JOSÉ AHIRTON BATISTA LOPES FILHO

PÓS TECH

MACHINE LEARNING ENGINEERING

APRENDIZADO NÃO SUPERVISIONADO

AULA 05

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

SUMÁRIO

FIAP

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

O QUE VEM POR AÍ?

Você conhece o LDA (Latent Dirichlet Allocation) ou Alocação Latente de Dirichlet? Este é um modelo estatístico probabilístico utilizado para descobrir tópicos ocultos em grandes corpora textuais.

Ele é amplamente utilizado em áreas como processamento de linguagem natural, mineração de texto e análise de dados. Nesta aula, vamos entender como esse modelo funciona e como utilizar a associação de similaridade em corpora textuais para agrupar palavras em grupos.

FIAP

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

HANDS ON

Vamos aprender na prática, como utilizar o algoritmo de LDA para realizar uma modelagem de tópicos no mundo de processamento de linguagem natural. Assim, vamos nos aprofundar nos seguintes tópicos em Python:

1. **Introdução ao Caso de Uso**
   - Apresentação do objetivo: modelagem de tópicos usando LDA.
   - Descrição do conjunto de dados: livros da franquia "O Mágico de Oz".
2. **Pré-Processamento de Dados**
   - Explicação das etapas de limpeza e preparação dos textos.
   - Ferramentas e técnicas utilizadas para o pré-processamento de linguagem natural.
3. **Implementação do LDA**
   - Passo a passo da implementação do algoritmo LDA no conjunto de dados.
   - Utilização de bibliotecas populares em Python, como scikit-learn.
4. **Análise dos Resultados**
   - Extração e visualização dos tópicos gerados pelo LDA.
   - Interpretação dos tópicos e discussão sobre a relevância dos resultados obtidos.
5. **Conclusão e Reflexão**
   - Resumo das descobertas e insights obtidos a partir da modelagem de tópicos.
   - Considerações sobre as aplicações práticas e futuras pesquisas na área de processamento de linguagem natural.

FIAP

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

SAIBA MAIS

O LDA (Latent Dirichlet Allocation) tem o objetivo de atribuir a um determinado documento probabilidades de pertencimento a um número k de tópicos. Como é uma tarefa não-supervisionada, não há conhecimento dos tópicos de antemão. Assim, eles emergirão durante o processo de modelagem. A principal ideia aplicada a esse algoritmo é a seguinte:

"Cada documento pode ser descrito por uma distribuição de tópicos e cada tópico pode ser descrito por uma distribuição de palavras".

Assim, podemos dizer que um documento foi gerado a partir de um conjunto de tópicos e cada tópico a partir de um conjunto de palavras.

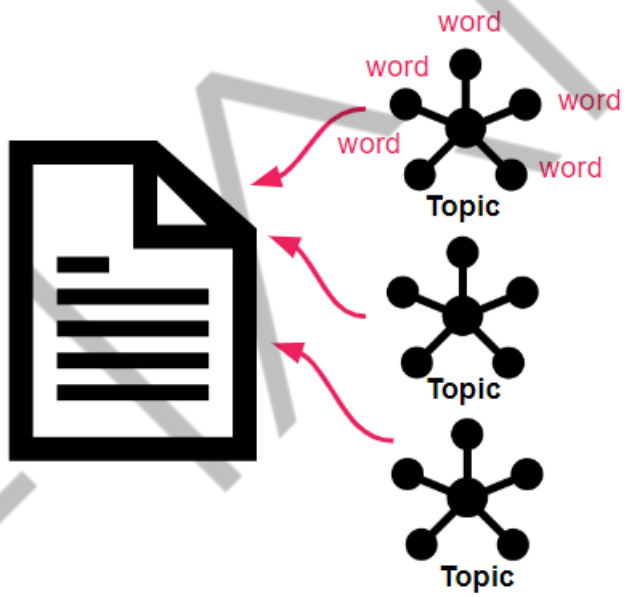

Figura 1 - Modelando documentos a partir de tópicos
Fonte: Elaborado pela autora (2024)

Agora, observe o texto a seguir:

"Muito se fala em como a inteligência artificial pode colocar diversas carreiras em extinção. Uma delas é a de agente de viagens --- você não precisa mais de dias de espera por um roteiro dos sonhos. Layla, uma IA para roteiro de viagem, já pode fazer isso por você. O aplicativo Just Ask Layla é capaz de construir roteiros de viagens completos de acordo com a necessidade do usuário" (OLGA, 2024).

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

Que tipo de classificação de tema você atribuiria a esse texto? Sobre inteligência artificial? Sobre turismo? Acredito que você colocaria as duas opções, certo? E como chegamos nessa conclusão? Pelas palavras contidas no texto? Perceba que ambos os grupos fazem parte do mesmo documento, ou seja, em um mesmo documento podem existir vários tópicos diferentes.

