



universidade de aveiro
theoria poiesis praxis

Introdução à Ciência de Dados

João Lourenço Marques: jjmarques@ua.pt

Processamento de linguagem natural (NLP)

- Modelação -



Data Science

Programa de "festas"!

Data Science/Ciência de Dados para Ciências Sociais

NLP /Modelação

1. NLP – conceitos fundamenais

2. Modelação

2.1.. Clusterização

2.2. Modelação de Tópicos

LDA

LSA

} Fatorização

Quando aparecer o símbolo “#” tem hiperlink



Data Science

NLP – conceitos
fundamentais

Data Science/Ciência de Dados

1

NLP – Conceitos Fundamentais

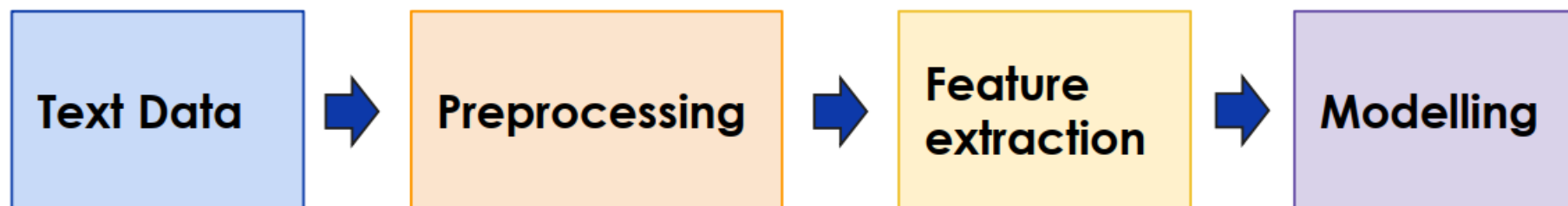


Data Science

NLP – Grandes etapas

NLP envolve as seguintes etapas:

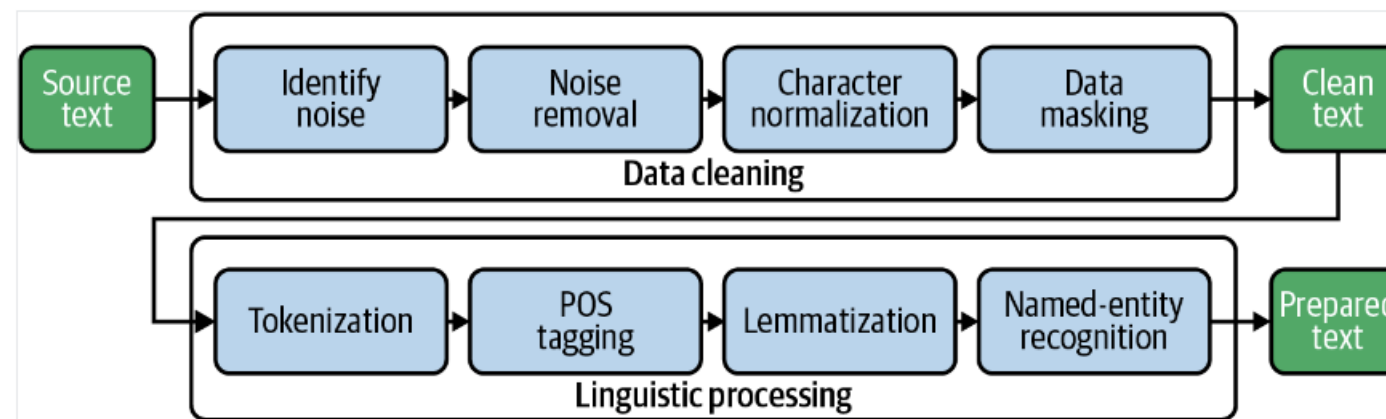
- Pré-processamento
- Vetorização
- Modelação





Data Science

NLP – Grandes etapas





Conceito fundamental do Processamento de Linguagem Natural (NLP)

- Vetorização de texto ...transformação de dados textuais em representações numéricas

permitindo que máquinas compreendam e processem a linguagem humana



Data Science

NLP – Grandes etapas



Técnicas de vectorização de texto - Alguns exemplos

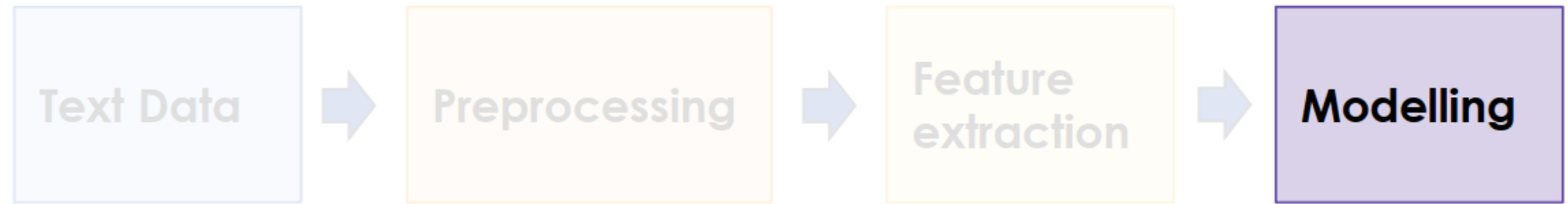
1. Binary Term Frequency (BTF)
2. Bag-of-words (BoW)
3. One-Hot Encoding (OHE)
4. Term Frequency-Inverse Document Frequency (TF-IDF)
5. Word2Vec
6. Global Vectors (GloVe)
7. FastText

<https://medium.com/@archit.saxena/text-to-numbers-the-art-of-text-vectorization-in-nlp-451ce5e845c8>



Data Science

NLP – Grandes etapas



(algumas) Abordagens de modelação

- Análise de Clusters (Clustering)
- Modelação de Tópicos (Topic Modeling - Matrix Factorization)
- *Geração de Texto (Text Generation)* -- RNNs, LSTMs, GPT e BERT.
- *Tradução Automática (Machine Translation)* - seq2seq, ANN's, Transformers:
- *Análise de sentimentos* ([#ver](#))



Data Science/Ciência de Dados

2.1

Modelação - Clustering



Clustering

O que é?

- Método que agrupa dados com características semânticas similares, por exemplo, por conjuntos de textos ou palavras
- Utilizado para descobrir padrões e tópicos em grandes volumes de dados textuais.

