

Tera

AULA 28

Unsupervised Learning:
Topic Analysis

Instrutor: [Raphael Ballet](#)

Background:

- Engenheiro de Controle e Automação (IMT)
- Mestre em Sistemas Aeroespaciais e Mecatrônica (ITA)
- Data Scientist - Elo7

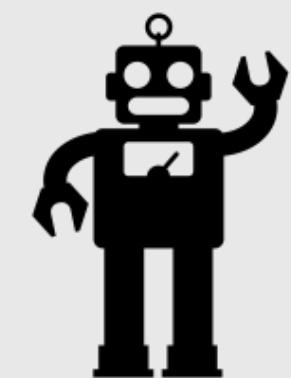
Interesses:



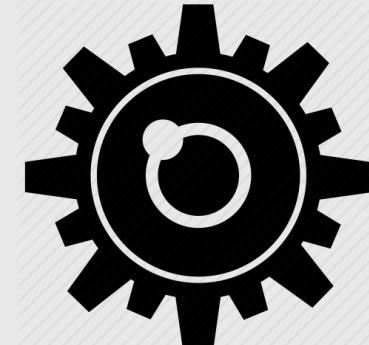
Drones



Aprendizado
de Máquina



Robótica



Visão
Computacional



Processamento de
Linguagem Natural



Sistemas de
recomendação

Planejamento:

1. Introdução a topic analysis
2. Non-Negative Matrix Factorization (NMF)
3. Latent Dirichlet Allocation (LDA)
4. Sistemas de recomendação – Topic Analysis
5. Case

1. INTRODUÇÃO:

- O que são tópicos de um documento?
 - Variáveis latentes (implícitas)
 - Temas

News

U.S. edition

Modern

Top Stories

Donald Trump
Elizabeth II
Harriet Tubman
Michael Strahan
Syria
Game of Thrones
Flint water crisis
Chicago Blackhawks
Curt Schilling
Zika virus

New York, New York

Suggested for you

World

U.S.

Elections

Business

Technology

Entertainment

Sports

Science

Health

Top Stories

[See realtime coverage](#)**Tubman's In. Jackson's Out. What's It Mean?**New York Times - 3 hours ago 

Harriet Tubman, left, will replace Andrew Jackson on the center of a new \$20 note. Credit Left to right: H. B. Lindsley/Library of Congress; U.S.

[Harriet Tubman to be first African-American on US currency](#) Reuters[Overnight Finance: Harriet Tubman bumps Jackson on the new \\$20](#) The Hill[Highly Cited: Tubman replacing Jackson on the \\$20, Hamilton spared](#) Politico[Featured: It's Official: Harriet Tubman Will Grace the \\$20 Bill](#) Smithsonian[Opinion: Applaud Hamilton call: Our view](#) USA TODAY[Wikipedia: United States twenty-dollar bill](#)**Cluster 1**

Related

[Harriet Tubman »](#)[Andrew Jackson »](#)**ESPN fires Curt Schilling following anti-transgender post on Facebook**

USA TODAY - 39 minutes ago

Broadcaster Curt Schilling has been fired by ESPN in the wake of outrage following a Facebook post against the transgender community.

Cluster 2**Charges against 3 in Flint water crisis 'only the beginning'**

CNN - 58 minutes ago

(CNN) Criminal charges against three men in Michigan on Wednesday marked a milestone in a crisis that's been years in the making, potentially harmed tens of thousands of people and cast a harsh spotlight on infrastructure issues across the country.

Cluster 3**Three dead, dozens injured in blast at chemical plant in Mexico**

Reuters - 21 minutes ago

MEXICO CITY A massive explosion rocked a major petrochemical facility of Mexican national oil company Pemex in the Gulf state of Veracruz on Wednesday, killing at least three people, injuring dozens more, and pumping a cloud of noxious chemicals into ...

Cluster 4

1. INTRODUÇÃO:

- Como uma pessoa agruparia os artigos?
 - **Temas** → política, música, esporte
 - **Período** → ano, mês, década
 - **Fonte** → jornal, blogs

1. INTRODUÇÃO:

- O que faz um artigo ser semelhante a outro?
 - Conjunto de palavras
 - Conjunto de temas / tópicos

