PROCESSAMENTO DE LINGUAGEM NATURAL

Modelagem de tópicos

TÓPICOS

1. Introdução à Modelagem de Tópicos
2. Modelos probabilísticos (LDA)
3. Avaliação

MOTIVAÇÃO

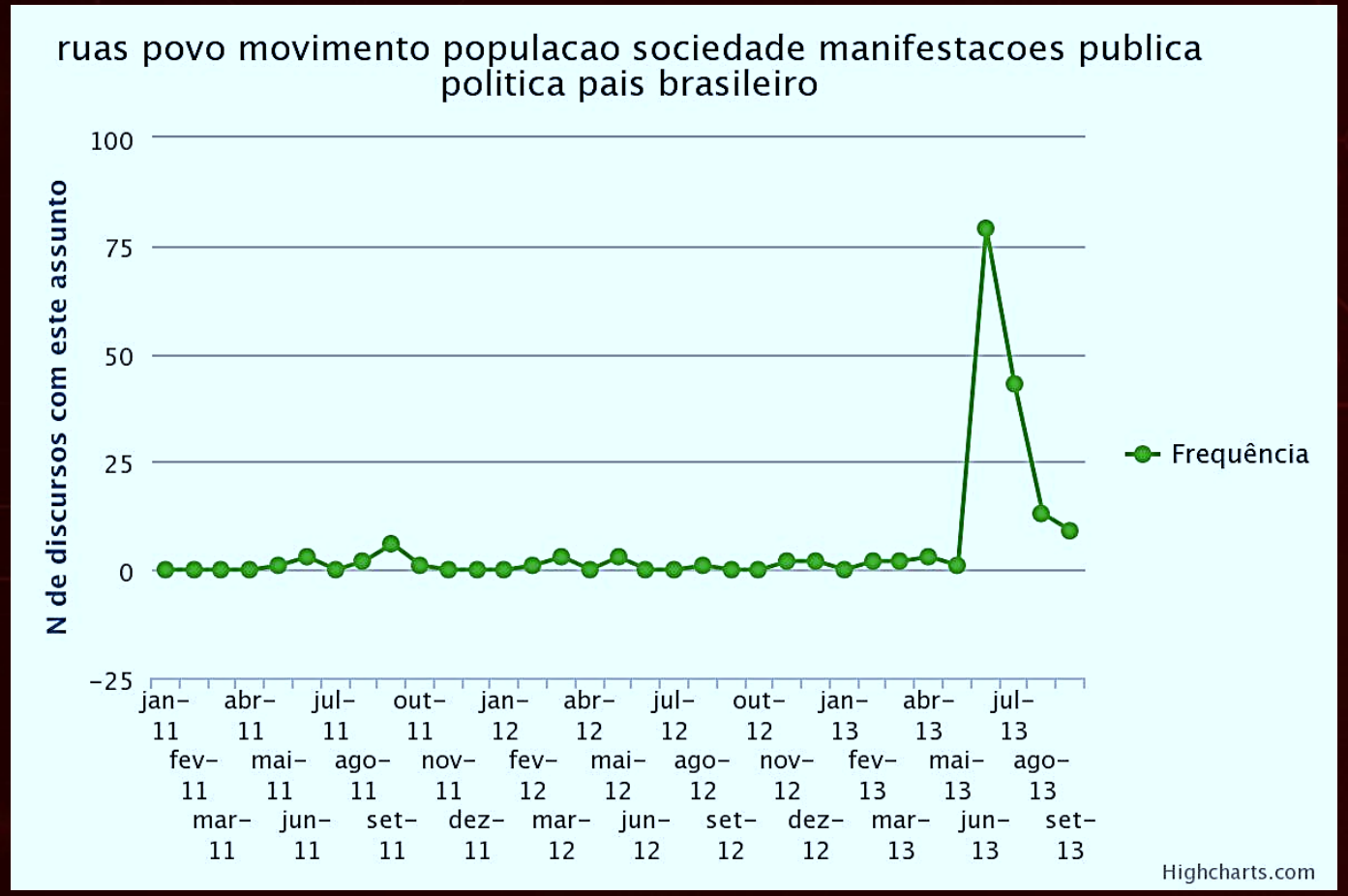
- Modelos de tópicos provêm métodos para automaticamente organizar, entender, buscar e sumarizar uma grande coleção de documentos.
- Descobrir padrões de tópicos ocultos que pervadem uma coleção de documentos textuais.
- Anotar documentos de acordo com seus tópicos.
- Usar as anotações para organizar, sumarizar e auxiliar na busca.