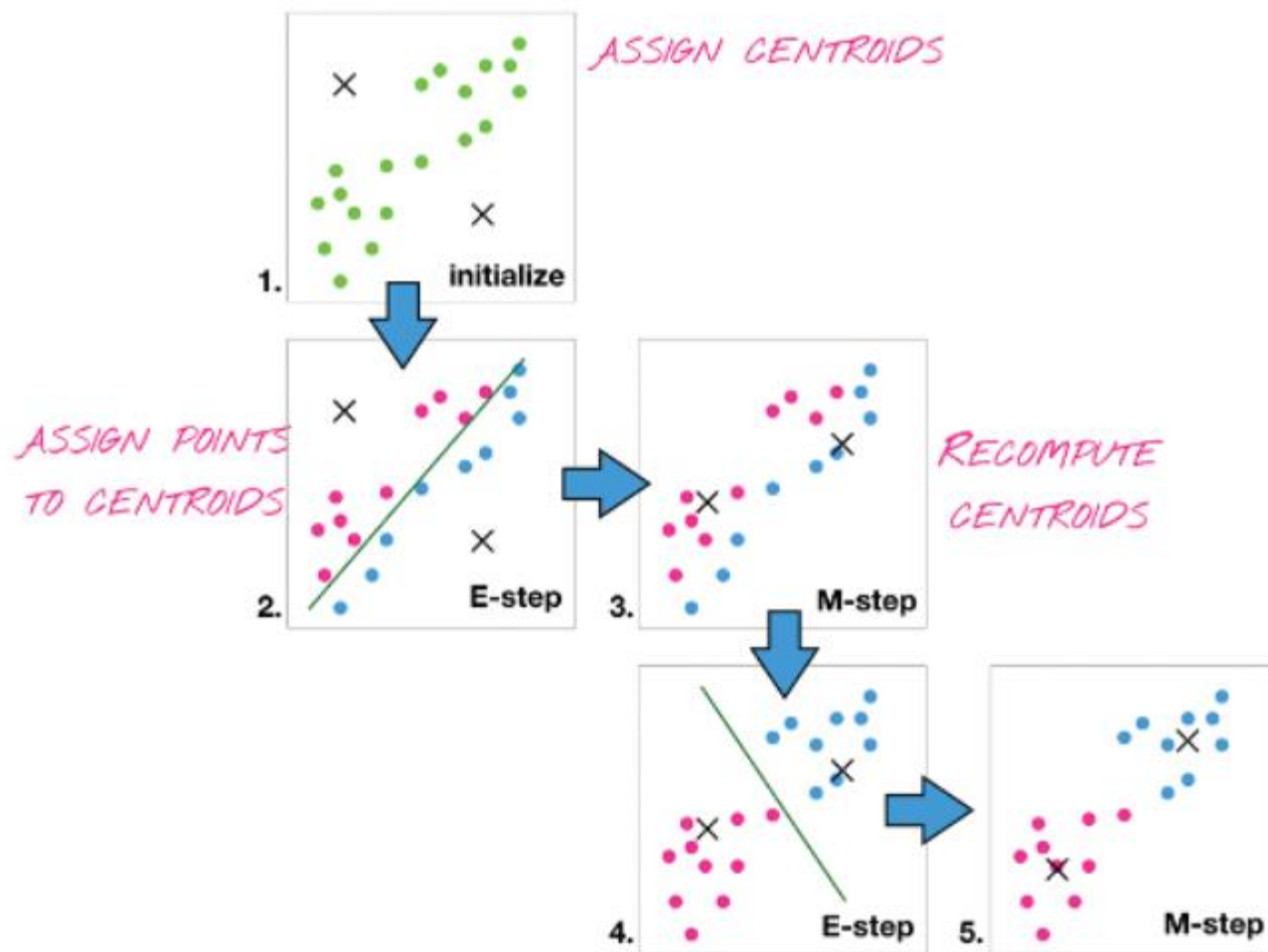
Para que serve?

- Dividir um conjunto de dados em grupos (clusters) de modo a que cada grupo contenha os dados mais semelhantes entre si do que com os de outros grupos.

Que métodos?

- **K-means:** Divide os dados em K grupos com base na proximidade do centroide.
- **Clustering Hierárquico:** Cria uma árvore de clusters baseada em distâncias.
- **DBSCAN:** Agrupa pontos que estão densamente próximos, marcando pontos isolados como ruídos.

k-Means clustering

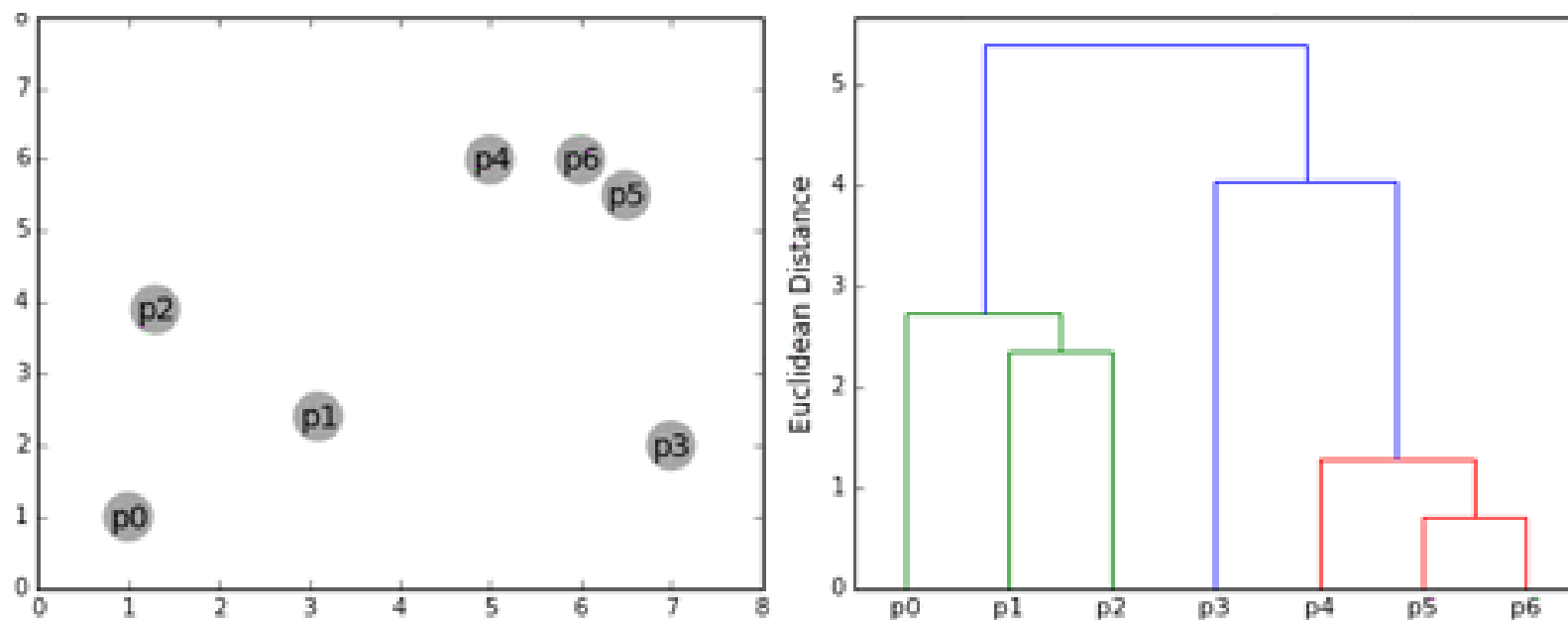


O algoritmo **K-Means** utiliza um número especificado de clusters (K), e segue um processo iterativo para atribuir a cada ponto de dados o cluster mais próximo com base na média dos pontos já presentes no cluster



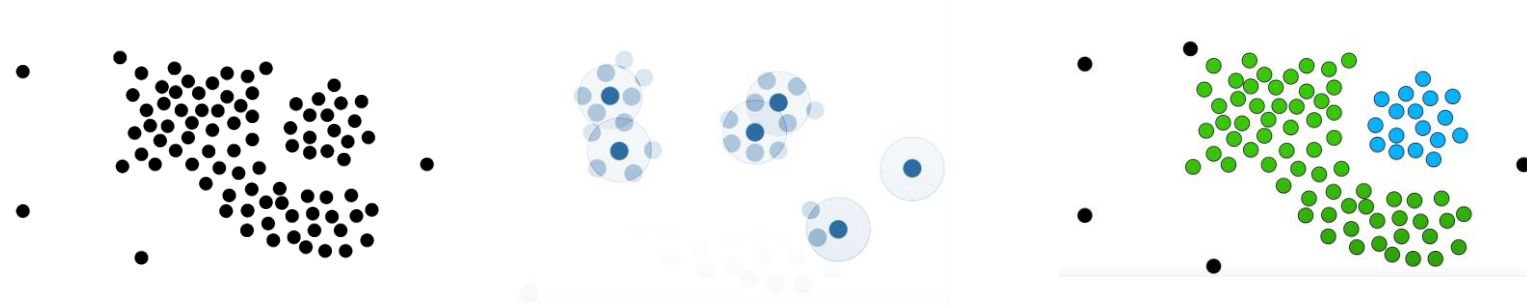
Clustering Hierárquico - Agglomerative Clustering

O algoritmo Hierárquico utiliza o **Dendograma** que identifica a forma como os agrupamentos se vão formados



