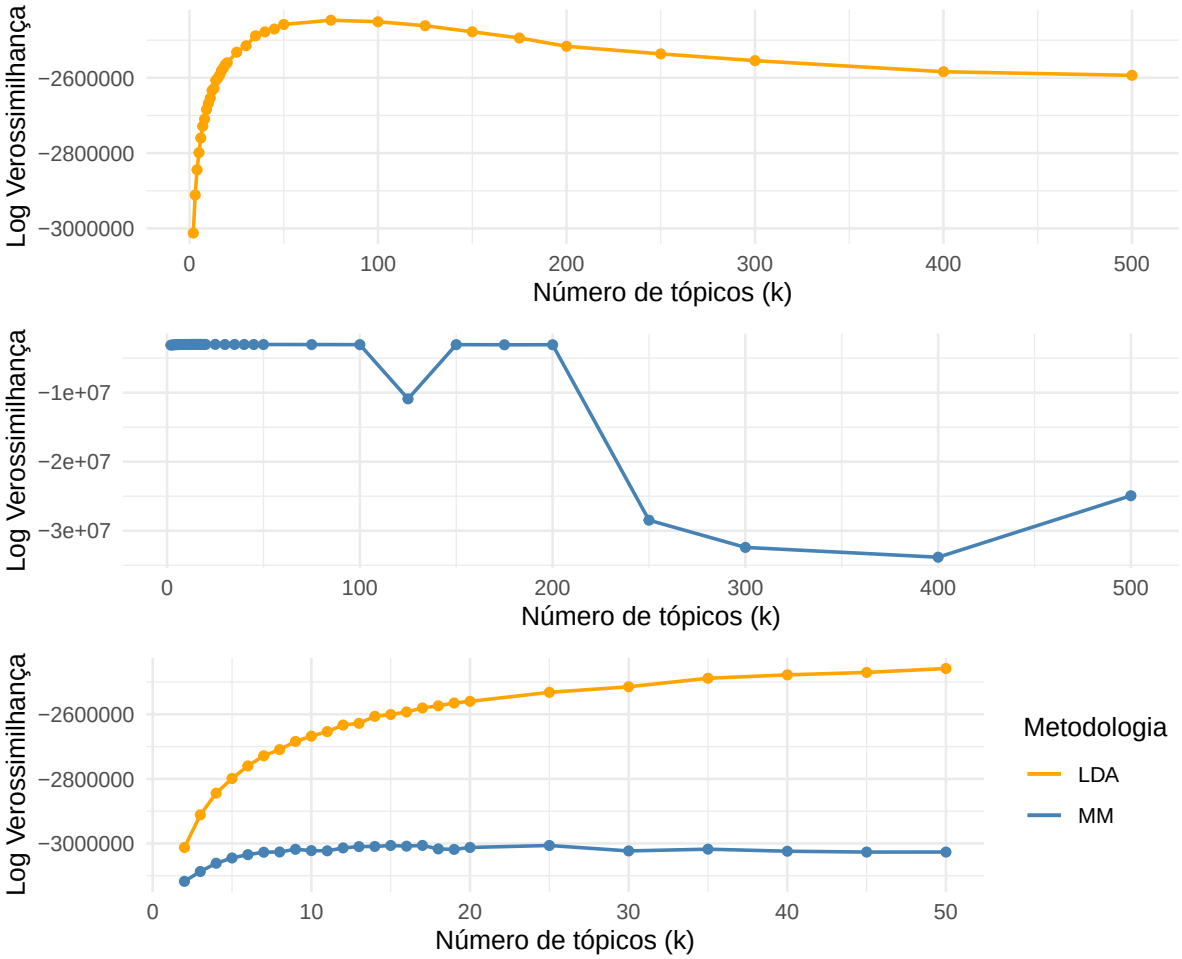

Figura 6 – *Associated Press* - Log Verossimilhança por número de tópicos.

A Figura 6 apresenta a log verossimilhança dos modelos ajustados considerando os diferentes números de tópicos e as duas metodologias. Acima estão as log verossimilhanças do LDA, no meio, as métricas no MM e, abaixo, estão os valores da medida com mais detalhes para K de 1 a 50. Lembrando que quanto maior o valor da métrica, mais adequado tende a ser o modelo, vemos que o LDA apresenta valores maiores para todos os números de tópicos quando comparado com o MM. No entanto, o pico do LDA, se dá no $K = 75$, enquanto o ápice do MM é o $K = 17$. Para o MM, no entanto, observamos que desde o $K = 9$, a métrica está razoavelmente estabilizada.