TÓPICOS DESCOBERTOS

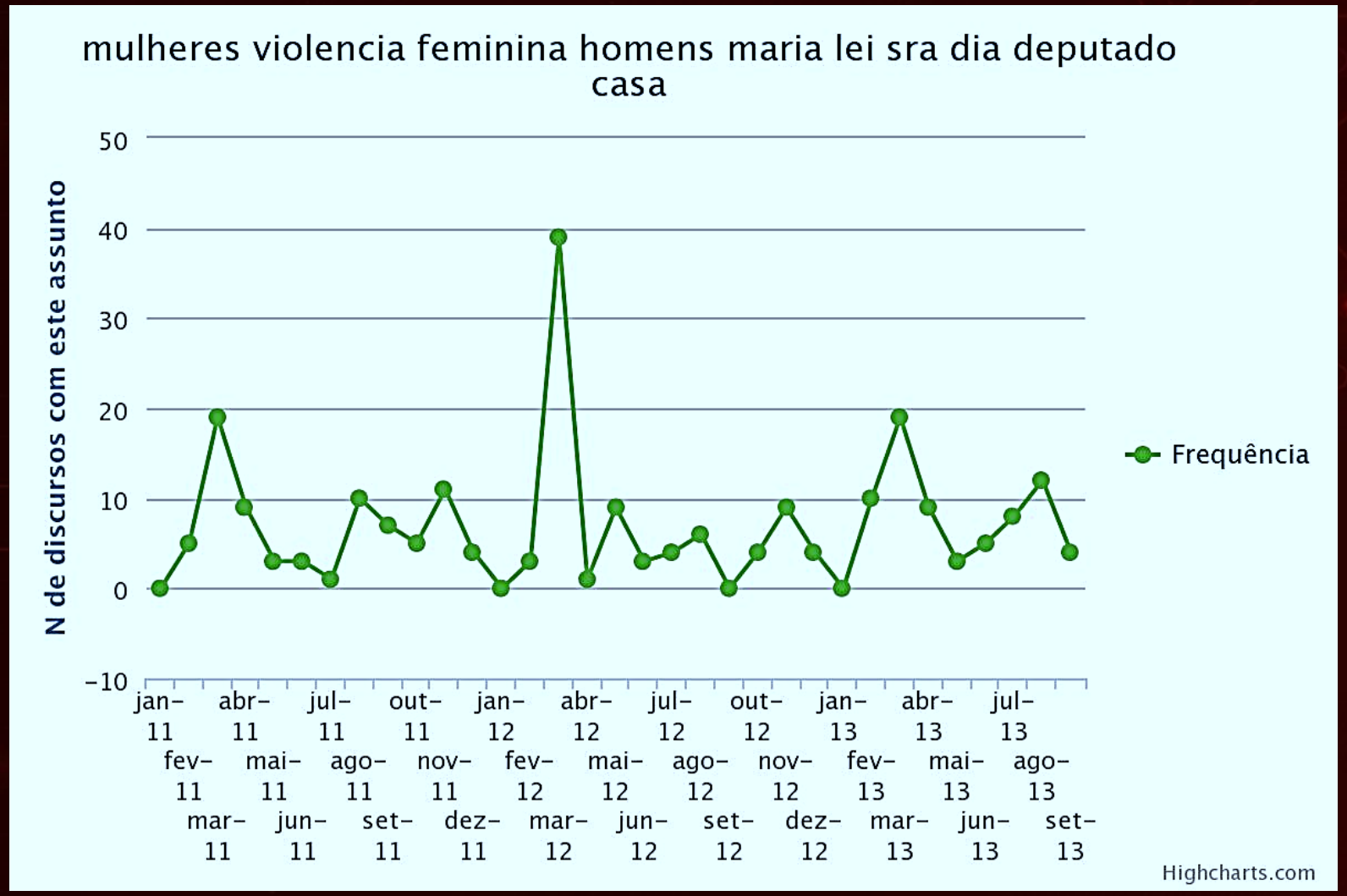
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Distribuição de tópicos sobre palavras [Blei, 2011]

