FACULDADE IBRATEC

Descoberta de Tópicos

Prof. Marcelo Pita

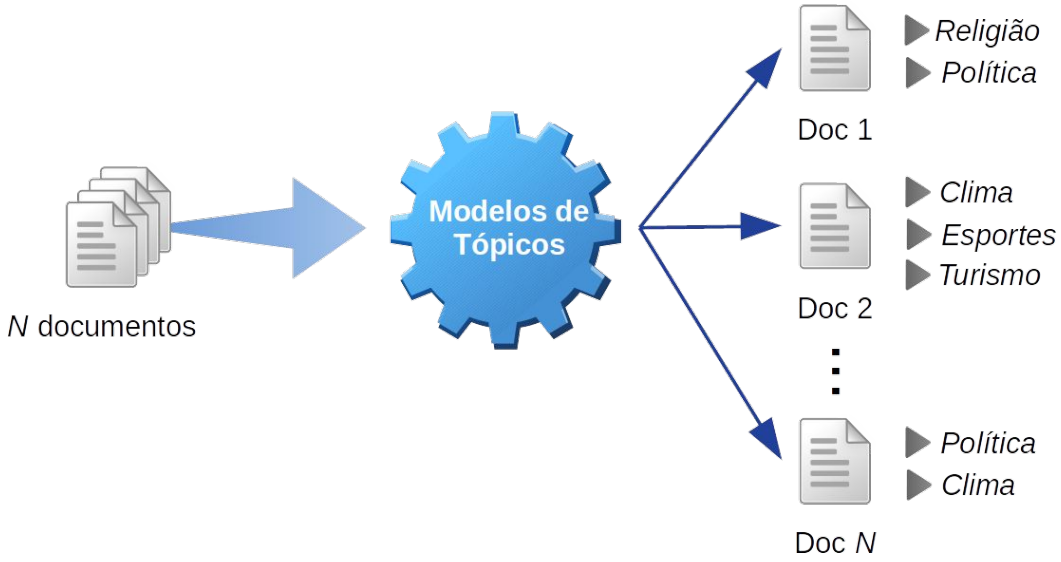
Pós em Inteligência Artificial

Processamento de Linguagem Natural

Descoberta de tópicos em texto

O que é um **modelo de tópicos**?

Modelo estatístico que visa descobrir “tópicos” em documentos de texto.



Descoberta de tópicos em texto

Para que serve um modelo de tópicos?

Organizar coleções de documentos automaticamente.

Exemplos de coleções:

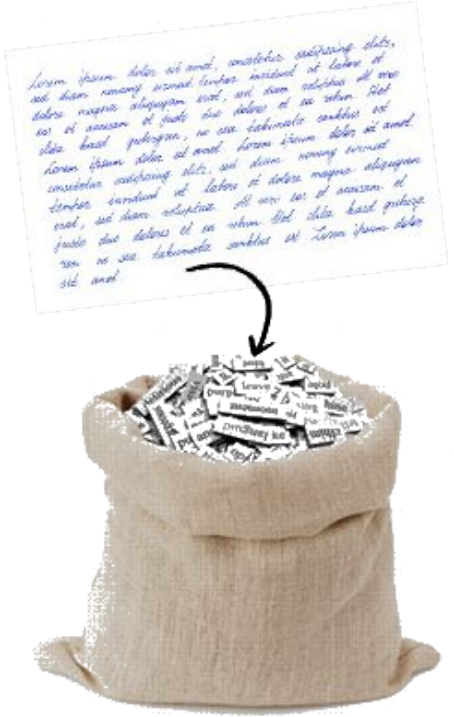
- Documento jurídicos em um tribunal
- Notícias de um jornal
- Artigos científicos de um periódico

Descoberta de tópicos em texto

Latent Dirichlet Allocation

Rede Bayesiana para descoberta de tópicos não conhecidos *a priori*.