1. INTRODUÇÃO:

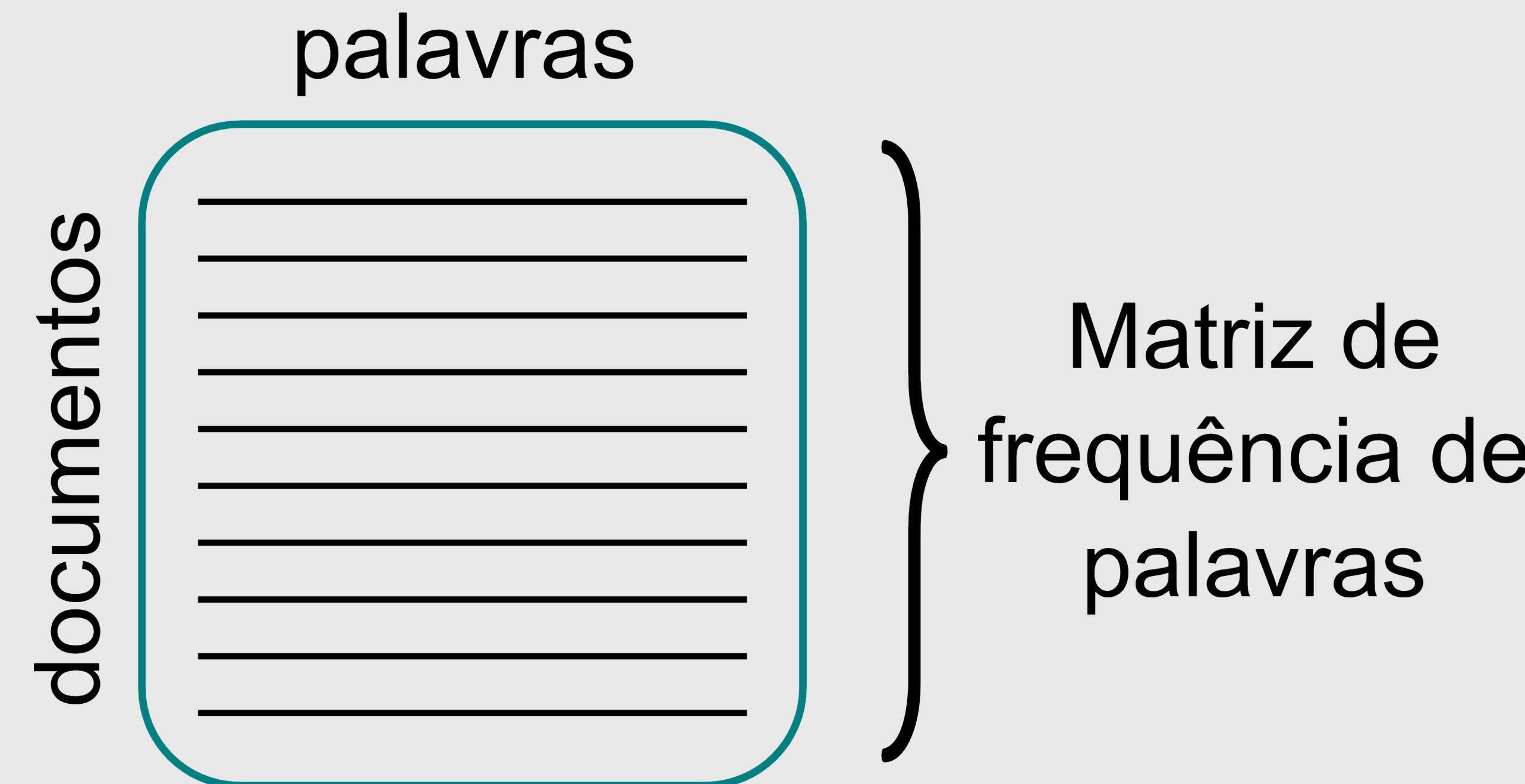
- Mas, a final, o que define um tópico?
 - Distribuição de palavras-chave
 - Documentos

1. INTRODUÇÃO:

- Como automatizamos esse processo?
 - Aprendizagem não supervisionada → Clustering

1. INTRODUÇÃO:

- Problema já conhecido:
 - Documentos + Palavras = Muitas dimensões



1. INTRODUÇÃO:

- Mas, já não lidamos com isso?
 - PCA + K-Means = Sucesso

1. INTRODUÇÃO:

- Problema principal do PCA:
 - Falta **interpretabilidade**:
 - Eixos são combinações lineares de palavras
 - Não podemos definir tópicos