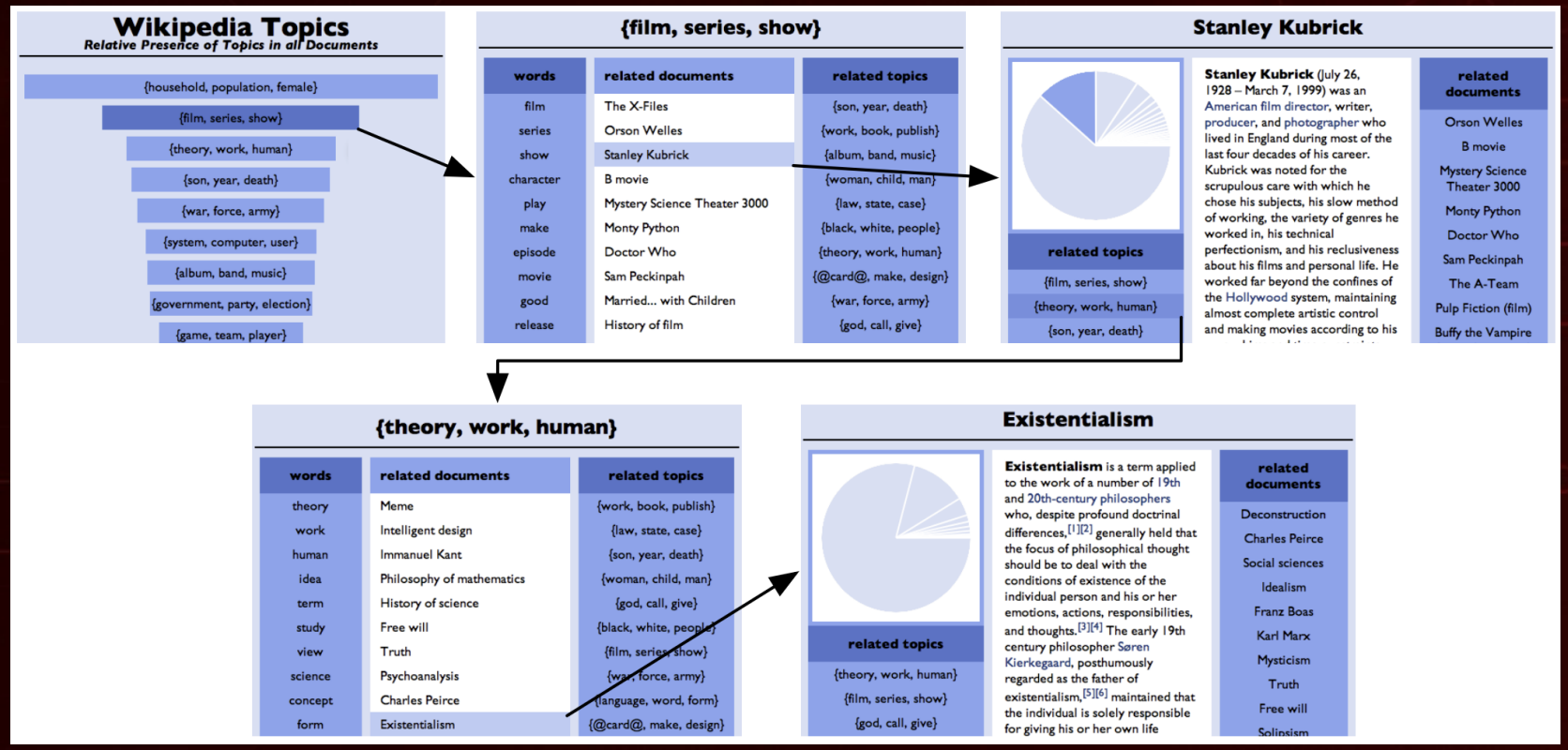
TÓPICOS DE DISCURSOS POLÍTICOS



TÓPICOS DE DISCURSOS POLÍTICOS



DOCUMENTOS E TÓPICOS



TÓPICOS

1. Introdução à Modelagem de Tópicos
2. Modelos probabilísticos (LDA)
3. Avaliação

MODELOS PROBABILÍSTICOS

- Modelos probabilísticos de tópicos são conjunto de algoritmos cujo o objetivo é descobrir estruturas temáticas em grandes coleções de documentos
- Representar uma coleção de documentos pelos tópicos é uma forma de reduzir o conjunto de descritores da coleção
- *Probabilistic Latent Semantic Indexing* (pLSI)
- *Latent Dirichlet Allocation* (LDA)
- *Hierarquical Dirichlet Process* (HDP)
- O LDA é o modelo base

