



Descoberta de Tópicos

Prof. Marcelo Pita

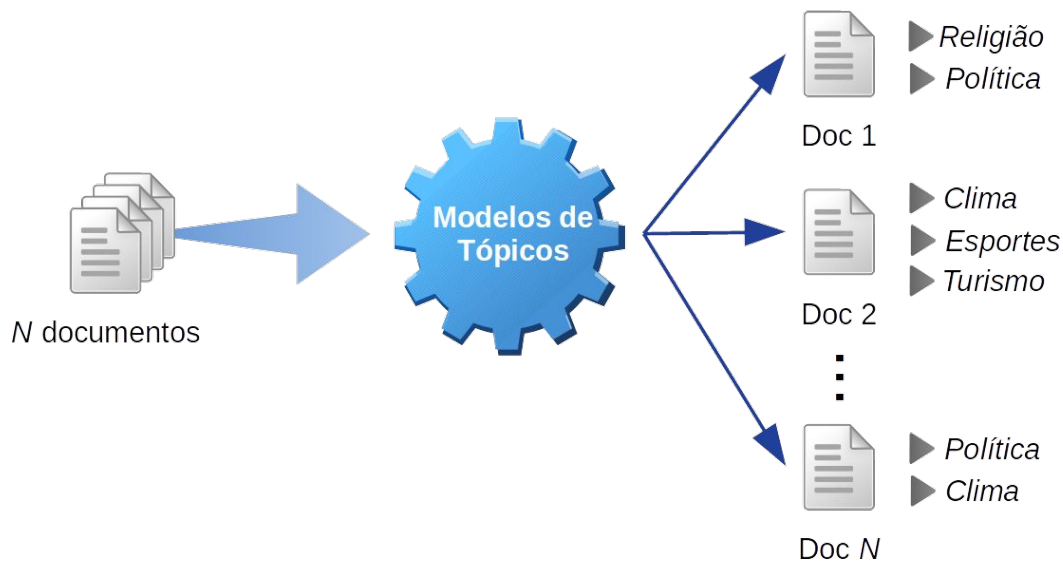
Pós em Inteligência Artificial
Processamento de Linguagem Natural



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*.

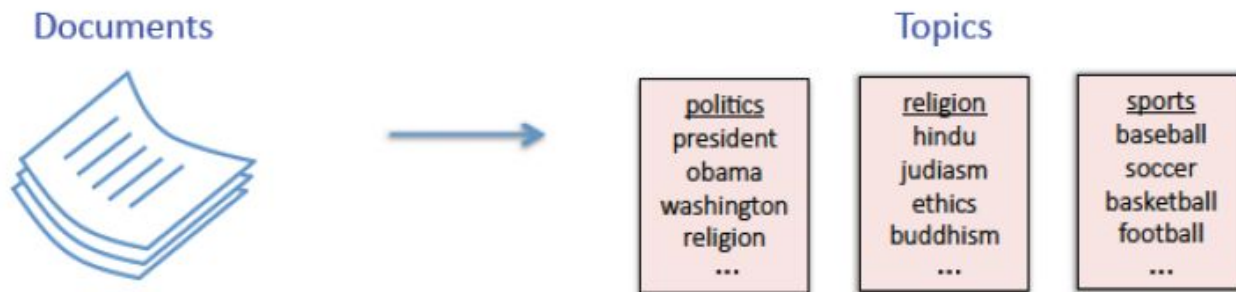
Documentos de texto são considerados “sacos de palavras” (*bag of words*)



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.

<u>FUTEBOL</u>	
<u>PALAVRA</u>	<u>Probab</u>
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

<u>AUTOMOBILISMO</u>	
<u>PALAVRA</u>	<u>Probab</u>
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

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

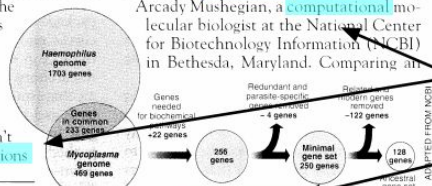
Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

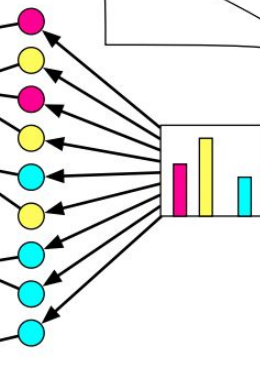
"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic** numbers game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



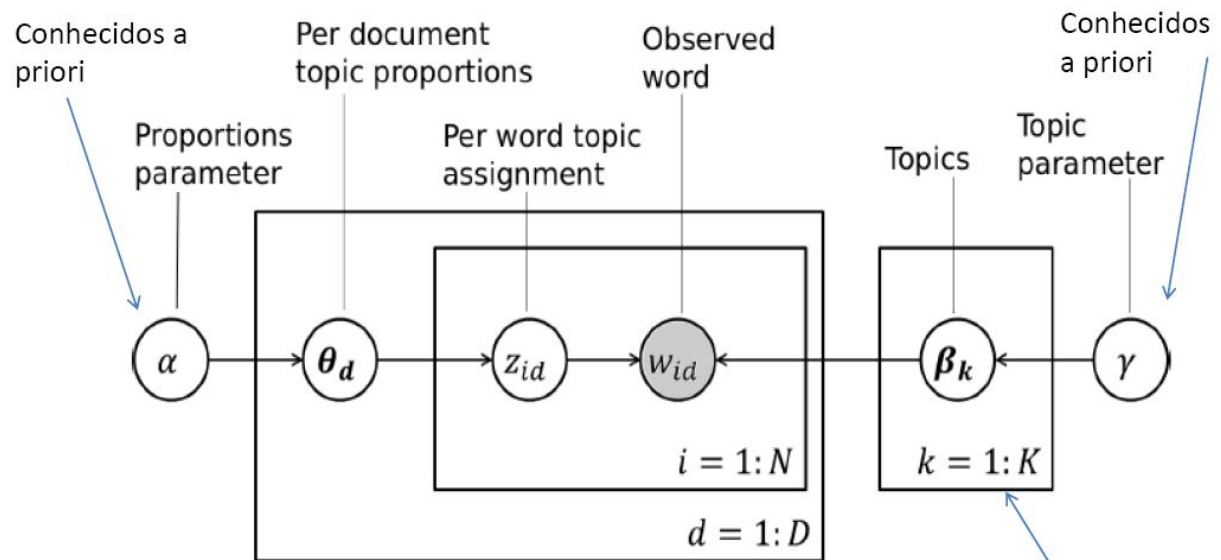
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions & assignments



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>