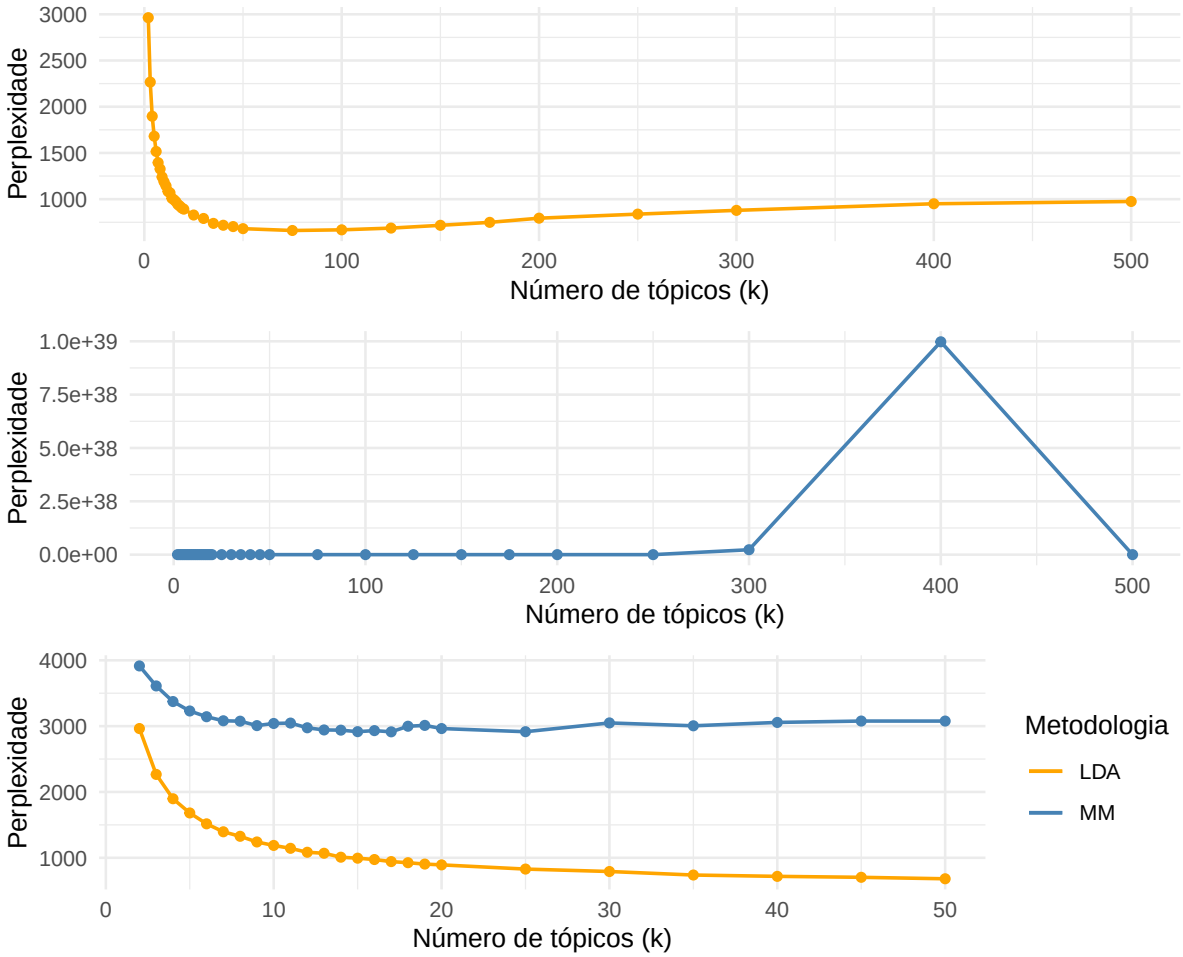

Figura 7 – *Associated Press* - Perplexidade por número de tópicos.

A Figura 7 ilustra a perplexidade dos modelos ajustados considerando os diferentes números de tópicos e as duas metodologias. Acima estão as perplexidades do LDA, no meio, as métricas no MM e, abaixo, estão os valores da medida com mais detalhes para K de 1 a 50. Como a perplexidade é uma medida parecida com a log verossimilhança, apesar da escala e de ser contrária em relação a interpretação, já que quanto menor o valor da métrica, mais adequado tende a ser o modelo, as conclusões se mantêm de maneira análoga ao que foi dito na métrica anterior.

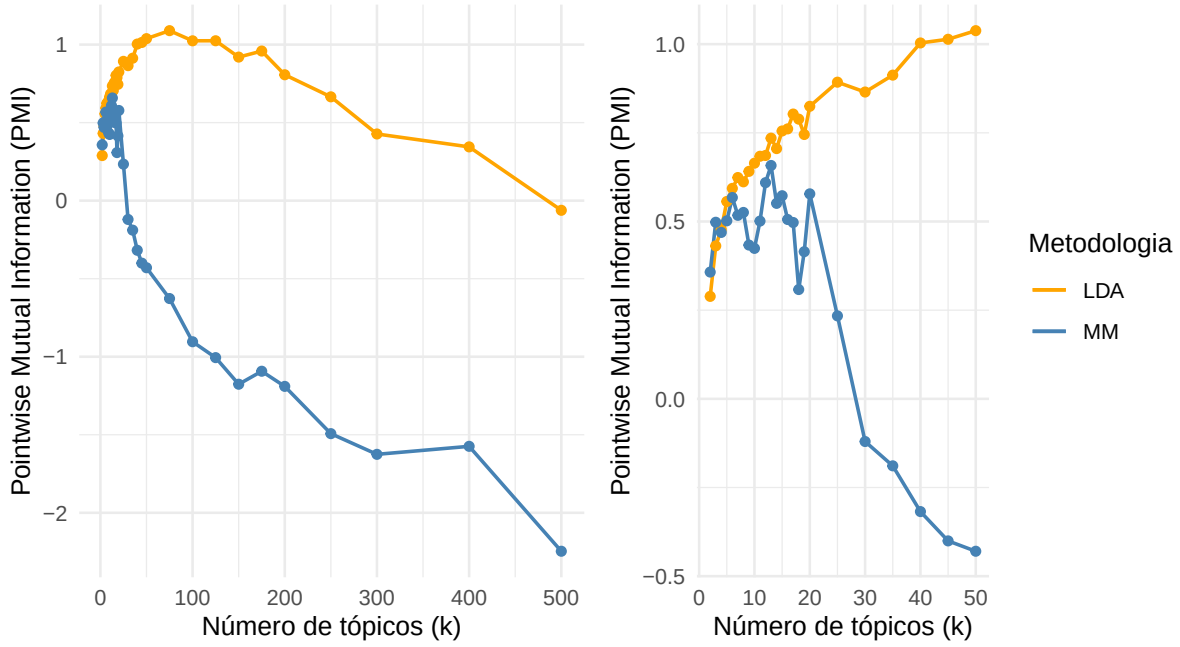

Figura 8 – *Associated Press* - PMI por número de tópicos.

