

# Sistemas de Recomendação

## Recomendadores Content-based

*Topic Modeling*



**PUC Minas**

Pós-Graduação Lato Sensu

**Nícollas Silva**

**Premissa:** as características dos itens revelam quais as características os usuários mais se interessam.

- Correlaciona os itens do domínio por meio de suas features.
  - Filmes: gêneros, atores, duração, ...
  - Músicas: categoria, autor, lançamento, ...
  - Carro: marca, cor, ano, ...

Ao aplicar estratégias de CB, devemos:

- Modelar os itens sobre suas características
- Modelar os usuários como um vetor agregado dessas características
- Computar a similaridade entre esses dois componentes

# O custo da dimensionalidade

- Representação no espaço vetorial:
  - Cada item é um vetor de componentes.
  - Cada usuário é uma combinação dos vetores dos itens.

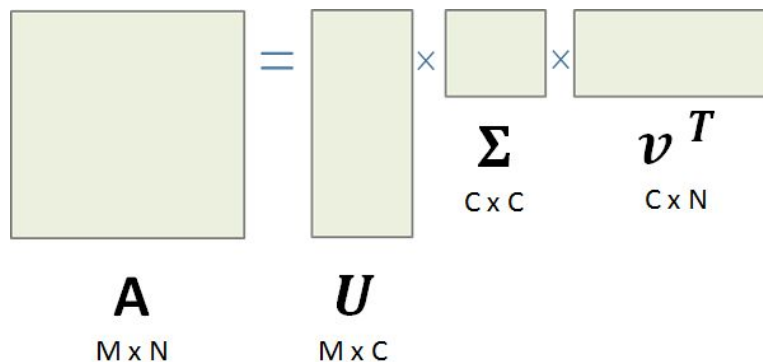
- **Problema:** O espaço altamente dimensional.
  - **Eficácia:** vai demorar para computar as similaridades.
  - **Eficiência:** vai ser difícil encontrar conceitos similares.

- Google Web N-grams  
[Franz and Brants, 2006]

# tokens	1,024,908,267,229
# sentences	95,119,665,584
# 1-grams	13,588,391
# 2-grams	314,843,401
# 3-grams	977,069,902
# 4-grams	1,313,818,354
# 5-grams	1,176,470,663

# O custo da dimensionalidade

- A redução de dimensionalidade é muito eficaz em SsR.
  - Métodos oriundos da Álgebra Linear e Estatística.
  - Modelam usuários e itens em um espaço semântico.


$$\begin{array}{c} \boxed{\phantom{A}} \\ \mathbf{A} \\ M \times N \end{array} = \begin{array}{c} \boxed{\phantom{U}} \\ \mathbf{U} \\ M \times C \end{array} \times \begin{array}{c} \boxed{\phantom{\Sigma}} \\ \mathbf{\Sigma} \\ C \times C \end{array} \times \begin{array}{c} \boxed{\phantom{v^T}} \\ \mathbf{v}^T \\ C \times N \end{array}$$