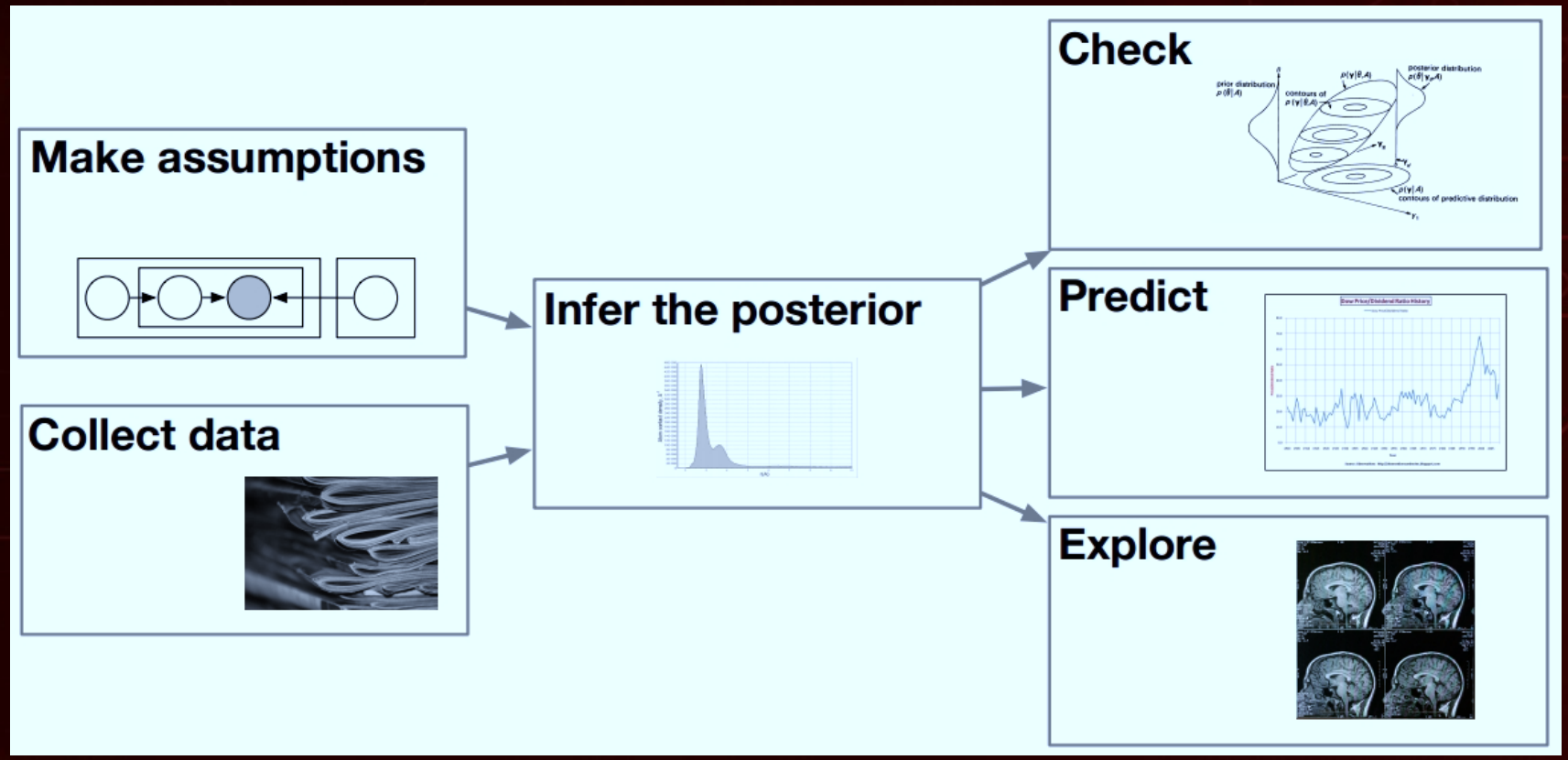
LDA

- *Latent Dirichlet Allocation* (LDA) é o método padrão para modelagem de tópicos
- Descrito por Blei et. al (2003):
  - https://www.seas.harvard.edu/courses/cs281/papers/blei-ng-jordan-2003.pdf
- Assume um modelo generativo:
  - Cada documento é uma mistura de tópicos
  - Cada tópico é uma mistura de termos
- *Correlated topic model* (CTM) é uma extensão do LDA