A Figura 8 traz o PMI dos modelos ajustados considerando os diferentes números de tópicos e as duas metodologias. Na esquerda estão todos os valores de K estudados, enquanto na direita estão os valores das métricas com mais detalhes para K de 1 a 50. Lembrando que quanto maior o valor da métrica, mais adequado tende a ser o modelo, em geral o LDA apresenta valores maiores da medida quando comparado com o tópico equivalente no MM. Porém, enquanto o melhor valor para o LDA é o $K = 75$, para o MM é o $K = 13$.

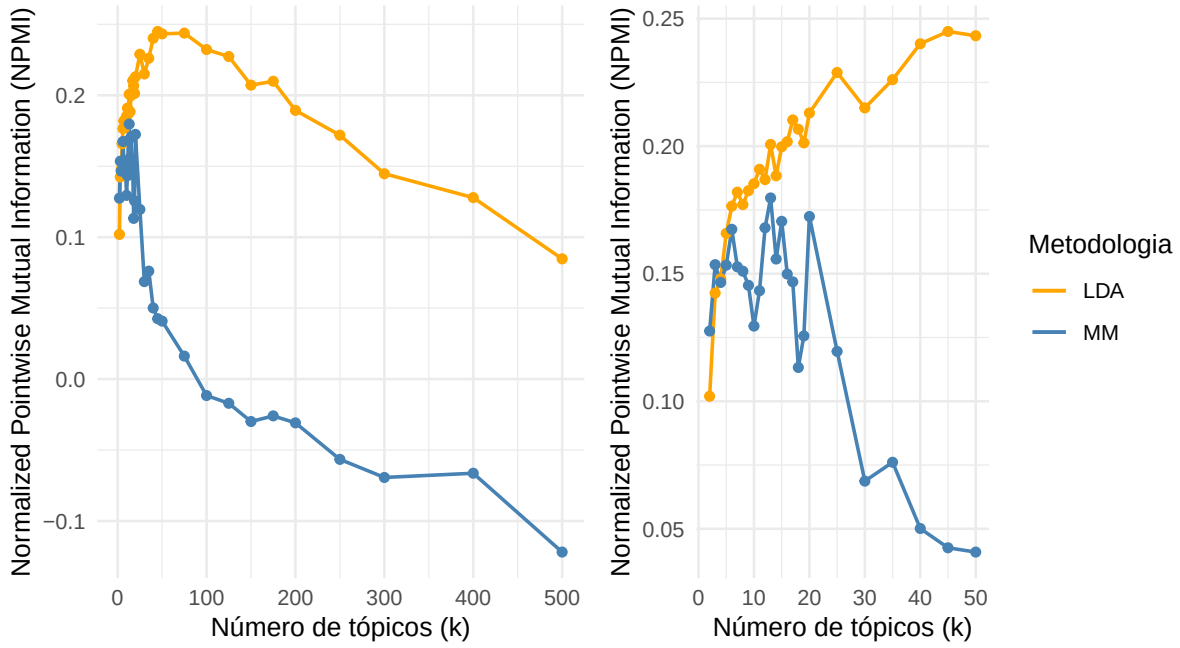

Figura 9 – *Associated Press* - NPMI por número de tópicos.

A Figura 9 mostra o NPMI dos modelos ajustados considerando os diferentes números de tópicos e as duas metodologias. Na esquerda estão todos os valores de K estudados, enquanto na direita estão os valores das métricas com mais detalhes para K de 1 a 50. Lembrando que quanto mais próximo de 1 o valor da métrica, maior tende a ser a associação entre as palavras dos tópicos e mais adequado o modelo ajustado. Assim como nas métricas analisadas anteriormente, o LDA, ao comparado com o MM, apresenta valores melhores para a métrica em quase todos os valores de Ks estudados. No entanto, novamente, o LDA nos retorna como maior valor da métrica um número de K maior, 45 contra 13 do MM.