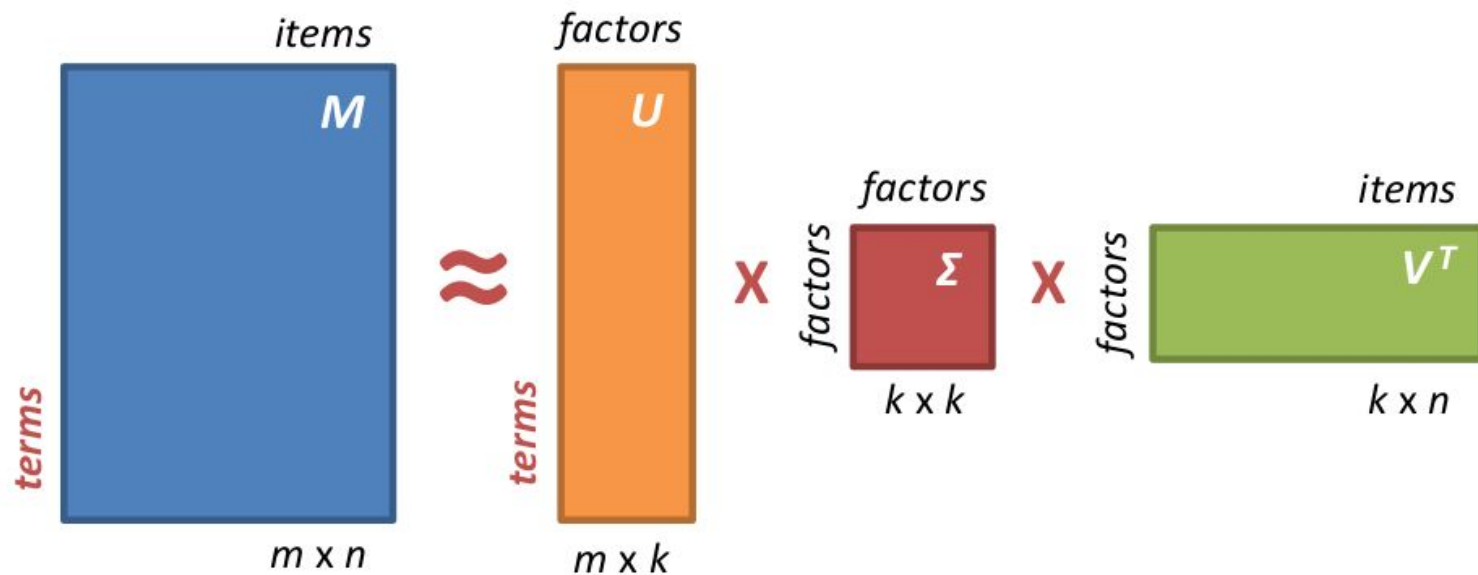
# Modelagem de Tópicos

- Os tópicos podem ou não ter um significado semântico.
  - São capazes de agregar termos similares.
  - Efetivos para métodos de filtragem onde buscamos por uma palavra-chave e encontramos diversos documentos relacionados.

Topic 1		Topic 2		Topic 3	
term	weight	term	weight	term	weight
biology	45692	space	67019	politics	24763
university	10576	nasa	9673	washington	11982
moth	5304	earth	5674	congress	7261
caterpillar	4927	moon	3455	president	5820

# Latent Semantic Analysis - LSA

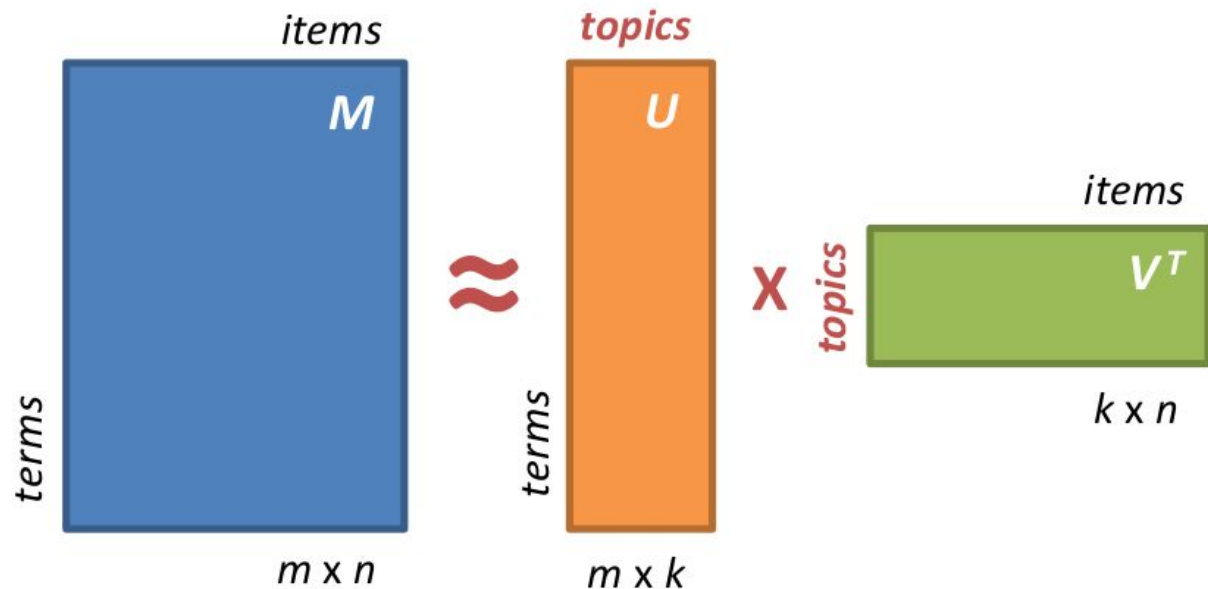
## Content-based filtering



Fonte: Slides da disciplina de Sistemas de Recomendação do prof. Rodrygo Santos - UFMG - 2016/2

# Modelagem de Tópicos

## Content-based filtering



Fonte: Slides da disciplina de Sistemas de Recomendação do prof. Rodrygo Santos - UFMG - 2016/2