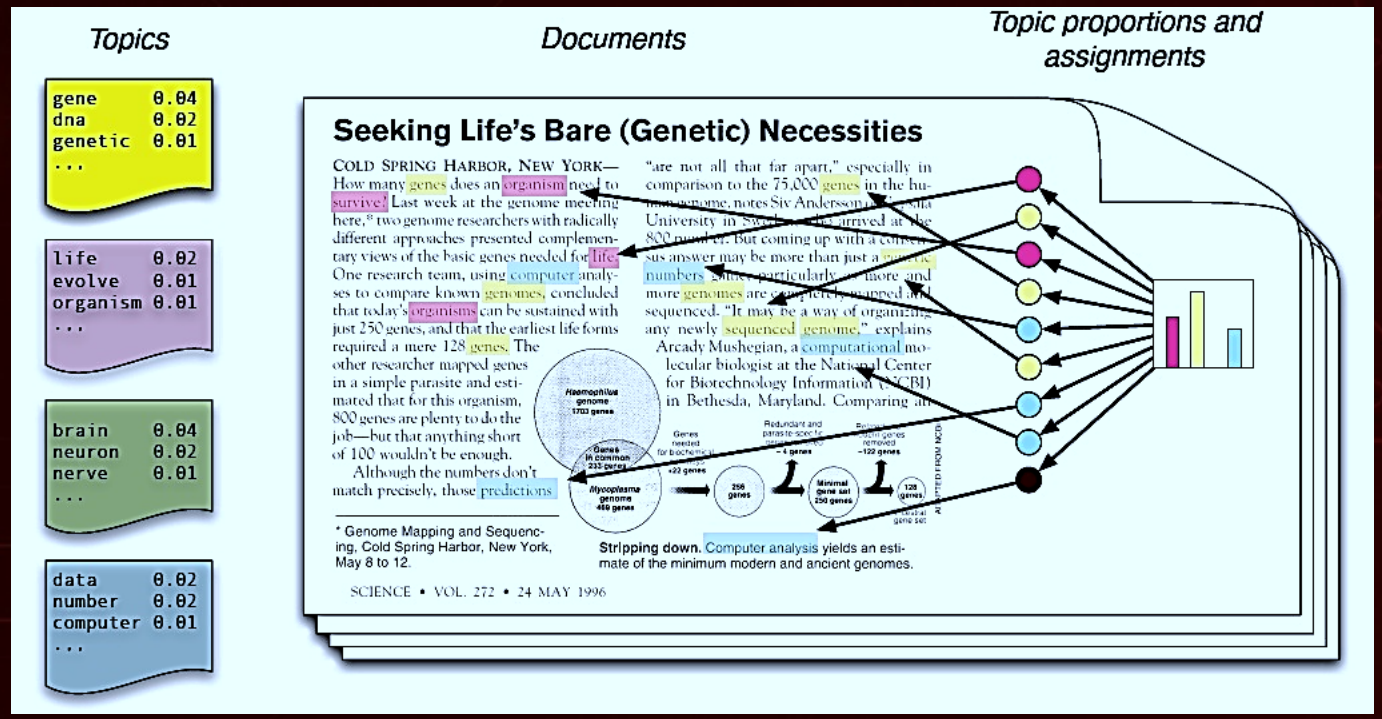
FUNCIONAMENTO

- Segundo o LDA, o *corpus* é resultado de um processo generativo.
- Cada documento é uma mistura de *K* assuntos.
- Cada assunto possui uma distribuição de probabilidade para os *V* termos do vocabulário.
- Os tópicos são distribuições de probabilidade sobre o amplo vocabulário hipotético.
- Com esse modelo, em hipótese, se geram os documentos, caso fossem conhecidos os parâmetros.
- É um modelo baseado em probabilidade condicional.