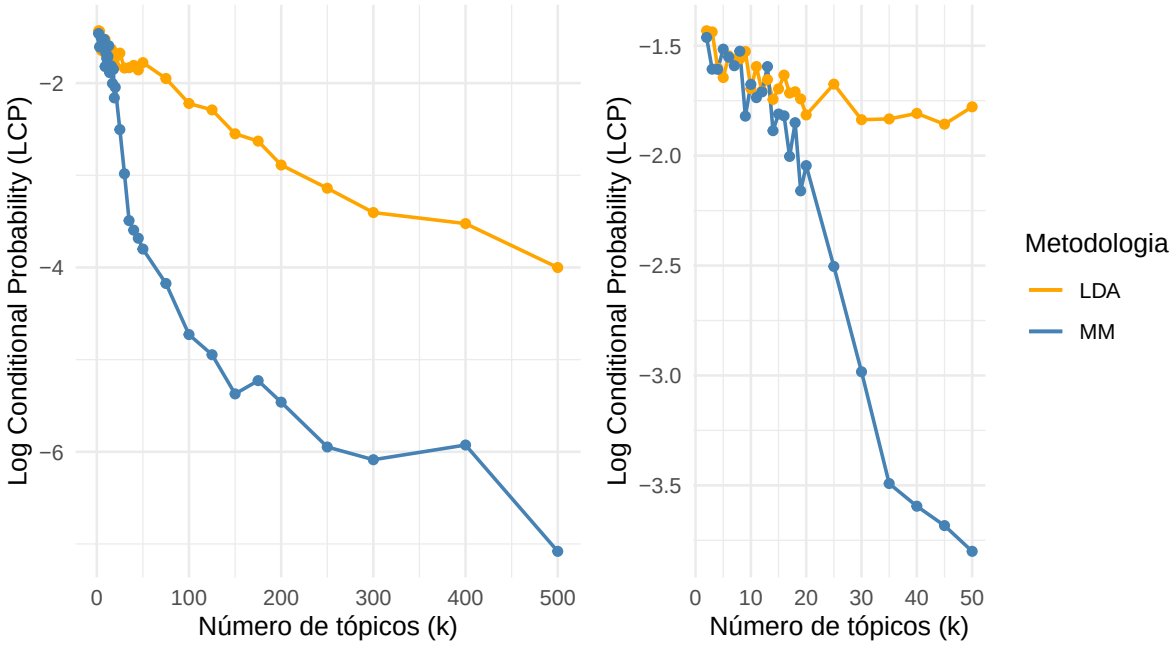

Figura 10 – *Associated Press* - LCP por número de tópicos.

A Figura 10 exibe o LCP dos modelos ajustados considerando os diferentes números de tópicos e as duas metodologias. Lembrando que quanto maior o valor da métrica, mais adequado tende a ser o modelo, em geral o LDA apresenta valores maiores da medida quando comparado com o tópico equivalente no MM. O melhor valor da medida se dá, para ambas as metodologias, no $K = 2$.

3.2.2 Considerações

Assim como no conjunto de dados visto na seção anterior, a métrica LCP se mostrou a mais restritiva escolhendo um número de tópicos menor que as outras métricas tanto para o MM quanto o LDA. Novamente, as métricas PMI e NPMI apresentaram comportamentos similares e o mesmo comentário vale para a log verossimilhança e a perplexidade. Tanto para o LDA, quanto para o MM, as quatro métricas escolheram número de tópicos parecidos (ou iguais) para cada metodologia, sendo o LDA com um número de tópicos ideal sugerido maior que o do MM em todas as métricas.

Ao analisar este mesmo banco de dados, Bhattacharya e Sil (2017) chegaram ao número de 25 tópicos como ótimo para o LDA. Para isso, eles levaram em consideração métricas como acurácia, precisão e sensibilidade. Por outro lado, Hou (2017) mostra que pela perplexidade, para o LDA, o $K = 100$ demonstra o melhor valor da métrica, mas destaca que a intuição e domínio do banco de dados é também de extrema importância para o pesquisador na definição desse número.

Mais uma vez, para verificarmos se nos modelos sugeridos pelas métricas, principalmente nos com valores de K maiores, há uma concentração em determinados tópicos dos documentos alocados, observamos que não há um acúmulo dos documentos apenas em determinados tópicos, ou seja, estão espalhados por todos os tópicos de maneira semelhante. No entanto, a fim de verificar os principais assuntos do conjunto de dados, iremos analisar as 5 palavras mais prováveis dos cinco tópicos mais representativos nos modelos com $K = 75$ e $K = 13$, para o LDA e MM, respectivamente. Números estes, escolhidos a partir das métricas de seleção, visto que o LDA nos retornou este valor de K para três das cinco métricas analisadas e o MM para duas, mas vemos que o $K = 13$ está contido também na estabilidade mencionada na log verossimilhança e perplexidade.

Tabela 7 – *Associated Press* - 5 palavras mais prováveis por tópico via LDA.

Tópico 1	**Tópico 34**	**Tópico 49**	**Tópico 37**	**Tópico 72**
dukakis	stock	soviet	police	air
campaign	market	gorbachev	shot	plane
jackson	index	union	killed	flight
democratic	exchange	moscow	city	airlines
republican	trading	mikhail	night	aircraft

Na Tabela 7 podemos visualizar, para o LDA, as 5 palavras mais prováveis nos cinco tópicos mais representativos, considerando o modelo ajustado com $K = 75$. Os tópicos 1, 34, 49, 37 e 72, possuem, respectivamente, 82, 74, 61, 58 e 55 documentos alocados dos 2246 totais. O tópico 1 apresenta "dukakis", que é o sobrenome do presidente dos Estados Unidos de 1989 a 1997, "democratic"e "republican"entre as suas cinco palavras mais prováveis, o que nos indica que os documentos deste tópico sejam do ramo de política. As palavras mais prováveis do tópico 34, como "market", "exchange"e "trading"são comumente utilizadas em notícias econômicas. O tópico 49 nos retornou dentre as palavras mais propícias, "gorbachev"e "mikhail", que foi um político russo, além de "soviet"e "union", ou seja, este tópico pode descrever as relações americano-soviéticas. As três palavras mais prováveis do tópico 37 são "police", "shot"e "killed", estas que, fazem alusão a notícias criminosas. Por fim, as cinco palavras mais propícias do tópico 72 são referentes a aviões e companhias aéreas.