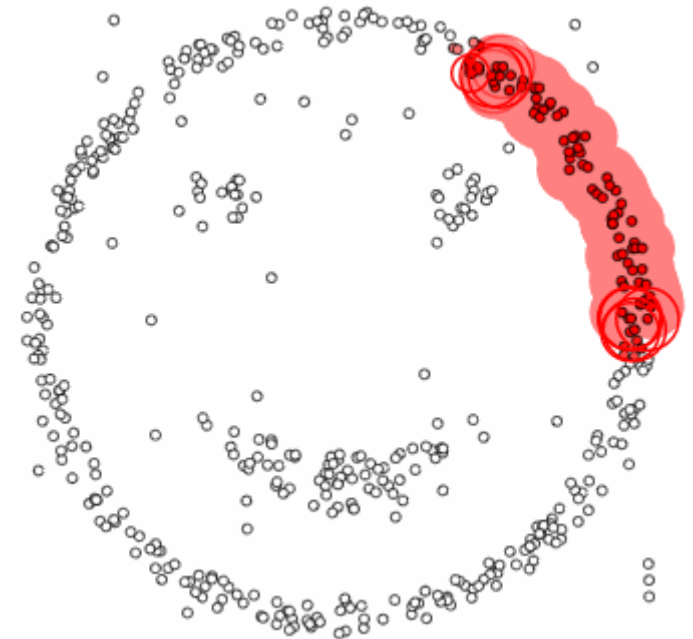
The agglomerative hierarchical clustering results (dendrogram)

(DBSCAN) - Density-based spatial clustering of applications with noise



O algoritmo DBSCAN utiliza dois parâmetros:

- **minPts:** O número mínimo de pontos (um limiar) agrupados para que uma região seja considerada densa.
- **eps (ϵ):** Uma medida de distância que será usada para localizar os pontos na vizinhança de qualquer ponto.





Data Science/Ciência de Dados

2.2

Modelação – Topic Modelling

Modelação de Tópicos (Topic Modelling)

O que é?

- Modelos estatísticos para **descobrir padrões** e **revelar tópicos abstratos** e **estruturas semânticas ocultas** dentro de um texto.

Para que serve?

- Extrair automaticamente os temas subjacentes de grandes volumes de texto.
- Facilitar a organização, compreensão e sumarização de grandes conjuntos de dados textuais.

Que Métodos de Modelos de Tópicos?

- **LDA (Latent Dirichlet Allocation)**: O mais comum, atribui tópicos a documentos e palavras a tópicos
- **HDP (Hierarchical Dirichlet Process)**: versão do LDA que determina o número de tópicos
- **NMF (Non-negative Matrix Factorization)**: fatoração de matrizes para identificar tópicos.
- **(LSA) Latent Semantic Analysis**: utiliza a decomposição em valores singulares (SVD, do inglês) para reduzir a dimensionalidade de matrizes



Topic Modelling

Baseiam-se na ideia de que documentos são misturas de tópicos, onde um tópico é uma distribuição de palavras.

.... abordagem não supervisionada

Tópicos são definidos como "um padrão recorrente de termos co-ocorrentes num corpus de texto.

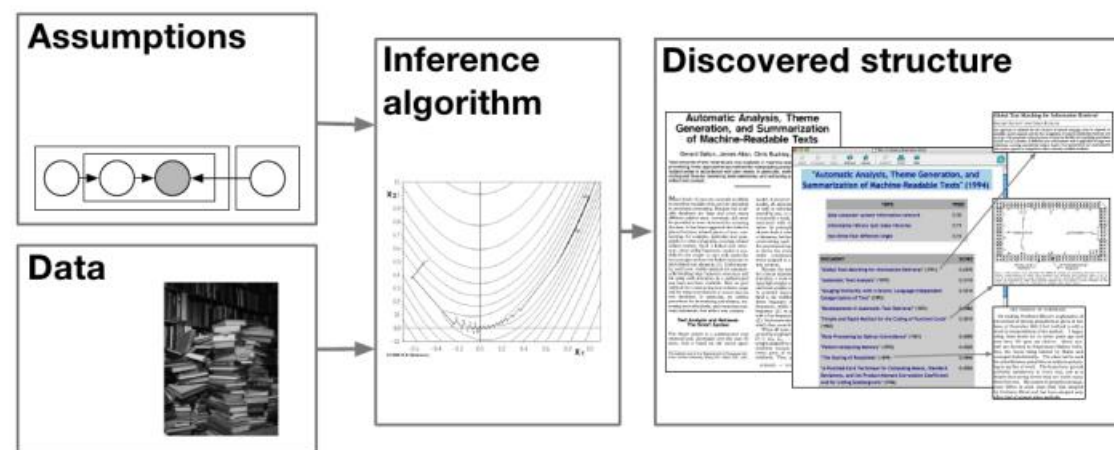
"saúde", "médico", "paciente", "hospital"
*para o **tópico Saúde,***

"fazenda", "colheitas", "trigo"
*para o **tópico Agricultura.***





Topic Modelling



Observação: uma coleção de textos

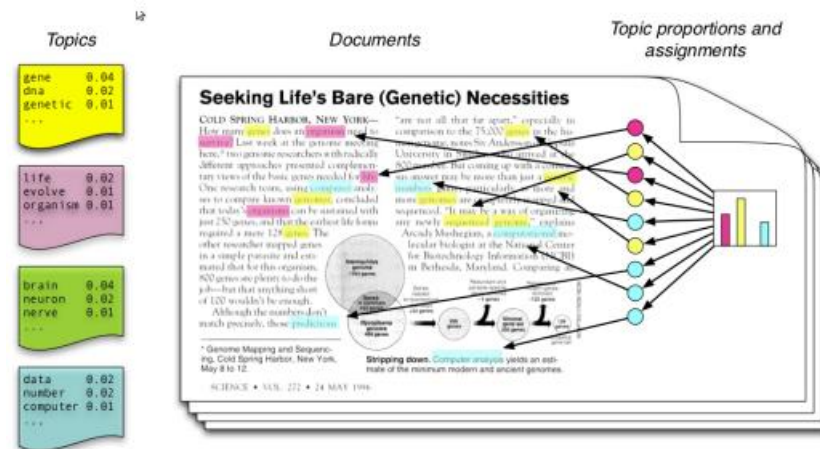
from David Blei, KDD-11 tutorial

Premissa: os textos foram gerados de acordo com modelos

Outputs: o modelo que gerou os textos

Encontrar as estruturas latentes de "**tópicos**" ou "**conceitos**" num **Corpus de texto**, que é obscurecido pelo ruído da "escolha da palavra"

Intuição



Cada documento é uma mistura de tópicos:

$$\sum_k p(z_m = k) = \sum_k \theta_{m,k} = 1$$

Cada palavra é extraída de um dos tópicos do documento:

$$p(w_{m,n}) = \sum_k p(w_{m,n}|z_{m,n} = k)p(z_{m,n} = k) = \sum_k \varphi_k(w_{m,n})\theta_{m,k}$$



1. Latent Dirichlet Allocation (LDA)

Conceitos fundamentais na modelo de LDA

(para mais detalhe consultar [LDA paper by Blei et. al. 2003](#))

- Uma **palavra** é o elemento fundamental dos dados textuais, retirada de um conjunto de vocabulário indexado por $\{1, \dots, V\}$.
- Um **documento** consiste em uma sequência de N palavras denotada por $w = (w_1, w_2, \dots, w_N)$, onde w_n representa a n -ésima palavra na sequência.
- Um **corpo de texto** ou **corpus** é uma coleção de M documentos, representada como $D = (w_1, w_2, \dots, w_M)$.
- Um **tópico** é caracterizado como uma distribuição sobre as palavras. O modelo LDA pressupõe que cada documento é uma mistura de vários tópicos.



1. Latent Dirichlet Allocation (LDA)

... é essencialmente um problema de agrupamento de texto.