1. INTRODUÇÃO:

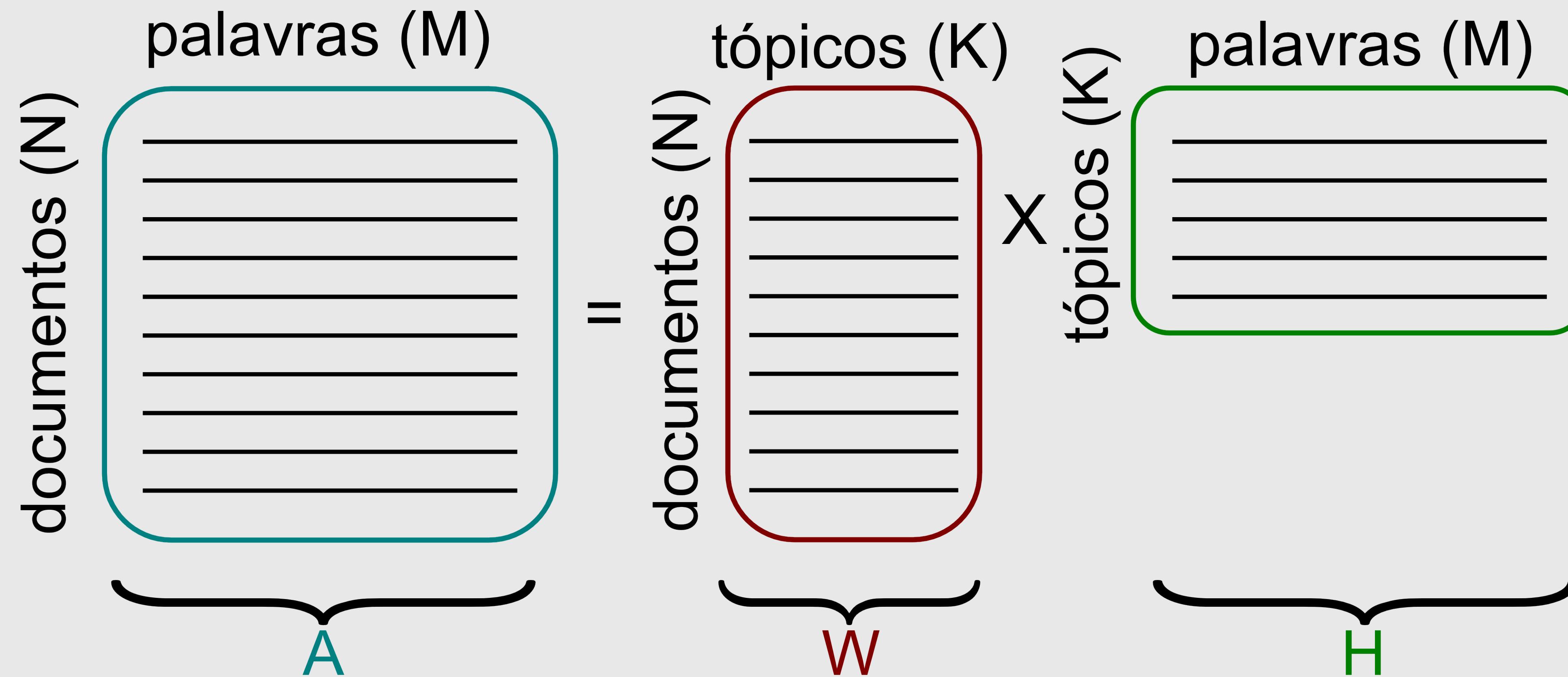
- Principais alternativas ao PCA:
 - Non-Negative Matrix Factorization (NMF)
 - Latent Dirichlet Allocation (LDA)

2. Non-Negative Matrix Factorization (NMF)

- Principal objetivo:
 - Decompor a matriz de frequência de palavras em representações de tópicos
 - Documentos são compostos de combinações de tópicos
 - Tópicos são compostos de combinações de palavras