Documentos de texto são considerados “sacos de palavras” (*bag of words*)

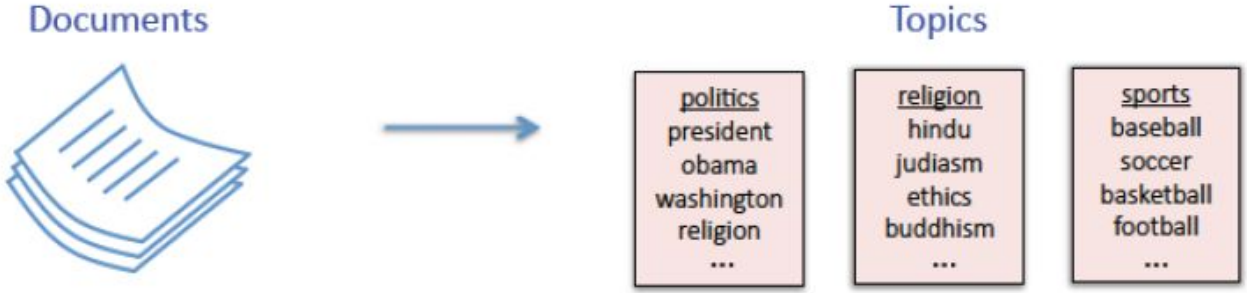


Descoberta de tópicos em texto

Latent Dirichlet Allocation

Passo 1:

Treinamento



Passo 2:

Inferência

Descoberta de tópicos em texto

Latent Dirichlet Allocation

Após o treinamento do modelo, tópicos são caracterizados como **distribuições de probabilidade sobre palavras**.

- Em um tópico sobre futebol, as palavras mais prováveis podem ser “gol”, “atacante”, “zagueiro”, “futebol”, “falta”.

Documentos são caracterizados como **distribuições de probabilidade sobre tópicos**.

Descoberta de tópicos em texto

Distribuição de probabilidade de palavras para os tópicos sobre futebol e automobilismo.