... o objetivo é estimar **dois conjuntos de distribuições** ao analisar o corpus:

- A distribuição de **palavras em cada tópico**.
- A distribuição de **tópicos sobre o corpus**

O LDA é um **modelo Bayesiano hierárquico** de três níveis (processo iterativo, chamado **amostragem de Gibbs**, para ajustar essas distribuições, tentando explicar a coleção de documentos observados).

... parte do princípio de que os documentos são produzidos a partir de uma mistura de tópicos e esses tópicos, por sua vez, geram palavras com base na sua distribuição de probabilidade.

Em vez de entrar em detalhes matemáticos complexos, é mais simples explicar o modelo com um [#exemplo](#):

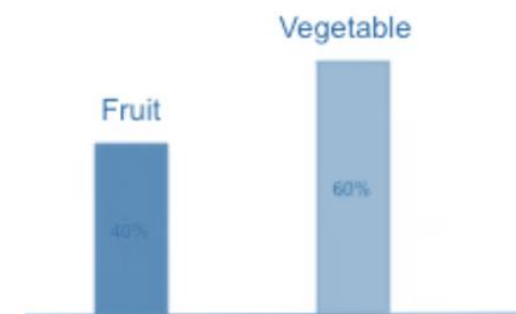


1. Latent Dirichlet Allocation (LDA)

O modelo LDA gera uma **distribuição de tópicos** para **cada documento** no corpus.

Por exemplo, um documento pode ser distribuído em dois tópicos: 40% no tópico "Fruta" e 60% no tópico "Legume".

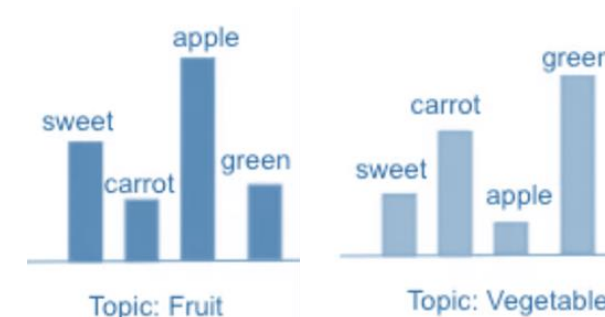
Documento: distribuição de tópicos



Cada **tópico** no modelo LDA é **uma distribuição de todas as palavras** do conjunto do vocabulário, que pode incluir palavras como "doce", "cenoura", "maçã", "verde". Dependendo do tópico, algumas palavras aparecerão com maior probabilidade.

Por exemplo, de forma intuitiva, a palavra "verde" teria uma probabilidade maior de aparecer no tópico "legume" do que no tópico "fruta".

Tópicos: distribuição de palavras

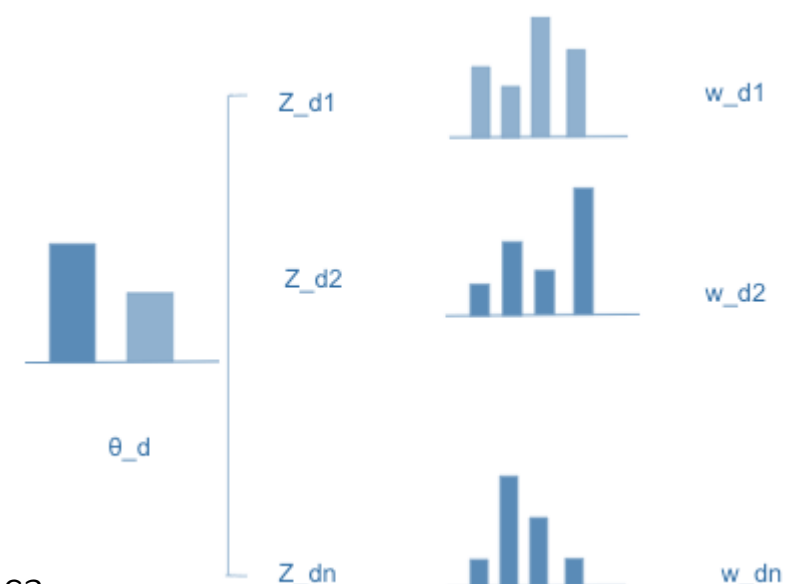




1. Latent Dirichlet Allocation (LDA)

Para o processo de geração de texto no modelo LDA, **precisamos definir duas variáveis latentes** que auxiliam na passagem do corpus para o documento e deste para a palavra.

- A primeira variável latente é θ , que representa a distribuição de tópicos em cada documento (por exemplo, 40% "Fruta", 60% "Legume").
- A segunda variável latente é Z ($Z \in \{1, 2, \dots, T\}$), que indica o tópico de cada palavra.



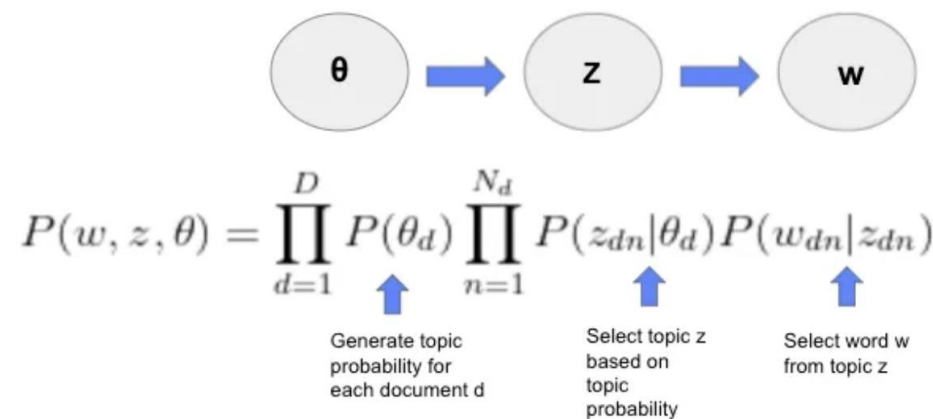
*Uma **variável latente** é aquela que não observamos diretamente nos dados, mas que revela estruturas ocultas nos dados, sendo útil na construção do modelo probabilístico.*



1. Latent Dirichlet Allocation (LDA)

- Para cada documento d , existe uma distribuição de tópicos θ_d .
- Cada palavra i no documento d (w_{di}), é gerada com base na distribuição de tópicos θ_d , no tópico Z para esta palavra di , e na distribuição da palavras sobre o tópico Z_{di} .
- Suponhamos que a primeira palavra é "verde" no documento d .
 - Ela é gerada especificando primeiro uma distribuição de tópicos: 60% "Fruta" e 40% "Legume" (θ_d). Então, para a primeira palavra, amostramos o tópico "Legume" (Z_{d1}).
 - No tópico "Legume", amostramos a palavra "verde" (w_{d1}) a partir da distribuição de tópicos.