2. Non-Negative Matrix Factorization (NMF)

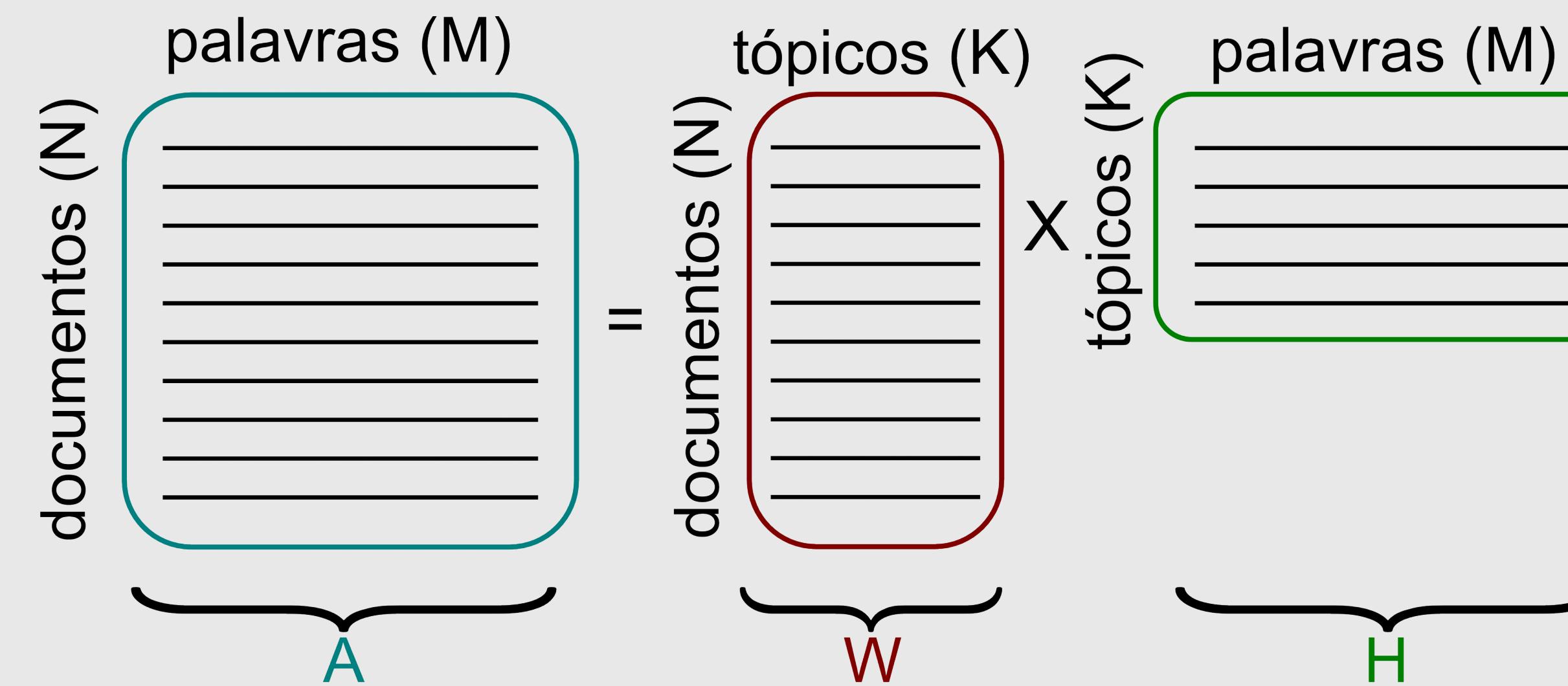
- NMF: Fatoração



2. Non-Negative Matrix Factorization (NMF)

- Matrizes:

- A: Matriz de frequênciа de termos (M) em documentos (N)
- W: Matriz de pesos → dist. tópicos (K) nos documentos
- H: Matriz de atributos → dist. palavras nos tópicos



2. Non-Negative Matrix Factorization (NMF)

- Principais características:
 - Precisa definir o número de tópicos
 - Matrizes A, W e H não podem ter valores negativos
 - Matrizes W e H podem reconstruir matriz A (aprox.)

2. Non-Negative Matrix Factorization (NMF)

- NMF pode ser utilizado em vários outros cenários:
 - Segmentação de fontes sonoras do áudio:
 - Documentos: áudio
 - Features: espectograma do áudio
 - Segmentação de imagens:
 - Documentos: imagem
 - Features: pixels

2. Non-Negative Matrix Factorization (NMF)

- Exemplo: notebook

2. Non-Negative Matrix Factorization (NMF)

- **Vantagens:**

- Tópicos são interpretáveis
- Naturalmente agregador (clustering)
- Tópicos gerados são, normalmente, mais coerentes
- Pode ser utilizado em outros contextos (ex: imagens, áudio etc)