Tabela 8 – *Associated Press* - 5 palavras mais prováveis por tópico via MM.

Tópico 2	**Tópico 11**	**Tópico 12**	**Tópico 7**	**Tópico 13**
government	percent	court	bush	people
police	million	federal	dukakis	fire
people	billion	government	campaign	water
president	market	trial	president	officials
united	stock	drug	percent	air

Na Tabela 8 podemos visualizar, para o MM, as 5 palavras mais prováveis nos cinco tópicos mais representativos, considerando o modelo ajustado com $K = 13$. Os tópicos 2, 11, 12, 7 e 13, possuem, respectivamente, 385, 369, 242, 219 e 197 documentos alocados dos 2246 totais. Os tópico 2 e 7, apresentam dentre as suas palavras mais propícias, algumas que são comumente vistas no ramo político, como "government"e "president"no 2 e "dukakis"e "campaign"no 7. Por outro lado, o tópico 11, apresenta palavras que fazem alusão a economia, como por exemplo "million", "billion"e "market". No tópico 12, dentre as palavras retornadas estão "court", "trial"e "drug", ou seja, nos indicam que são notícias que envolvam algum tipo de sentença, julgamento e afins. Por fim, no tópico 13, como as palavras "people", "fire", e "officials"estão nas mais propícias, nos indicam serem notícias que englobam incêndios.

CAPÍTULO

4

CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho apresentamos duas metodologias que são usualmente utilizadas para análise de tópicos em textos, o LDA e o MM. Apresentamos como esses modelos podem ser estimados na perspectiva Bayesiana e descrevemos e comparamos o comportamento de diferentes métricas que são usualmente utilizadas para escolher o valor mais adequado do número de tópicos.

Para os conjuntos de dados estudados, os modelos ajustados mais adequados para o LDA apresentaram valores maiores de tópicos quando comparado com o MM, o que dá origem a modelos mais complexos e com muitos parâmetros, sendo de mais difícil interpretabilidade e com uma segmentação muito mais detalhada dos documentos, sendo útil para um sistema de recomendação. Por outro lado, na grande maioria dos casos, ao compararmos os modelos das duas metodologias nos mesmos valores de tópicos, os do LDA retornaram valores melhores das métricas analisadas do que os do MM, que pode ser mais apropriado quando o objetivo é apenas classificação de textos.

Em relação as cinco métricas estudadas, pelos dois bancos de dados trabalhados, percebemos que a métrica LCP se mostrou a mais restritiva escolhendo um número de tópicos menor que as outras métricas tanto para o MM quanto o LDA. Pela baixa quantidade de tópicos selecionados, essa métrica não mostrou um bom desempenho na seleção do número de tópicos nas situações consideradas. Por outro lado, como era de se esperar, visto que a diferença entre as duas métricas é basicamente a escala, a log verossimilhança e a perplexidade apresentaram resultados iguais e que, em geral, quanto maior o número de tópicos, melhor tende a ser o modelo. Sendo assim, foram as métricas que sugeriram os maiores valores de tópicos nas duas metodologias. Por fim, o PMI e o NPMI também apresentaram resultados semelhantes e se mostraram as melhores medidas para a definição do número de tópicos dos modelos, visto que não são tão restritivas quanto o LCP, mas também não aparentaram ter o viés de escolha do número de tópicos tão grande quanto a log verossimilhança e a perplexidade. Para o MM, em

especial, reforçarmos que a log verossimilhança e a perplexidade se estabilizaram em números de tópicos próximos ao valor ideal pelo PMI e NPMI em ambos os conjuntos de dados analisados.

Como sugestão para estudos futuros, seria interessante verificar o comportamentos do modelo *Hierarchical Dirichlet Process* (HDP), visto que esta metodologia considera o número de tópicos como variável aleatória a ser estimada junto com os outros parâmetros do modelo, não precisando ser pré-estabelecido, e sem o esforço computacional de termos que estimar modelos com diferentes números de tópicos para escolher o mais adequado entre eles. Também, podem ser feitos estudos acerca da influência dos valores dos hiper-parâmetros na estimação dos modelos.

Além disso, Röder, Both e Hinneburg (2015) abordam o conceito de coerência de cada tópico, que pode ser uma alternativa para estudos futuros pensando na verificação da qualidade dos modelos.

Outros estudos comparativos também podem ser conduzidos considerando textos com comprimentos mais extremos. Os resultados aqui exibidos, consideram textos de médio comprimento, mas o desempenho das metodologias, assim como dos critérios de seleção, pode ser diferente em textos muitos curtos (textos de twitter, por exemplo) ou muito longos.

REFERÊNCIAS

BHATTACHARYA, I.; SIL, J. Sparse representation based query classification using LDA topic modeling. **Suresh Chandra Satapathy Vikrant Bhateja**, p. 621, 2017. Citado nas páginas 44 e 50.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet allocation. **The Journal of Machine Learning Research**, JMLR. org, v. 3, p. 993–1022, 2003. Citado nas páginas 22 e 30.

BOUMA, G. Normalized (pointwise) mutual information in collocation extraction. **Proceedings of GSCL**, Potsdam, v. 30, p. 31–40, 2009. Citado nas páginas 22 e 36.

BROOKS, S. P.; GELMAN, A. General methods for monitoring convergence of iterative simulations. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 7, n. 4, p. 434–455, 1998. Citado na página 29.