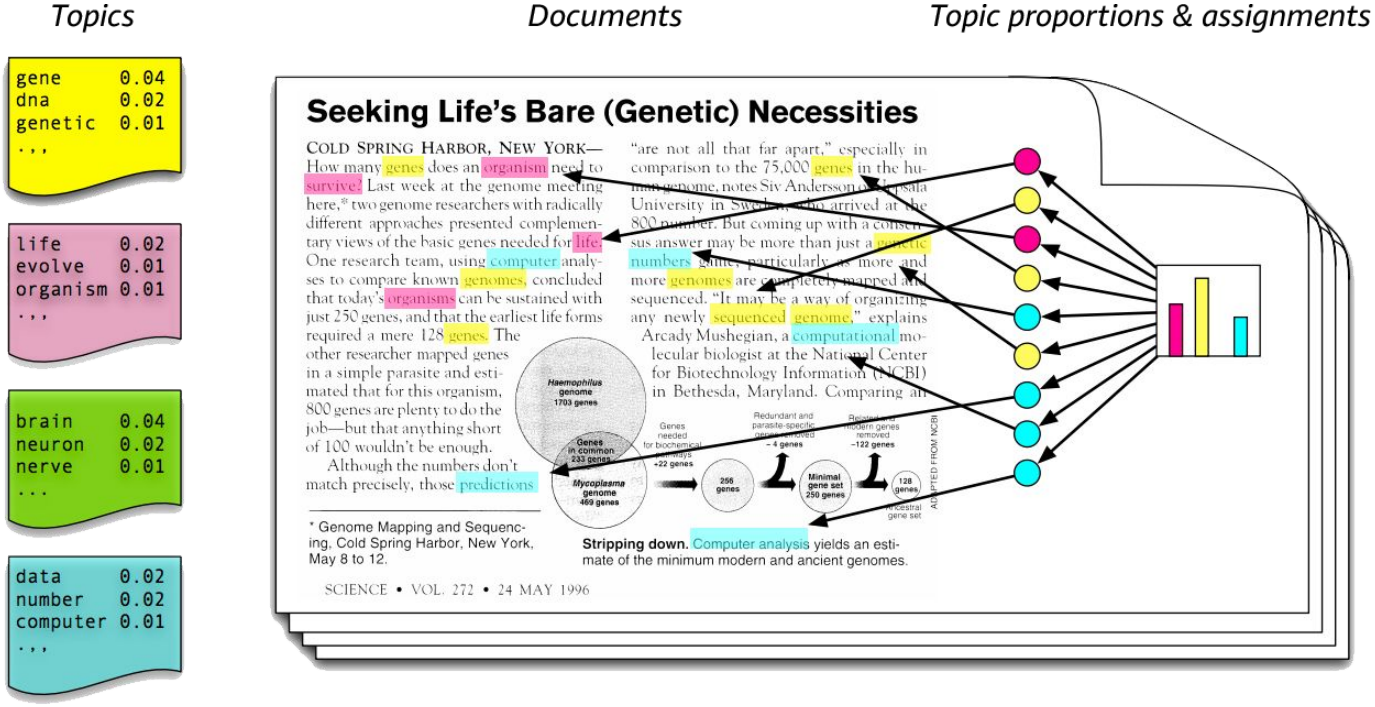
FUTEBOL

PALAVRA	Probab
Gol	0.10
Neymar	0.001
Treinador	0.05
Vitoria	0.03
Empate	0.03
Campeão	0.01
Atacante	0.02
Zagueiro	0.01
Falta	0.02
Pênalti	0.01
....	....
Velocidade	0.001
Circuito	0.0001
Freio	0.0001
Motor	0.0000
Massa	0.0000
.....	....
Total	1

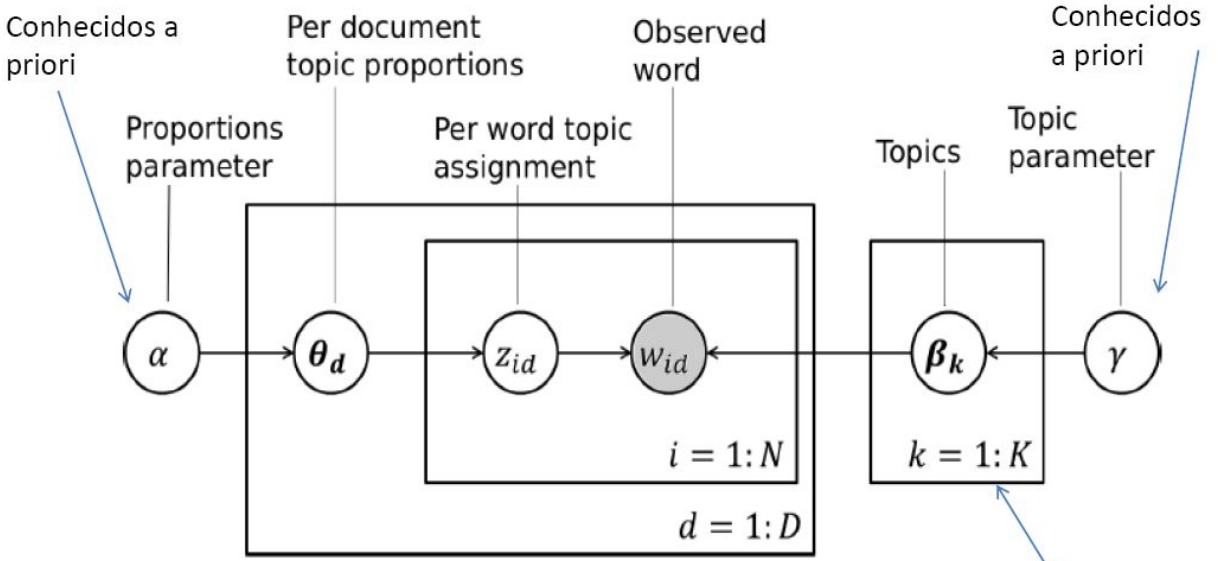
AUTOMOBILISMO

PALAVRA	Probab
Gol	0.0000
Neymar	0.0000
Treinador	0.001
Vitoria	0.05
Empate	0.01
Campeão	0.04
Atacante	0.0000
Zagueiro	0.0000
Falta	0.0000
Pênalti	0.0000
....	
Velocidade	0.15
Circuito	0.07
Freio	0.04
Motor	0.08
Massa	0.04
.....	....
Total	1

Descoberta de tópicos em texto



Descoberta de tópicos em texto



- Nodes are random variables; edges indicate dependence.
- Shaded nodes indicate *observed* variables.

K vetores M-dim
Cada um = dist de probab sobre as M palavras do dicion.

Descoberta de tópicos em texto

<Prática>