Matematicamente, o processo pode ser especificado pela equação Bayesiana:


$$P(w, z, \theta) = \prod_{d=1}^D P(\theta_d) \prod_{n=1}^{N_d} P(z_{dn}|\theta_d) P(w_{dn}|z_{dn})$$

Generate topic probability for each document d

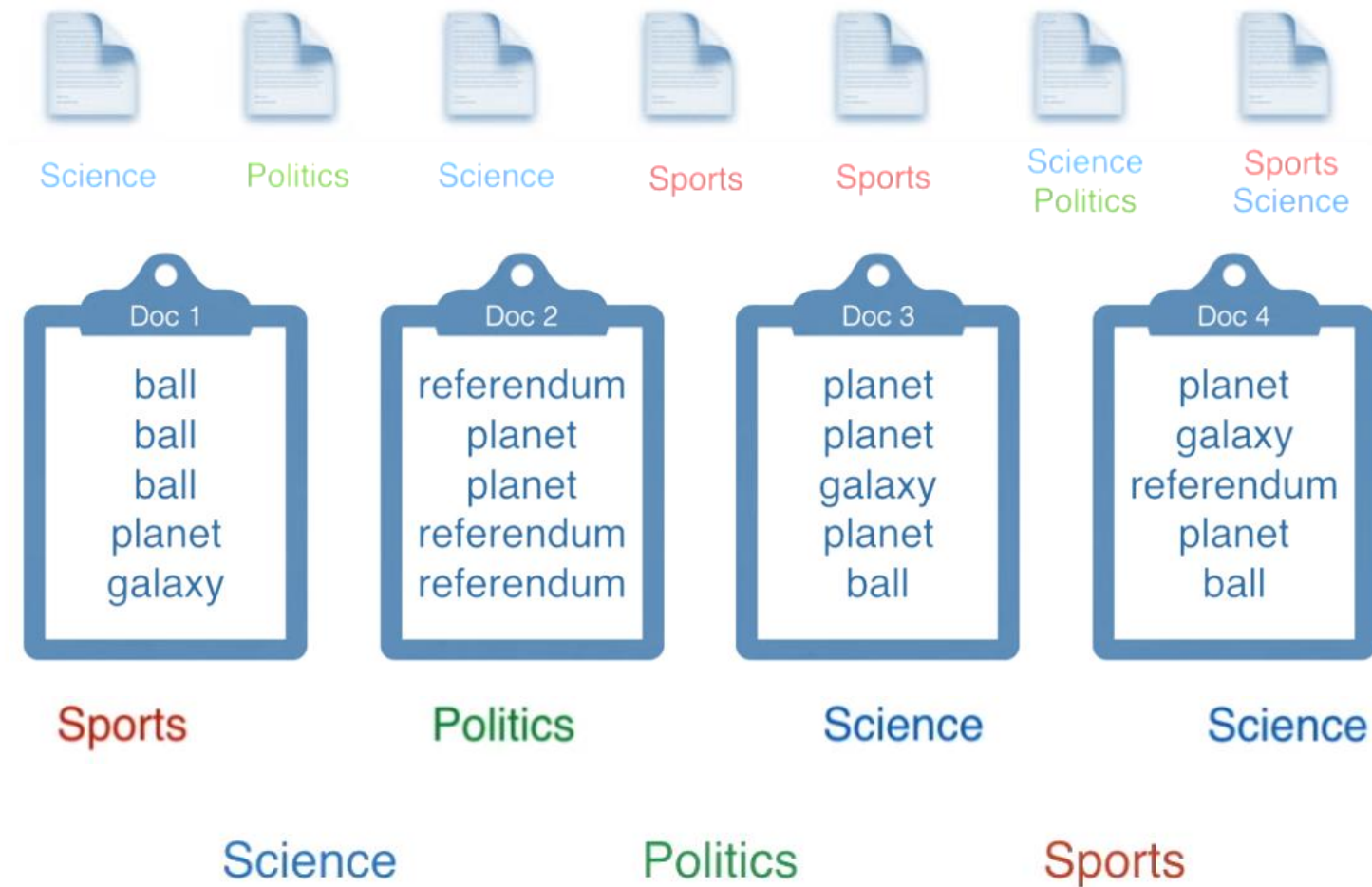
Select topic z based on topic probability

Select word w from topic z



1. Latent Dirichlet Allocation (LDA)

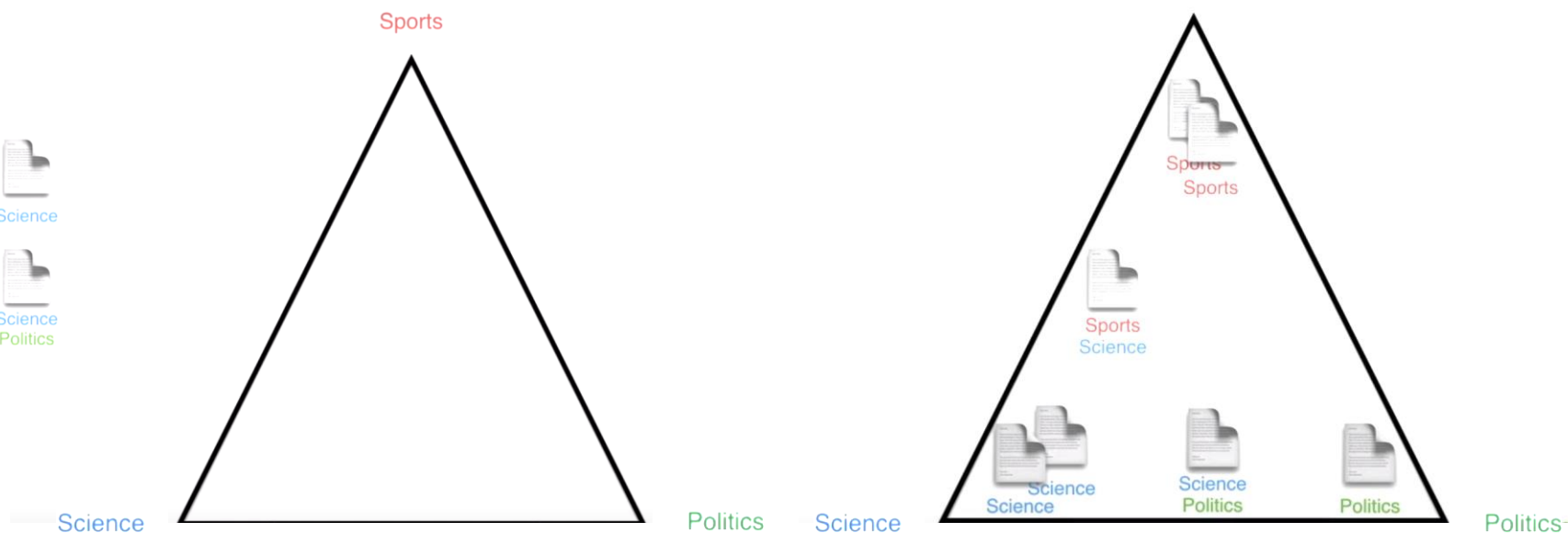
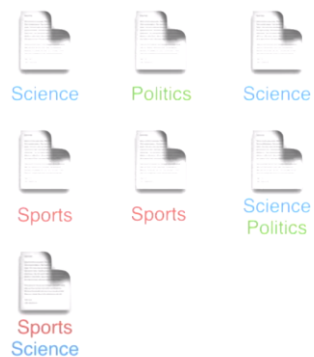
Um outro exemplo