CASELLA, G.; GEORGE, E. I. Explaining the Gibbs sampler. **The American Statistician**, Taylor & Francis, v. 46, n. 3, p. 167–174, 1992. Citado na página 21.

CHANG, J.; GERRISH, S.; WANG, C.; BOYD-GRABER, J.; BLEI, D. Reading tea leaves: How humans interpret topic models. **Advances in Neural Information Processing Systems**, v. 22, 2009. Citado na página 35.

COWLES, M. K.; CARLIN, B. P. Markov chain Monte Carlo convergence diagnostics: a comparative review. **Journal of the American Statistical Association**, Taylor & Francis, v. 91, n. 434, p. 883–904, 1996. Citado na página 29.

FALEIROS, T. d. P.; LOPES, A. d. A. *et al.* Modelos probabilísticos de tópicos: desvendando o latent Dirichlet allocation. São Carlos, SP, Brasil., 2016. Citado na página 30.

FEINERER, I. Introduction to the tm package text mining in R. **Accessible en ligne: http://cran. r-project. org/web/packages/tm/vignettes/tm. pdf**, 2013. Citado na página 24.

GITHUB. 2020. Disponível em: <https://gist.github.com/alopes/5358189> . Citado na página 23.

GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 101, n. suppl 1, p. 5228–5235, 2004. Citado nas páginas 22, 28 e 33.

HORNIK, K.; GRÜN, B. topicmodels: An R package for fitting topic models. **Journal of Statistical Software**, American Statistical Association, v. 40, n. 13, p. 1–30, 2011. Citado na página 33.

HOU, J. 2017. Disponível em: <https://www.rpubs.com/JanpuHou/294548> . Citado nas páginas 44 e 50.

KAGGLE. 2020. Disponível em: <https://www.kaggle.com/datasets/maneesh99/sports-datasetbbc> . Citado na página 37.

MEI, Q.; SHEN, X.; ZHAI, C. Automatic labeling of multinomial topic models. In: **Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2007. p. 490–499. Citado nas páginas 22 e 35.

MIMNO, D.; WALLACH, H.; TALLEY, E.; LEENDERS, M.; MCCALLUM, A. Optimizing semantic coherence in topic models. In: **Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2011. p. 262–272. Citado nas páginas 22 e 36.

MULLEN, L. A.; BENOIT, K.; KEYES, O.; SELIVANOV, D.; ARNOLD, J. Fast, consistent tokenization of natural language text. **Journal of Open Source Software**, The Open Journal, v. 3, n. 23, p. 655, 2018. Citado na página 24.

NEWMAN, D.; SMYTH, P.; WELLING, M.; ASUNCION, A. Distributed inference for latent Dirichlet allocation. **Advances in Neural Information Processing Systems**, v. 20, 2007. Citado nas páginas 22 e 34.

PEEL, D.; MCLACHLAN, G. J. Robust mixture modelling using the t distribution. **Statistics and Computing**, Springer, v. 10, p. 339–348, 2000. Citado nas páginas 21 e 25.

PHAN, X.-H.; NGUYEN, L.-M.; HORIGUCHI, S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: **Proceedings of the 17th International Conference on World Wide Web**. [S.l.: s.n.], 2008. p. 91–100. Citado na página 33.

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**. [S.l.: s.n.], 2015. p. 399–408. Citado na página 54.

SYED, S.; SPRUIT, M. Full-text or abstract? examining topic coherence scores using latent Dirichlet allocation. In: IEEE. **2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)**. [S.l.], 2017. p. 165–174. Citado na página 35.

TEH, Y.; JORDAN, M.; BEAL, M.; BLEI, D. Sharing clusters among related groups: Hierarchical Dirichlet processes. **Advances in Neural Information Processing Systems**, v. 17, 2004. Citado na página 22.

TEH, Y.; NEWMAN, D.; WELLING, M. A collapsed variational bayesian inference algorithm for latent Dirichlet allocation. **Advances in Neural Information Processing Systems**, v. 19, 2006. Citado na página 33.

TITTERINGTON, D. M.; SMITH, A. F.; MAKOV, U. E. Statistical analysis of finite mixture distributions. **(No Title)**, 1985. Citado na página 25.

WALLACH, H. M.; MURRAY, I.; SALAKHUTDINOV, R.; MIMNO, D. Evaluation methods for topic models. In: **Proceedings of the 26th Annual International Conference on Machine Learning**. [S.l.: s.n.], 2009. p. 1105–1112. Citado nas páginas 22 e 34.

APÊNDICE

A

DISTRIBUIÇÃO DE DIRICHLET

A distribuição *Dirichlet*, que pode ser denotada por $Dirichlet(\alpha)$ e é a *priori* conjugada da distribuição $Multinomial$, é a utilizada para representar a distribuição de tópicos e sua função de densidade é representada por

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}, \tag{A.1}$$

em que $\theta = (\theta_1, \cdots, \theta_k)$ é um vetor aleatório de dimensão k, $\sum_{i=1}^{k}\theta_i = 1$, sendo que $\theta_i \geq 0$. Ainda, o vetor paramétrico $\alpha = (\alpha_1, \cdots, \alpha_k)$ é um vetor de dimensão k e $\Gamma(\cdot)$ é a função Gama.