MODELOS PROBABILÍSTICOS

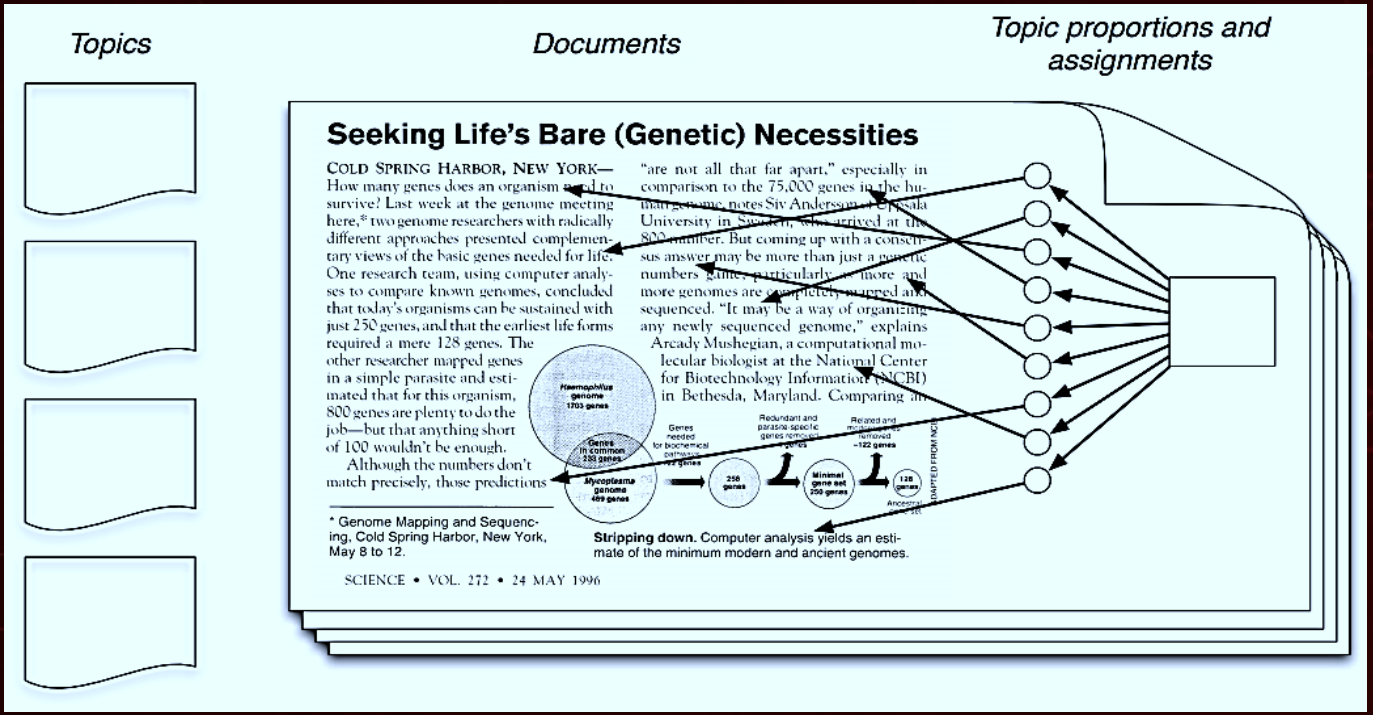


MODELO LDA



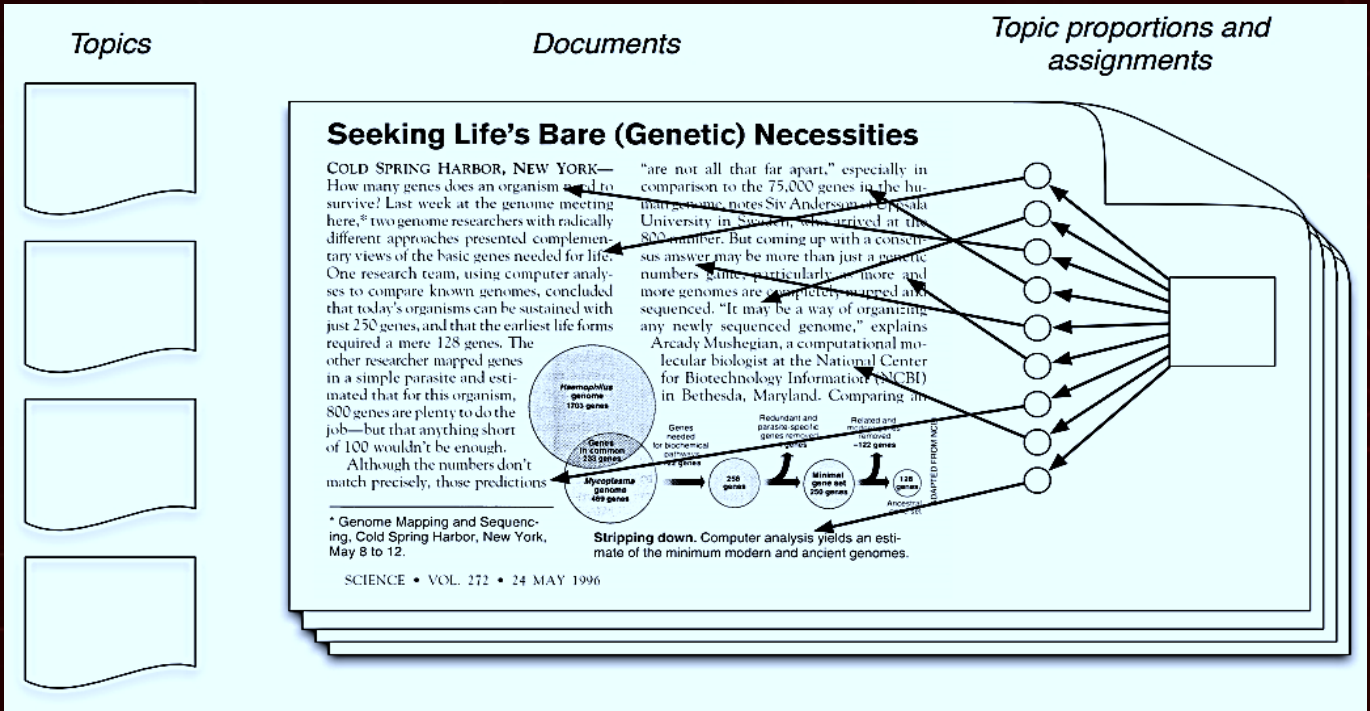
- Cada tópico é uma distribuição de palavras
- Cada documento é uma mistura de tópicos
- Cada palavra é uma amostra de um desses tópicos