2. Non-Negative Matrix Factorization (NMF)

- **Desvantagens:**

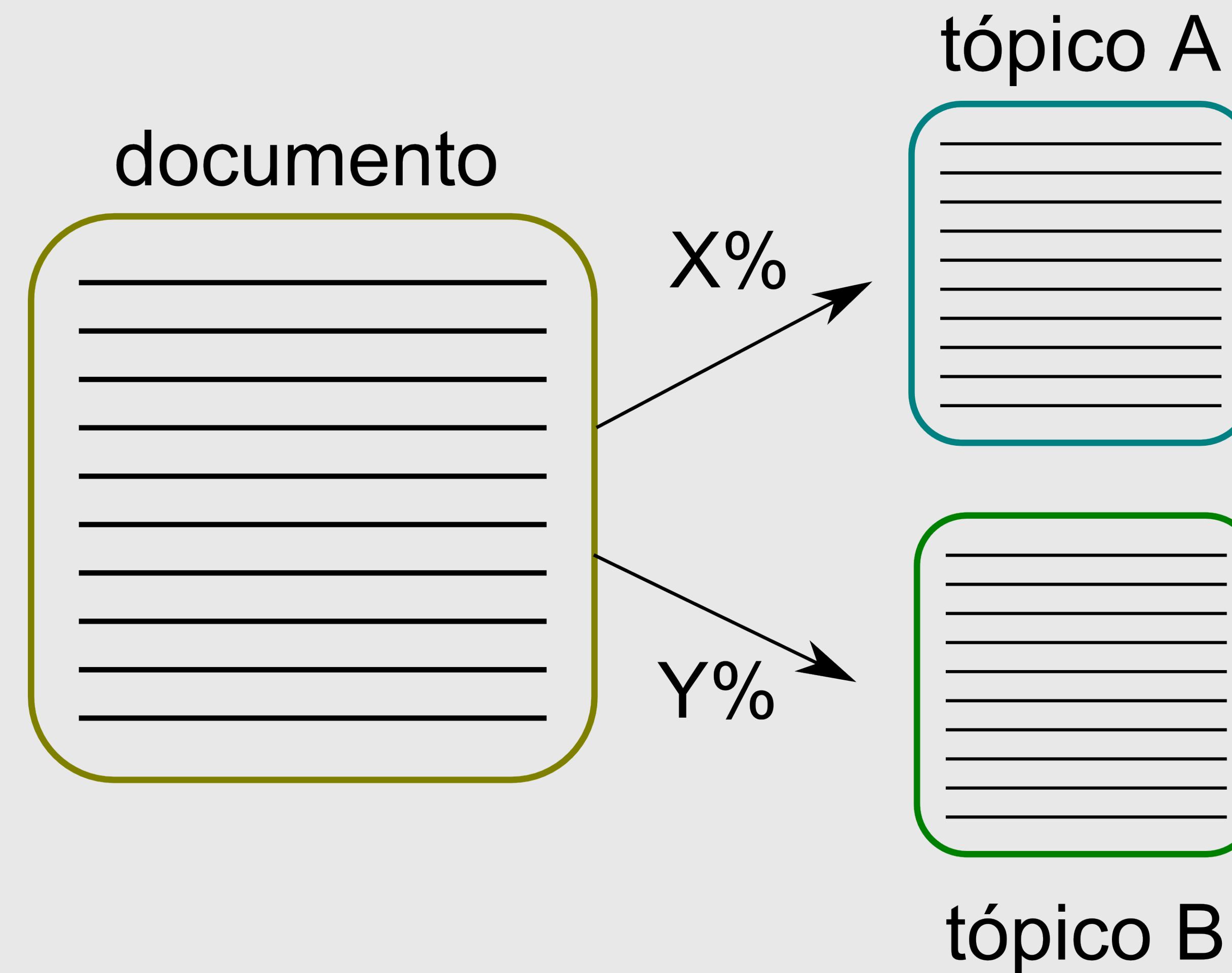
- Solução aproximada
- Pode causar overfitting
- Limitação de utilizar apenas features positivas

3. Latent Dirichlet Allocation (LDA)

- Método probabilístico
- Representa documentos como uma mistura de tópicos:
 - Cada tópico possui uma probabilidade associada dentro do documento
 - Cada palavra do documento possui uma probabilidade associada de pertencer a um tópico
- Precisa definir o número de tópicos (igual NMF)