A partir da Equação (A.1), pode-se provar que

- $E(\theta_i) = \frac{\alpha_i}{\sum_{i=1}^{k}\alpha_i}$;
- $Var(\theta_i) = \frac{\alpha_i(\sum_{i=1}^{k}\alpha_i)-\alpha_i^2}{(\sum_{i=1}^{k}\alpha_i)^3+(\sum_{i=1}^{k}\alpha_i)^2}$;
- $Cov(\theta_i, \theta_j) = -\frac{\alpha_i\alpha_j}{(\sum_{i=1}^{k}\alpha_i)^3+(\sum_{i=1}^{k}\alpha_i)^2}$.

APÊNDICE

B

CÓDIGOS DE PROGRAMAÇÃO

```
################## BIBLIOTECAS #################

library(ggplot2)
library(tibble)
library(tidyr)
library(readr)
library(dplyr)
library(tidytext)
library(topicmodels)
library(LDA)

################## BANCO DE DADOS ASSOCIATED PRESS #################

data("AssociatedPress", package = "topicmodels")

ap_td <- tidy(AssociatedPress)

ap_dtm <- ap_td %>%
  anti_join(stop_words, by = c(term = "word")) %>%
  cast_dtm(document, term, count)

dados = as.matrix(ap_dtm)

dados_ap = data.frame(dados)

################## EXEMPLO LDA K = 2 #################
```

```
### AJUSTANDO O MODELO
ap_lda2_ap <- LDA(dados_ap, k = 2, method="Gibbs")

### TOP 10 TERMOS DE CADA TÓPICO
ap_top_terms2 <- terms(ap_lda2_ap, 10)

### CRIANDO A MATRIZ DE COORRELAÇÃO DAS PALAVRAS
bow_ap <- matrix(unlist(dados_ap), nrow = nrow(dados_ap))
coorre_ap<-matrix(0,ncol(bow_ap),ncol(bow_ap))

for (i in 1:nrow(coorre_ap)){
  for (j in i:ncol(coorre_ap)){
    coorre_ap[i,j]<-sum(bow_ap[,i]>0 & bow_ap[,j]>0)+0.001
    if (j>i) coorre_ap[j,i]<-coorre_ap[i,j]}}
#Na diagonal principal temos o número de documentos em que cada palavra
ocorre e fora da diagonal principal temos o número de documentos em que o
par de palavras (i,j) acontece

probs_ocor_ap<-coorre_ap/nrow(bow_ap)
#Aqui de fato é a probabilidade de cada palavra e dos pares de palavras

### MÉTRICAS

#VEROSSIMILHANÇA
vero2_lda <- ap_lda2_ap@loglikelihood

#PERPLEXIDADE
perp2_lda <- exp(-vero2_ap/sum(dados_ap))

#Ajustando as top 10 palavras para o cálculo das demais métricas
ap_top_terms_2 = t(ap_top_terms2)
ap_top_terms_2_v2 = matrix(unlist(ap_top_terms_2), nrow = nrow(ap_top_terms_2))
ap_top_terms_2_v3<-matrix(0,nrow(ap_top_terms_2_v2),ncol(ap_top_terms_2_v2))

for (i in 1:K_2){
  for (j in 1:10){
    ap_top_terms_2_v3[i,j] = which(colnames(dados_ap) == ap_top_terms_2_v2[i,j])
```

```
  }}
pa_top_terms2 = ap_top_terms_2_v3


pmi_lda<-NULL
npmi_lda<-NULL
lcp_lda<-NULL
modelos_ap<-c(2) #vetor de ks
for (mod in 1:length(modelos_ap)){
  K<-modelos_ap[mod]
  top_termos<-get(paste("pa_top_terms",eval(modelos_ap[mod]),sep=""))

  soma1<-0
  soma2<-0
  soma3<-0
  for (topicos in 1:K){
    for (i in 1:(length(top_termos[topicos,])-1)){
      for (j in (i+1):length(top_termos[topicos,])){

        soma1<-soma1+
          log((probs_ocor_ap[top_termos[topicos,i],top_termos[topicos,j]])/
                ((probs_ocor_ap[top_termos[topicos,i],top_termos[topicos,i]])*
                   (probs_ocor_ap[top_termos[topicos,j],top_termos[topicos,j]])))


        soma2<-soma2+
          (log((probs_ocor_ap[top_termos[topicos,i],top_termos[topicos,j]])/
                 ((probs_ocor_ap[top_termos[topicos,i],top_termos[topicos,i]])*
                    (probs_ocor_ap[top_termos[topicos,j],top_termos[topicos,j]])))
             /-log(probs_ocor_ap[top_termos[topicos,i],top_termos[topicos,j]]))

        soma3 <-soma3+
          log((probs_ocor_ap[top_termos[topicos,i],top_termos[topicos,j]])/
                ((probs_ocor_ap[top_termos[topicos,i],top_termos[topicos,i]])))


      }}}
  pmi_lda[mod]<-soma1/(K*45)
  npmi_lda[mod]<-soma2/(K*45)
```

```
  lcp_lda[mod]<-soma3/(K*45)
}
pmi_lda
npmi_lda
lcp_lda

################## EXEMPLO MM K = 2 ##################

#Simula valores de uma distribuição discreta para simular as variáveis S