MODELO LDA



- Na realidade, observamos apenas os documentos
- As outras variáveis são latentes

MODELO LDA



- **Nosso objetivo é inferir essas variáveis latentes**
- **Computar a distribuição condicionada aos documentos**

$$P(topicos\_palavras, documentos\_tópicos, atribuições \mid documentos)$$

PROCESSO GENERATIVO

- Processo imaginário que descreve como os documentos são criados.
- Formalmente, define um tópico como uma distribuição sobre o vocabulário de palavras.
- Assume que os tópicos são especificados antes de qualquer dado ser gerado.
- Para cada documento da coleção, escolhemos uma palavra da seguinte forma:
    1. Aleatoriamente escolhe uma distribuição sobre os tópicos.
    2. Para cada palavra no documento:
        1. Aleatoriamente escolhe um tópico da distribuição de tópicos.
        2. Aleatoriamente escolhe uma palavra da correspondente distribuição sobre o vocabulário.

MODELO LDA

- Passo 1: aleatoriamente reatribuir todos os z_{iw} levando em conta
  - Proporções dos tópicos *j* no documento *i*
  - Distribuições de cada palavra *w* do vocabulário no tópico
- Passo 2: aleatoriamente reatribuir as distribuições de tópicos em cada documento *i* a partir dos novos valores de z_{iw}
- Passo 3: Repetir o procedimento para todos os documentos
- Passo 4: aleatoriamente reatribuir as distribuições das palavras por tópicos a partir de todos os z_{iw} do *corpus*

$$\Pi_i = [\Pi_{1i}\ \Pi_{i2}, \ldots, \Pi_{ik}],$$

TÓPICOS

1. Introdução à Modelagem de Tópicos
2. Modelos probabilísticos (LDA)
3. Avaliação

AVALIANDO MODELOS DE TÓPICOS

- A forma mais comum de avaliar os modelos probabilísticos de tópicos é calculando o logaritmo da verossimilhança do modelo
- Outra métrica é o cálculo da perplexidade do modelo

$$perplexidade(w) = \exp\left(-\frac{\log p(wZ|\alpha,\beta)}{\log \sum_{j=1}^{m} n_{d_j}}\right)$$

- Bom para comparar modelos probabilísticos, entretanto, os valores obtidos não necessariamente condizem com a correta relação entre os tópicos encontrados e os assuntos descritos na coleção

AVALIANDO MODELOS DE TÓPICOS

- No trabalho de (Newman, 2010) foi proposto um método automático baseado na informação mútua entre pares de palavras que formam o tópico
- Uma coleção de referência é utilizada para calcular a coocorrência entre os termos relacionados
- Quanto maior a similaridade média entre os pares das palavras, mais coerente é o tópico

CONCLUSÕES

- Modelos probabilísticos oferecem ferramentas eficientes para organizar, buscar e entender uma vasta quantidade de informações
- Um *framework* que abre possibilidade de expansão!
- Modelos Probabilísticos de tópicos têm um tratamento matemático rigoroso
- Na perspectiva de um desenvolvedor de aplicações práticas, criar um modelo generativo e derivá-lo, a fim de obter um algoritmo de inferência implementável, é uma tarefa difícil

O QUE VIMOS?

- Introdução
- Modelos de linguagem
- N-gramas

PRÓXIMA VIDEOAULA

- ***Embeddings***

REFERÊNCIAS

- Seminário sobre Modelos Probabilísticos de Tópicos
  - Prof. Thiago P. Faleiros (UnB)
- Curso de Processamento de Linguagem Natural
  - Profa. Helena Caseli (UFSCar)
- Curso de Processamento de Linguagem Nat
  - Prof. Thiago Pardo (ICMC-USP)