Isso pode ser um problema para os paradigmas da classificação e agrupamento, pois o objetivo dessas técnicas é definir uma classe para categorizar os dados. Por isso, temos como uma possível solução para esse problema: o LDA, em que a técnica atribui para um determinado documento probabilidades de pertencimento a um número “k” de tópicos.

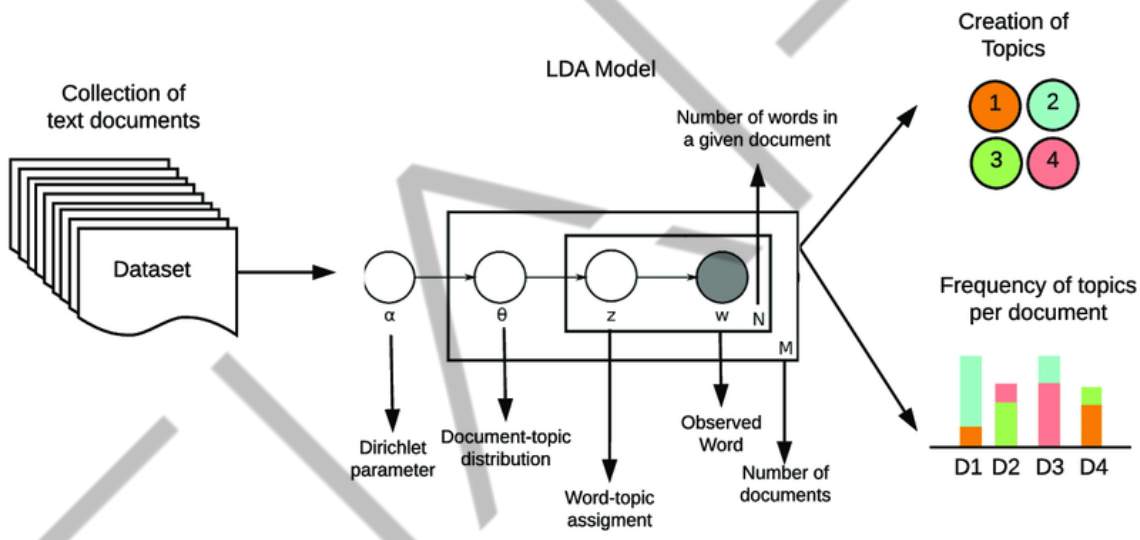

Figura 2 – Exemplo de LDA
Fonte: Bakrey (2023)

Como o LDA descobre os tópicos?

Imagine um conjunto de 1000 palavras e 1000 documentos. Assuma que cada documento possui, em média, 500 dessas palavras. Como descobrir a categoria a que cada documento pertence? Uma maneira é conectar cada documento a cada palavra com base em sua aparição no documento.

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

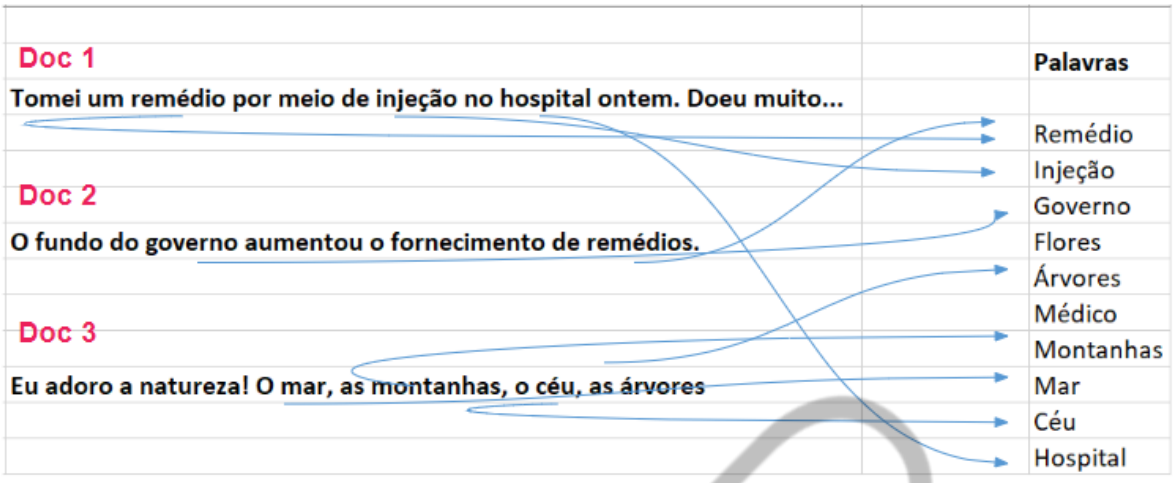

Figura 3 – Descobrindo tópicos por número de aparição de palavras no documento
Fonte: Elaborado pela autora (2024)

Entretanto, essa abordagem **não é escalável**, pois ficaria impossível, visualmente falando, conseguir identificar todas as relações. Então, como resolveremos? É possível utilizar uma camada oculta nomeada como **"latent"**, supondo que existam k tópicos que apareçam em todos os documentos.