rDiscreta<-function(p){
  u<-runif(1)
  P<-cumsum(p)
  val<-sum(P<u)+1
  val}

#Simula vetores da distribuição Dirichlet
rDiric<-function(gama){
  X<-rgamma(length(gama),gama,1)
  Y<-X/sum(X)
  return(Y)}

y_ap<-as.matrix(dados_ap) #documentos

#Valores dos hiperparâmetros das distribuições a priori (alfa e beta)
a0<-1
b0<-1


K_2<-2 #fixar o número de tópicos


#Chute inicial (aleatório) para o vetor S
set.seed(nrow(dados_ap))
S_2_ap<-sample(1:K_2,nrow(y_ap),replace=T)
S_tot_2_ap<-S_2_ap
table(S_2_ap)
n_k_2_ap<-NULL
for (k in 1:K_2) n_k_2_ap[k]<-sum(S_2_ap==k)
```

```
posteriori_theta_2_ap <-matrix(rDiric(n_k_2_ap+a0),nrow=K_2,ncol=1)
#posteriori Dirichlet correta da probabilidade de cada tópico
posteriori_theta_tot_2_ap<-posteriori_theta_2_ap

#Primeiro passo das probabilidades de cada palavra dentro de cada tópico da
posteriori Dirichlet dela.
m_2_ap =  matrix(data=NA,nrow=K_2,ncol=ncol(dados_ap))
for (k in 1:K_2) m_2_ap[k,]<-apply(as.matrix(y_ap[S_2_ap==k,],
ncol=ncol(y_ap)),2,sum)

posteriori_phi_k_2_ap<-NULL
for (k in 1:K_2) posteriori_phi_k_2_ap<-rbind(posteriori_phi_k_2_ap,
rDiric(m_2_ap[k,] + b0))
posteriori_phi_tot_2_ap = posteriori_phi_k_2_ap

### GIBBS

library(compiler)

iter<-500
for (it in 1:iter){
  cat('\n', K_2, it)
  #Atualiza os Ss
  enableJIT(3)
  for (i in 1:nrow(dados_ap)){
    log_probs1_ap<-NULL
    for (top in 1:K_2) {
      log_probs1_ap[top]<-dmultinom(y_ap[i,], size = sum(y_ap[i,]),
      prob=posteriori_phi_k_2_ap[top,], log = TRUE)}
    log_probs_ap<-log(c(posteriori_theta_2_ap))+log_probs1_ap
    log_probs_ap<-log_probs_ap-max(log_probs_ap)
    probs_ap<-exp(log_probs_ap)
    probs_ap<-probs_ap/sum(probs_ap)
    S_2_ap[i]<-rDiscreta(probs_ap)}
  S_tot_2_ap<-rbind(S_tot_2_ap,S_2_ap)

  #Atualiza theta
  for (k in 1:K_2){
```

```
    n_k_2_ap[k]<-sum(S_2_ap==k)}
  posteriori_theta_2_ap<-rDiric(n_k_2_ap+a0)
  posteriori_theta_tot_2_ap<-cbind(posteriori_theta_tot_2_ap,
  posteriori_theta_2_ap)

  #Atualiza phi
  for (k in 1:K_2){
    m_2_ap[k,]<-apply(as.matrix(y_ap[S_2_ap==k,],ncol=ncol(y_ap)),2,sum)
    posteriori_phi_k_2_ap[k,] = rDiric(m_2_ap[k,] + b0)}

}

#Theta da última iteração
theta_2 = posteriori_theta_tot_2_ap[,501]

###MÉTRICAS

soma1_2<-0
soma2_2<-0
soma3_2<-0
posicoes_2 <- matrix (0,nrow = K_2, ncol = 10)

for (topicos in 1:K_2){
  for (i in 1:(length(posicoes_2[topicos,])-1)){
    for (j in (i+1):length(posicoes_2[topicos,])){

      posicoes_2[topicos,]<-rev(tail(order(posteriori_phi_k_2_ap[topicos,]), 10))

      soma1_2<-soma1_2+
        log((probs_ocor_ap[posicoes_2[topicos,i],posicoes_2[topicos,j]])/
              ((probs_ocor_ap[posicoes_2[topicos,i],posicoes_2[topicos,i]])*
                 (probs_ocor_ap[posicoes_2[topicos,j],posicoes_2[topicos,j]])))


      soma2_2<-soma2_2+
        (log((probs_ocor_ap[posicoes_2[topicos,i],posicoes_2[topicos,j]])/
               ((probs_ocor_ap[posicoes_2[topicos,i],posicoes_2[topicos,i]])*
                  (probs_ocor_ap[posicoes_2[topicos,j],posicoes_2[topicos,j]]))))/
           -log(probs_ocor_ap[posicoes_2[topicos,i],posicoes_2[topicos,j]]))
```

```
      soma3_2 <-soma3_2+
        log((probs_ocor_ap[posicoes_2[topicos,i],posicoes_2[topicos,j]])/
              ((probs_ocor_ap[posicoes_2[topicos,i],posicoes_2[topicos,i]])))


    }}}
pmi_mm_2<-soma1_2/(K_2*45)
npmi_mm_2<-soma2_2/(K_2*45)
lcp_mm_2<-soma3_2/(K_2*45)


soma4_2<-0
soma5_2<-0

for (topicos in 1:K_2){

  soma4_2<-soma4_2+
    n_k_2_ap[topicos]*log(theta_2[topicos])

  for (i in 1:ncol(y_ap)){

    soma5_2<-soma5_2+
      m_2_ap[topicos,i]*log(posteriori_phi_k_2_ap[topicos,i])

  }}

vero_mm_2 = soma4_2+soma5_2
perp_mm_2 = exp(-vero_2/sum(y_ap))
```


ICMC
USP
SÃO CARLOS