# Latent Dirichlet Allocation - LDA

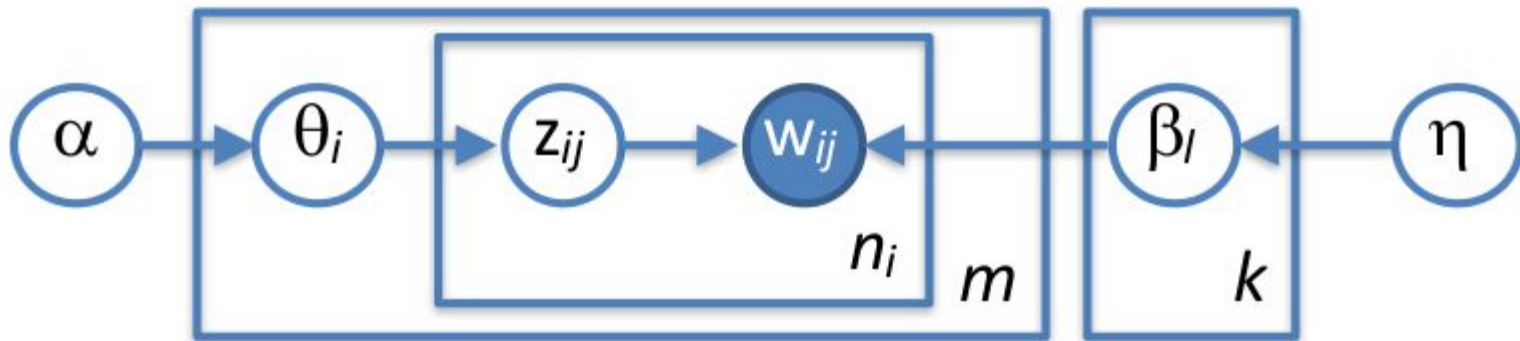
- Assume que conjuntos de observações podem ser explicados por grupos *não-observáveis* que agrupa os similares.
  - Um documento é uma mistura de alguns tópicos.
  - Cada palavra é referente a um desses tópicos.
- LDA é um modelo quase idêntico ao LSA
  - Assume que os tópicos derivam da distribuição de Dirichlet.
  - O espaço de Dirichlet assume a esparsidade do modelo.

# Latent Dirichlet Allocation - LDA

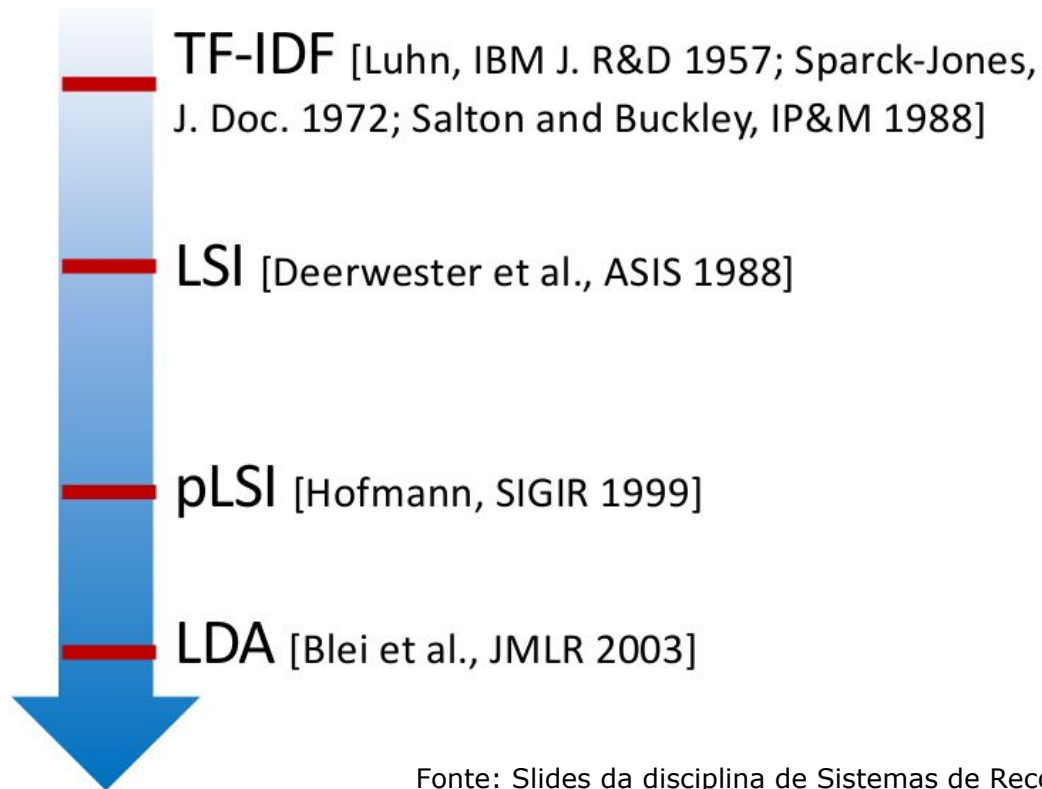


PUC Minas

- Na notação *plate*, o LDA é comumente representado por:
  - Entidades (documentos, termos e tópicos) que se repetem.
  - Variáveis que controlam a escolha desses elementos.



# Redução da dimensionalidade



Fonte: Slides da disciplina de Sistemas de Recomendação do prof. Rodrygo Santos - UFMG - 2016/2

- Computar similaridades entre os tópicos extraídos
  - Usuários e itens são representados sobre o mesmo espaço latente.
  - O espaço dimensional é muito menor que o original.
- Modelos estado-da-arte para filtragem baseada em conteúdo
  - Essas abordagens superam os modelos apresentados anteriormente.
  - Elas são eficientes e eficazes.