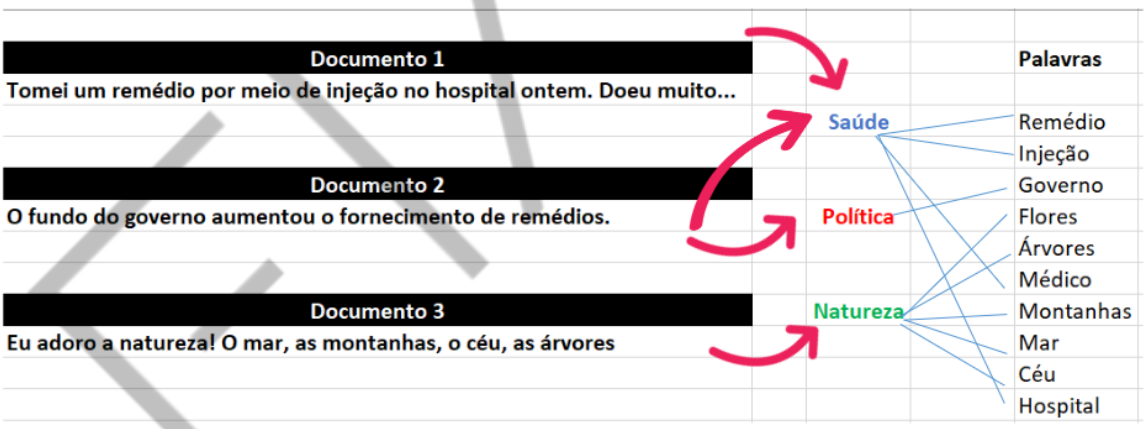

Figura 4 – Descobrindo tópicos por número "k" de tópicos existentes no documento
Fonte: Elaborado pela autora (2024)

Assim, posso usar essa informação **conectando palavras a tópicos**, dependendo quão bem essa palavra se ajuste a esse tópico, e então conectar os tópicos aos documentos com base nos tópicos abordados em cada documento.

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

Figura 5 – Como os tópicos são representados
Fonte: Elaborado pela autora (2024)

Cada documento será atribuído a um determinado tópico de acordo com a probabilidade de pertencimento: quanto maior a probabilidade de pertencer a um determinado tópico x, esse será considerado o tema que representa esse documento.

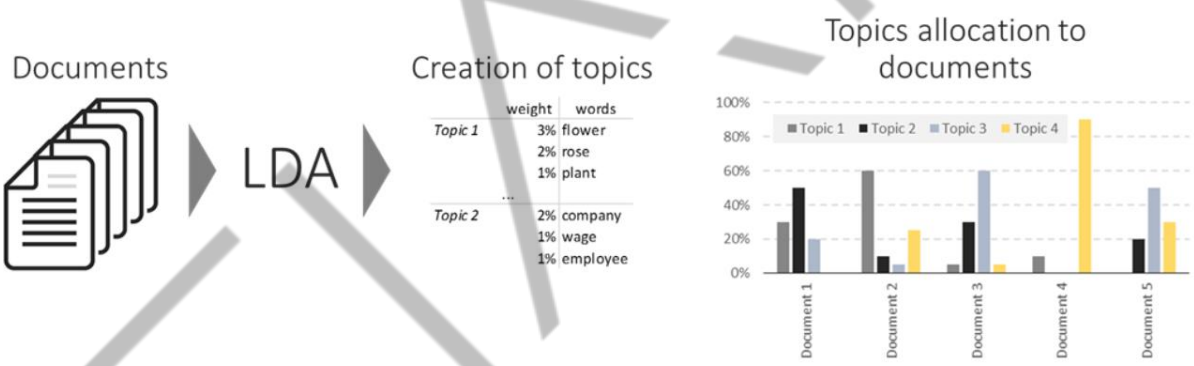

Figura 6 – Criação dos tópicos
Fonte: Elaborado pela autora (2024)