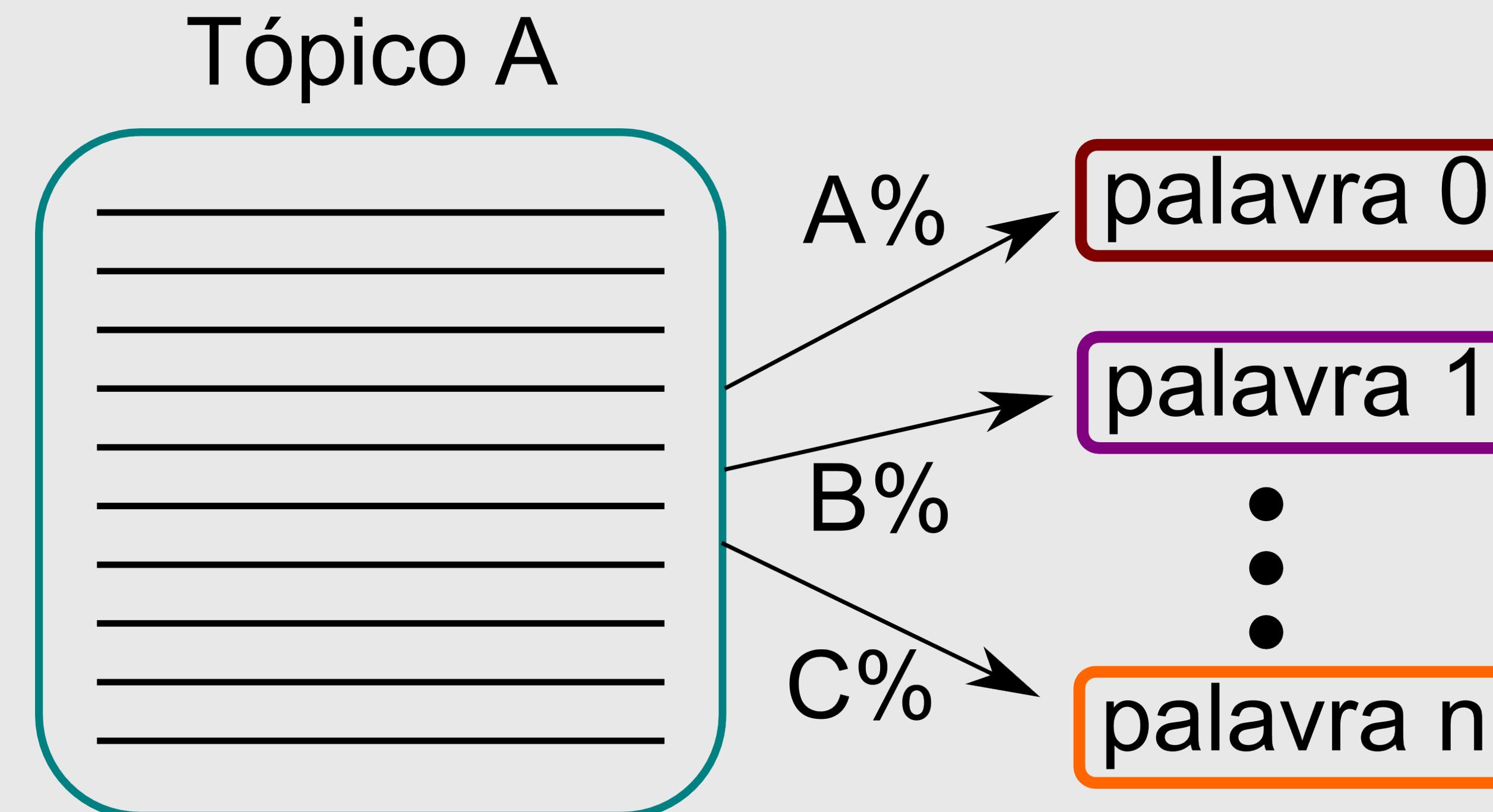
3. Latent Dirichlet Allocation (LDA)

- Documento → Místura de tópicos



3. Latent Dirichlet Allocation (LDA)

- Tópicos → Mistura de palavras



3. Latent Dirichlet Allocation (LDA)

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

I

3. Latent Dirichlet Allocation (LDA)

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

- Tópico A: Comida
- Tópico B: Animais

I

3. Latent Dirichlet Allocation (LDA)

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

- Documento 1: Apenas tópico A
- Documento 2: Apenas tópico B
- Documento 3: Mistura dos tópicos A e B

I

3. Latent Dirichlet Allocation (LDA)

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

- Qual o tópico associado a palavra “Fish” no documento 3?
 - $P('Fish' | \text{tópico A}) = 0.75 (3 - A, 1 - B)$
 - $P('Fish' | \text{tópico B}) = 0.25$

T

3. Latent Dirichlet Allocation (LDA)

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

- Qual a probabilidade de cada tópico no documento 3?
 - $P(\text{tópico A} \mid \text{Documento 3}) = P(\text{tópico B} \mid \text{Documento 3}) = 0.5$

I

3. Latent Dirichlet Allocation (LDA)

- Exemplo:

Document 1		Document 2		Document 3	
Eat	A	Cat	B	Cat	B
Fish	A	Dog	B	Eat	A
Vegetables	A	Pet	B	Fish	?
Fish	A	Pet	B	Cat	B
Eat	A	Fish	B	Fish	A

- Portanto, podemos concluir que “Fish” está contido no tópico A.