1. Latent Dirichlet Allocation (LDA)

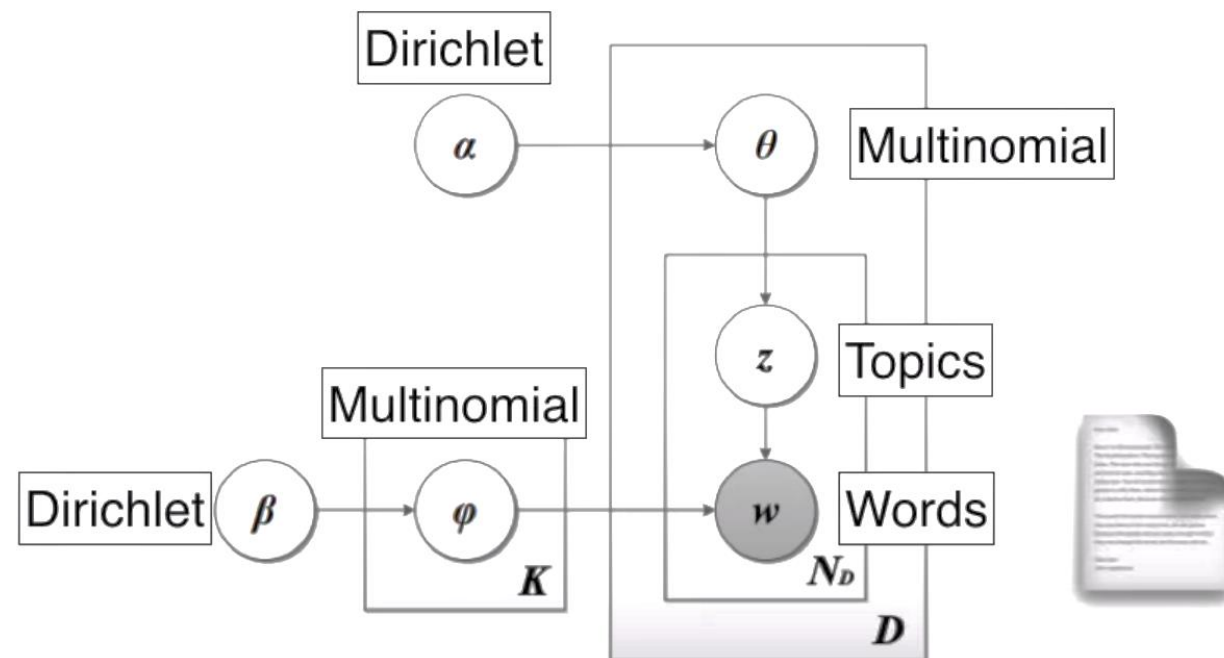
LDA





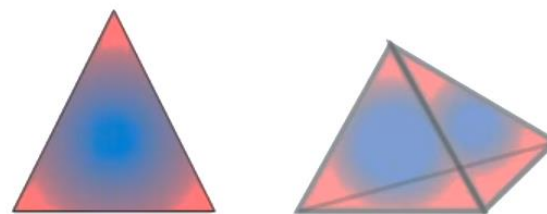
1. Latent Dirichlet Allocation (LDA)

Racional do LDA

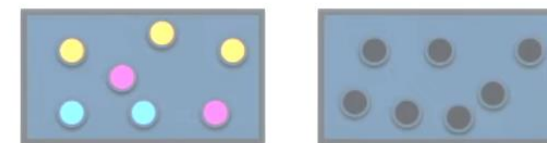


Probabilidade de um documento

$$P(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{j=1}^M P(\theta_j; \alpha) \prod_{i=1}^K P(\varphi_i; \beta) \prod_{t=1}^N P(Z_{j,t} | \theta_j) P(W_{j,t} | \varphi_{Z_{j,t}})$$



Topics Words
Distribuições de Dirichlet

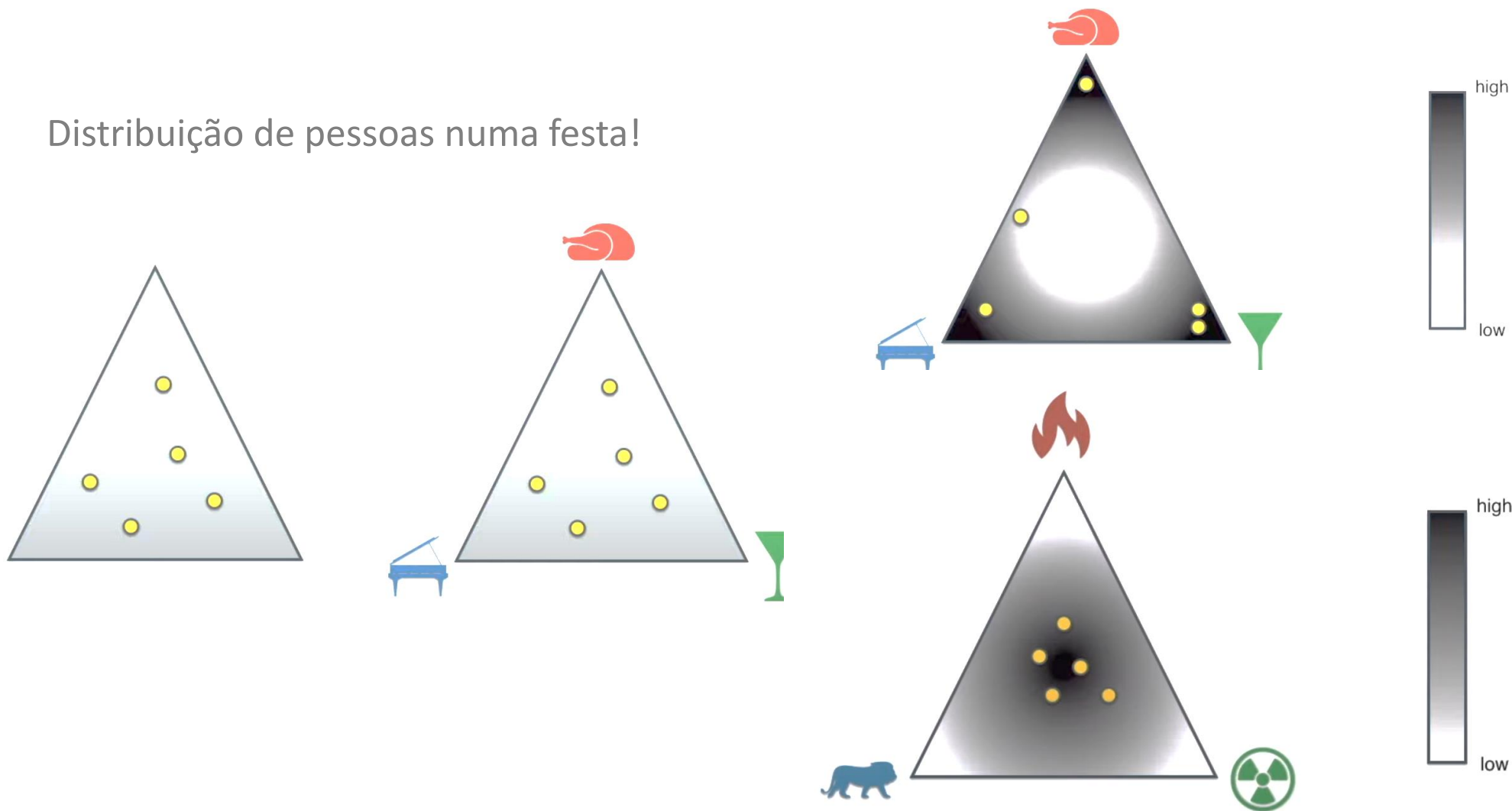


Topics Words
Distribuições Multinominais



1. Latent Dirichlet Allocation (LDA)

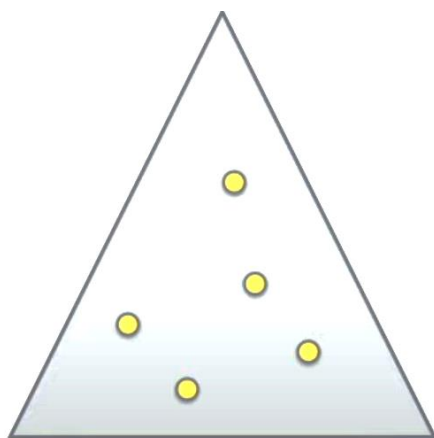
Distribuição de pessoas numa festa!