Funcionamento do LDA

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

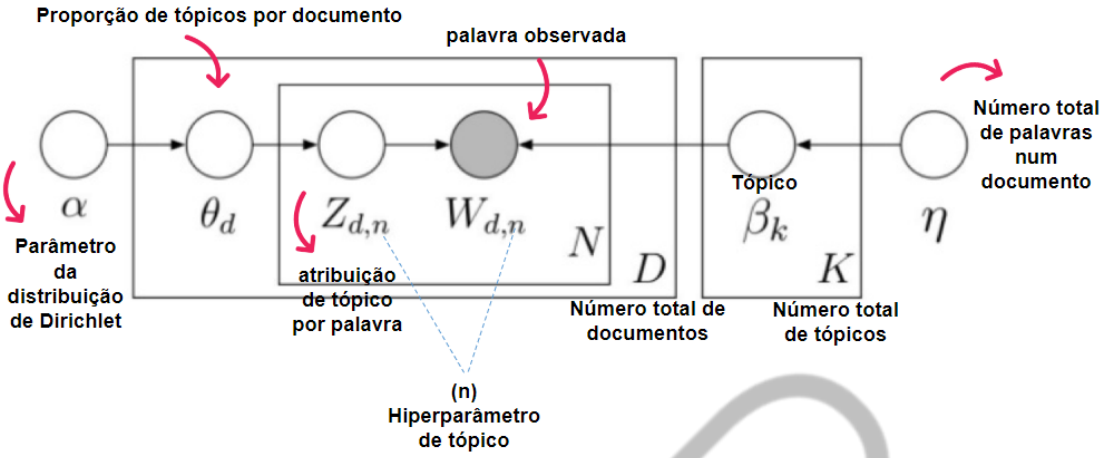

Figura 7 – Funcionamento do LDA
Fonte: Elaborado pela autora (2024)

- **Modelagem Probabilística:** o LDA assume que cada documento na coleção é uma mistura de tópicos abstratos e cada tópico é composto por um conjunto de palavras com alta probabilidade de co-ocorrência.
- **Distribuições Latentes:** o modelo estima duas distribuições latentes:
  - **Distribuição por Tópico por Documento:** define a proporção de cada tópico em cada documento.
  - **Distribuição por Palavra por Tópico:** define a probabilidade de cada palavra aparecer em cada tópico.
- **Aprendizado Iterativo:** o LDA utiliza um algoritmo iterativo para estimar essas distribuições latentes, buscando a configuração que melhor explica a co-ocorrência das palavras nos documentos.

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

MERCADO, CASES E TENDÊNCIAS

Que tal dar uma conferida nesse artigo que traz um experimento utilizando o Latent Dirichlet Allocation (LDA) para modelar informações na tecnologia Blockchain? Os resultados do modelo Latent Dirichlet Allocation são analisados com base em vários termos-chave extraídos e documentos-chave encontrados para cada tópico.

Esses tópicos podem orientar os pesquisadores e pesquisadoras a buscarem pesquisas em tendências específicas e também encontrar lacunas de pesquisa em diversas tecnologias associadas à Tecnologia Blockchain. Veja mais aqui.

FIAP

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

O QUE VOCÊ VIU NESTA AULA?

Nesta aula exploramos a modelagem de Tópicos com o Algoritmo LDA, um algoritmo que une a clusterização e o processamento de linguagem natural (PLN) para gerar agrupamentos valiosos de corpus de documentos em tópicos.

FIAP

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

REFERÊNCIAS

BAKREY, M. **All about Latent Dirichlet Allocation (LDA) in NLP**. 2023. Disponível em: <https://mohamedbakrey094.medium.com/all-about-latent-dirichlet-allocation-lda-in-nlp-6cfa7825034e>. Acesso em: 15 jul. 2024.

CASTRO, B. Y. S. **Como modelar tópicos através de Latent Dirichlet Allocation (LDA) através da biblioteca Gensim**. 2020. Disponível em: <https://medium.com/somos-tera/como-modelar-tópicos-através-de-latent-dirichlet-allocation-lda-através-da-biblioteca-gensim-1fa17357ad4b>. Acesso em: 15 jul. 2024.

OLGA, J. **Conheça Layla**, uma IA para roteiro de viagem. 2024. Disponível em: <https://epocanegocios.globo.com/inteligencia-artificial/noticia/2024/06/conheca-layla-uma-ia-para-roteiro-de-viagem.ghtml>. Acesso em: 15 jul. 2024.

PIRES, L. **Modelagem de Tópicos em Python utilizando o Modelo de Alocação Latente de Dirichlet (LDA)**. 2021. Disponível em: <https://medium.com/leti-pires/modelagem-de-tópicos-em-python-utilizando-o-modelo-de-alocação-latente-de-dirichlet-lda-3276a469f421>. Acesso em: 15 jul. 2024.

SETH, N. **Part 2**: Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn. 2021. Disponível em: <https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/>. Acesso em: 15 jul. 2024.

FIAP

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

PALAVRAS-CHAVE

Inteligência Artificial. LDA. Similaridade de Palavras.

FIAP

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com

PDF exclusivo para Marcio Anastacio da Luz Junior - rm354541
marcioluzj@gmail.com