3. Latent Dirichlet Allocation (LDA)

- O método é repetido para todas as palavras múltiplas vezes
- O algoritmo para quando não houver mais variação
(convergência)
→ Método gerador

3. Latent Dirichlet Allocation (LDA)

- Método gerador:
 - Supõe que os documentos são gerados por um modelo probabilístico
 - Objetivo é aproximar esse modelo

3. Latent Dirichlet Allocation (LDA)

- Método gerador: Vantagens
 - Podemos amostrar a partir do modelo encontrado
 - Em outras palavras, podemos gerar novos documentos “articiais”

3. Latent Dirichlet Allocation (LDA)

- Exemplo: notebook

3. Latent Dirichlet Allocation (LDA)

- **Vantagens:**

- Tópicos são interpretáveis
- Possui maior generalização em comparação com o NMF
- Permite variação de tópicos e palavras (distribuição)
- Permite gerar documentos novos

3. Latent Dirichlet Allocation (LDA)

- **Desvantagens:**
 - Tópicos tendem a ser menos coerentes comparado ao NMF

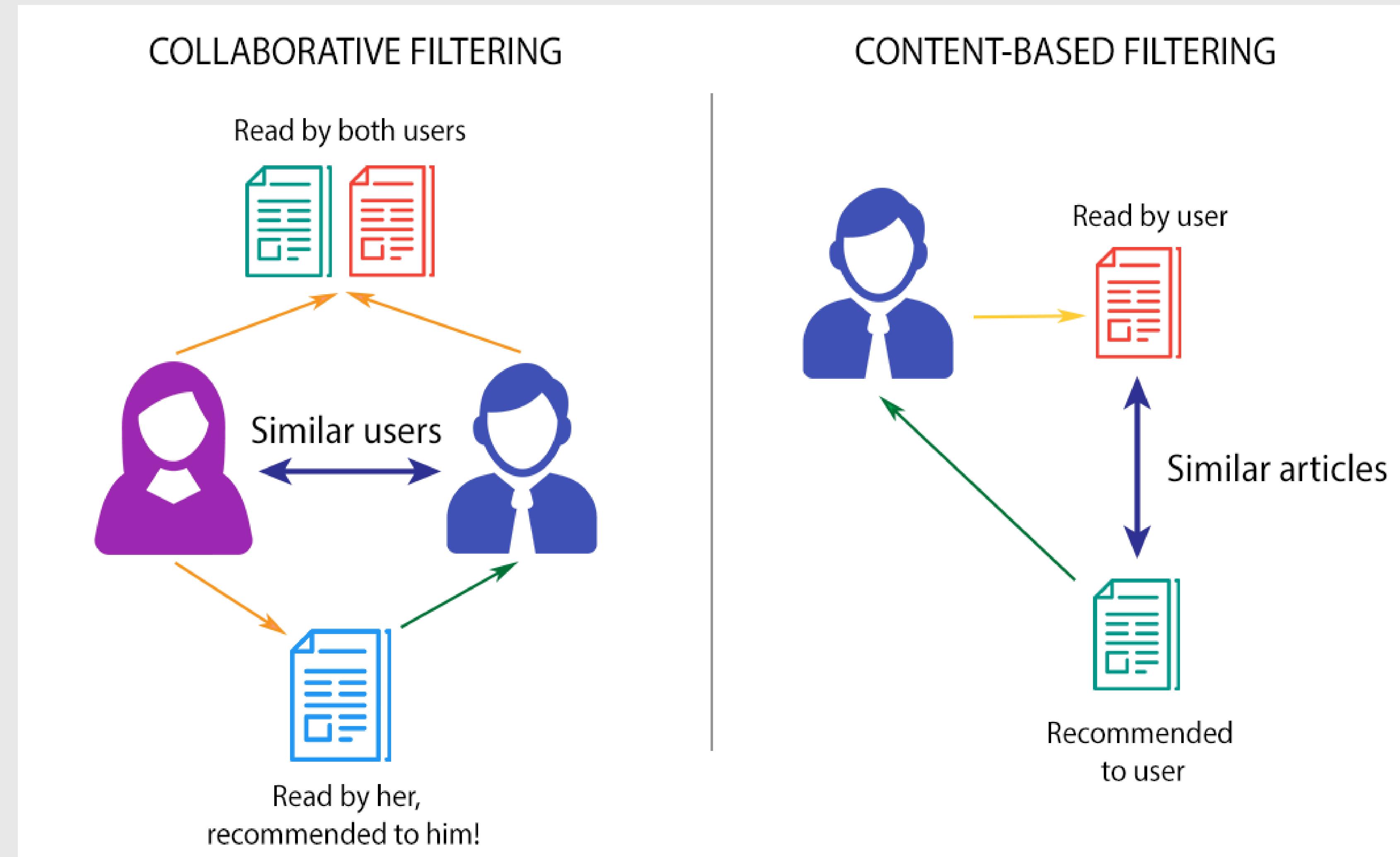
3. Latent Dirichlet Allocation (LDA)