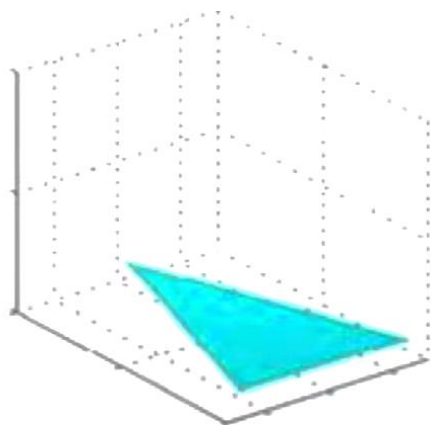


1. Latent Dirichlet Allocation (LDA)

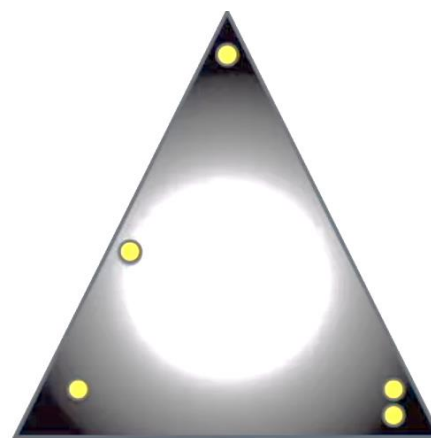
Exemplos de distribuições de Dirichlet $f(x_1, \dots, x_K; \alpha_1, \dots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^K x_i^{\alpha_i - 1}$



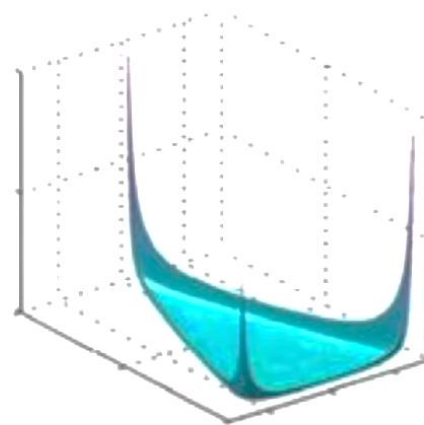
$\alpha = 1$
Uniforme



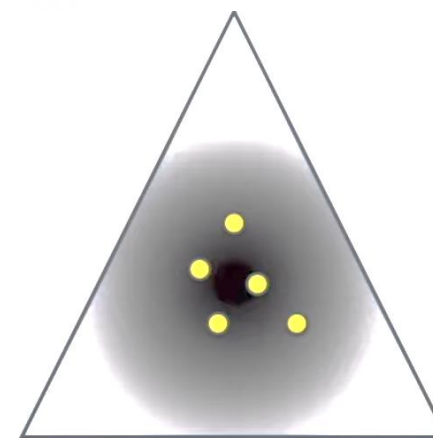
1, 1, 1



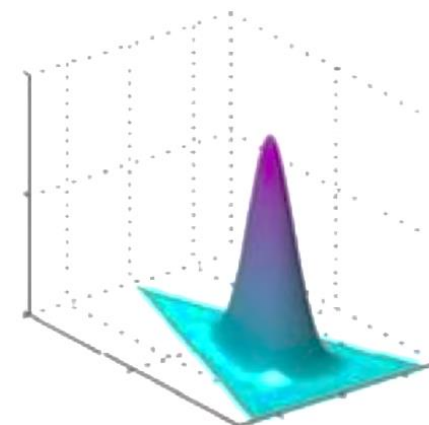
$\alpha < 1$



0.7, 0.7, 0.7



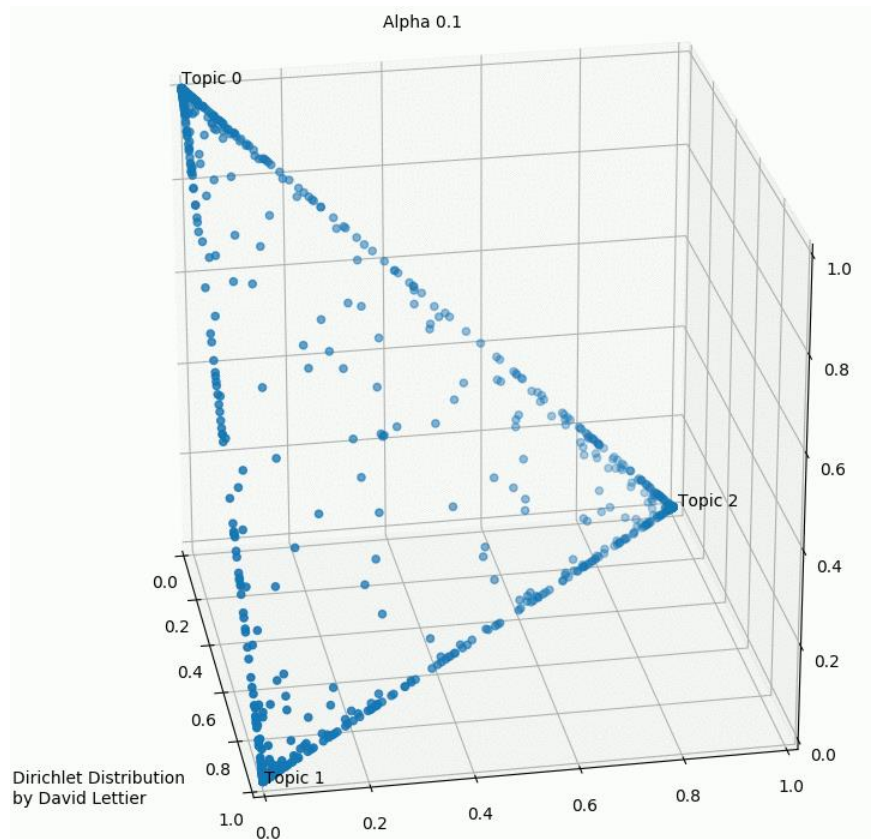
$\alpha > 1$



5, 5, 5



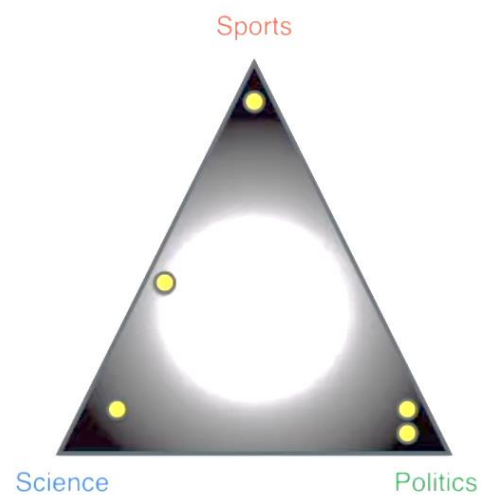
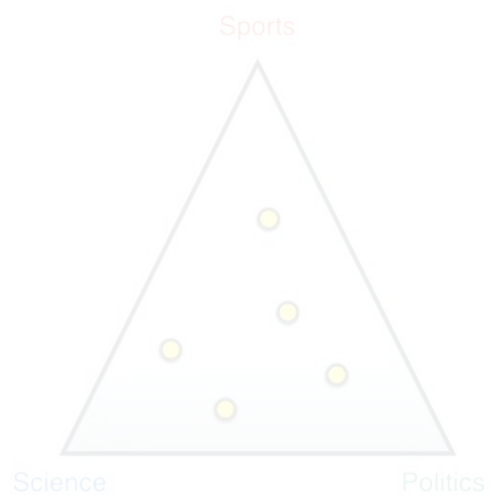
1. Latent Dirichlet Allocation (LDA)