- Exemplo: notebook

4. Sistemas de Recomendação

- Existem 2 grandes grupos:
 - Proximidade de documentos (produtos, músicas, filmes etc)
 - Proximidade entre usuários (filtro colaborativo)

4. Sistemas de Recomendação



4. Sistemas de Recomendação

- Proximidade de documentos:
 - Distância entre documentos
 - Similaridade de temas (tópicos)

4. Sistemas de Recomendação

- Proximidade de documentos:
 - Distância entre documentos
 - Similaridade de temas (tópicos)



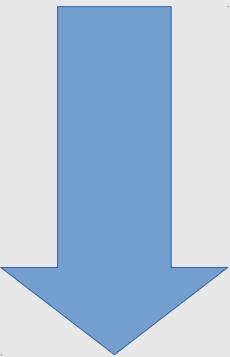
**Clustering
Topic Analysis**

4. Sistemas de Recomendação

- Proximidade entre usuários (filtro colaborativo):
 - Usuários semelhantes consomem documentos semelhantes

4. Sistemas de Recomendação

- Proximidade entre usuários (filtro colaborativo):
 - Usuários semelhantes consomem documentos semelhantes



**Clustering
Topic Analysis**

4. Sistemas de Recomendação

- Proximidade entre usuários (filtro colaborativo):
 - **Documentos:**
 - Histórico de consumo do usuário (compra, avaliação, leitura etc)
 - **Atributos / Features:**
 - Lista de itens de consumo (produtos, livros, músicas, filmes etc)

I 4. Sistemas de Recomendação

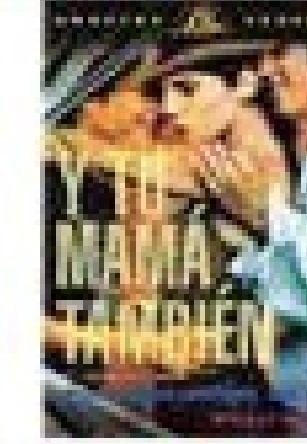
- Exemplo: Recomendação de filmes

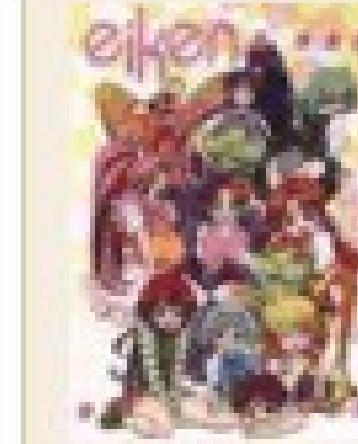
[Close](#)

Other Movies You Might Enjoy

[Amelie](#)

[Add](#)
★ ★ ★ ★ Not Interested

[Y Tu Mama Tambien](#)

[Add](#)
★ ★ ★ ★ Not Interested


Eiken has been added to your Queue at position 2.
This movie is available now.
[Move To Top Of My Queue](#)

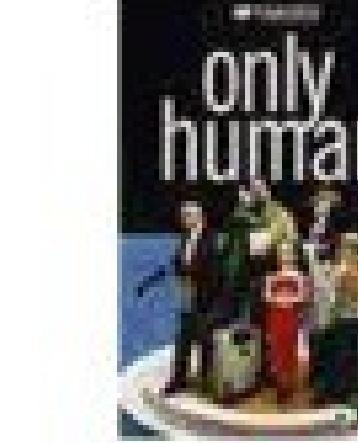
[< Continue Browsing](#) [Visit your Queue >](#)

[Guys and Balls](#)

[Add](#)
★ ★ ★ ★ Not Interested

[Mostly Martha](#)

[Add](#)
★ ★ ★ ★ Not Interested

[Only Human](#)

[Add](#)
★ ★ ★ ★ Not Interested

[Russian Dolls](#)

[Add](#)
★ ★ ★ ★ Not Interested

[Close](#)

I 4. Sistemas de Recomendação

- Exemplo: Recomendação de filmes



T

OBRIGADO!