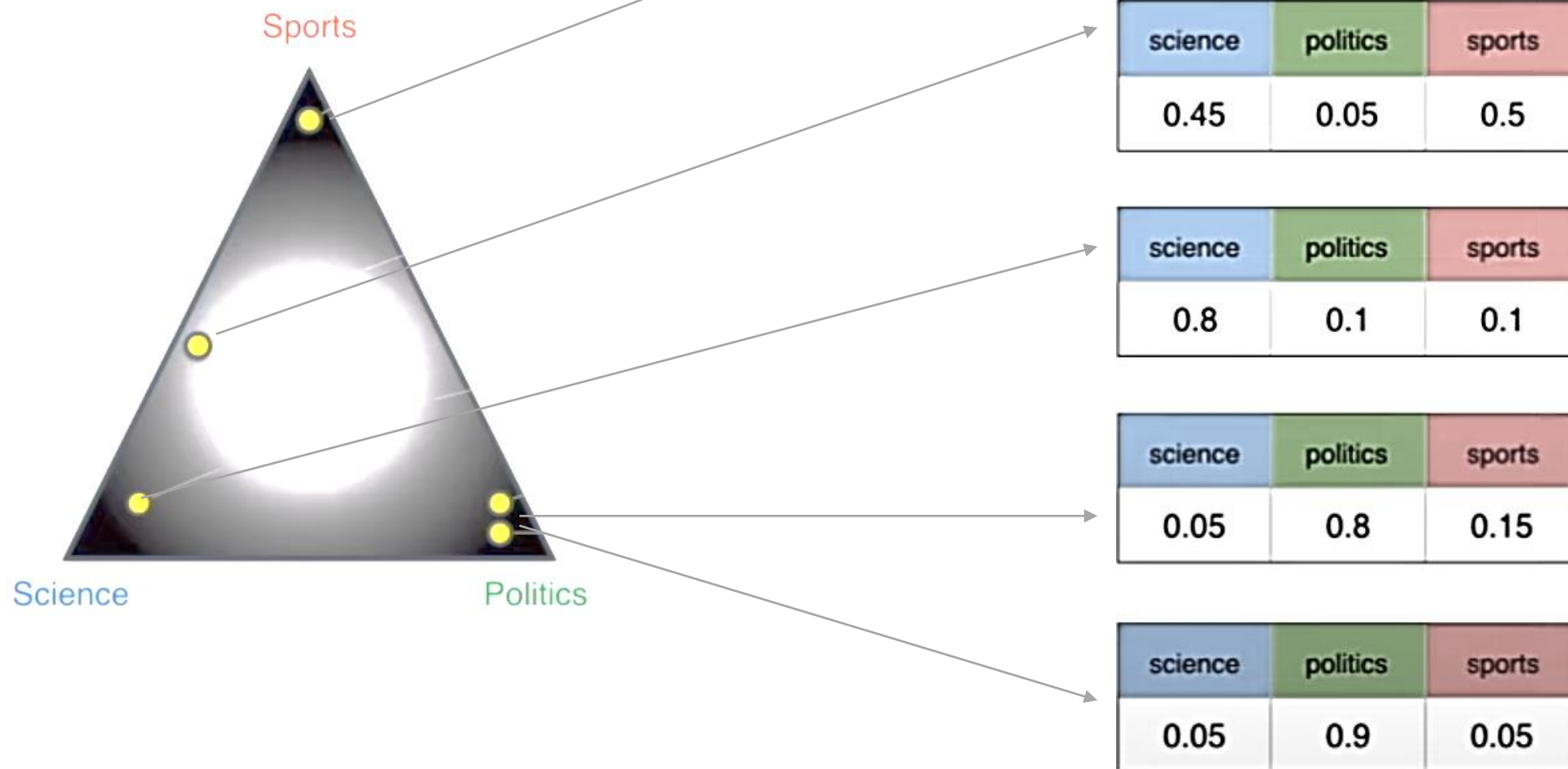
1. Latent Dirichlet Allocation (LDA)

Que tópicos?





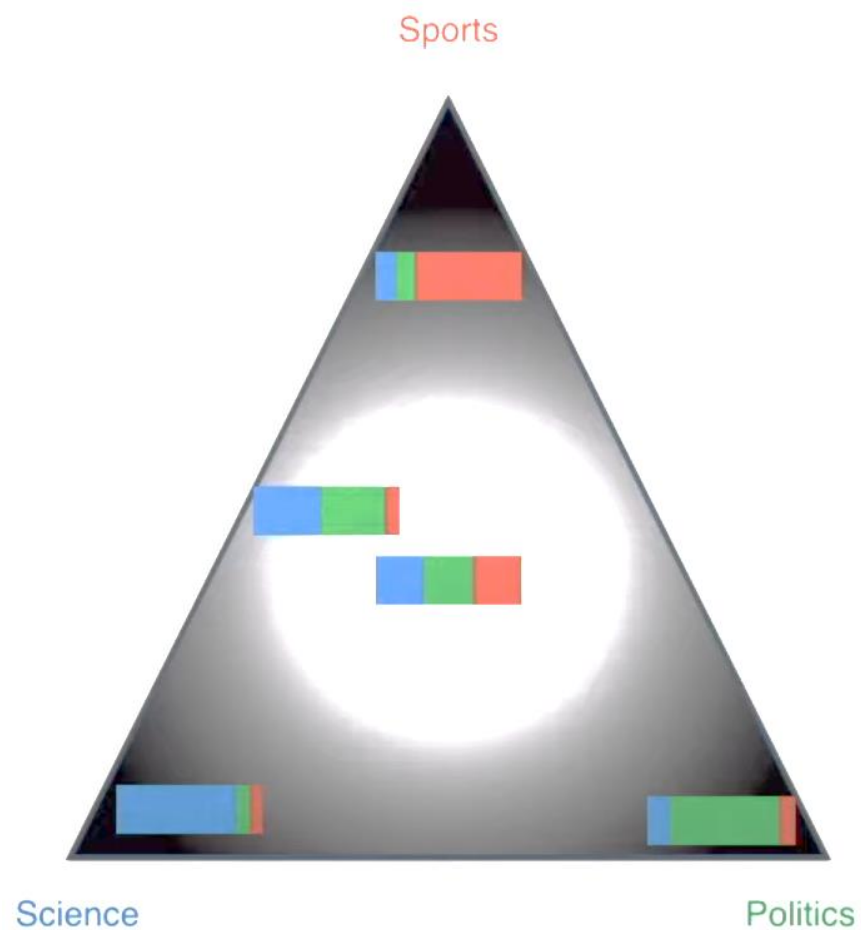
1. Latent Dirichlet Allocation (LDA)





1. Latent Dirichlet Allocation (LDA)

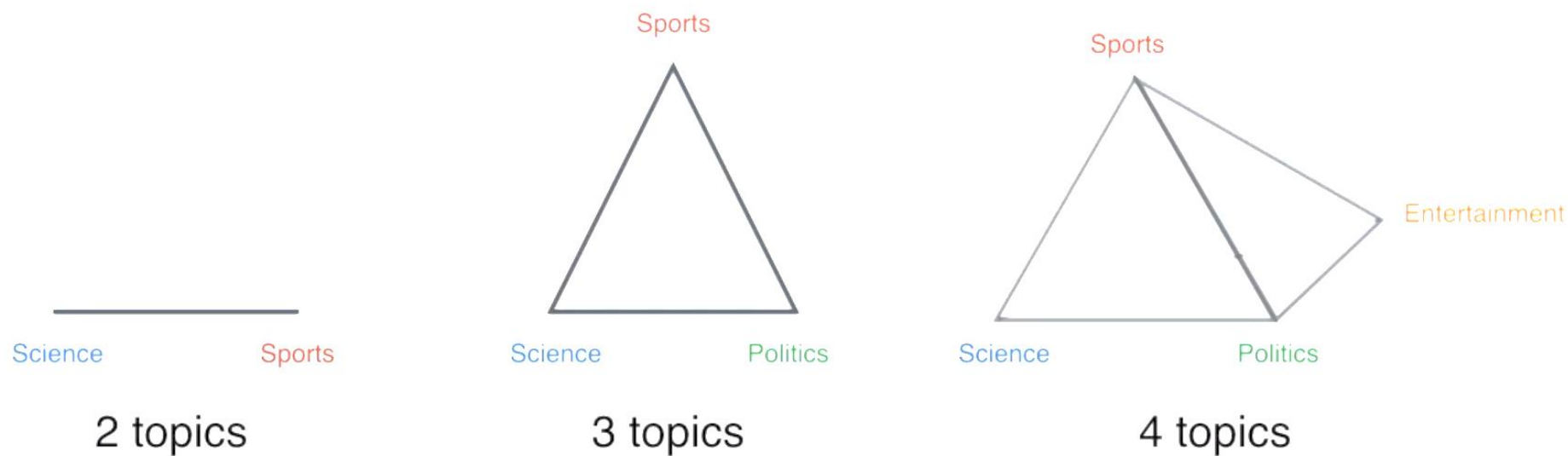
Distribuição de distribuições





1. Latent Dirichlet Allocation (LDA)

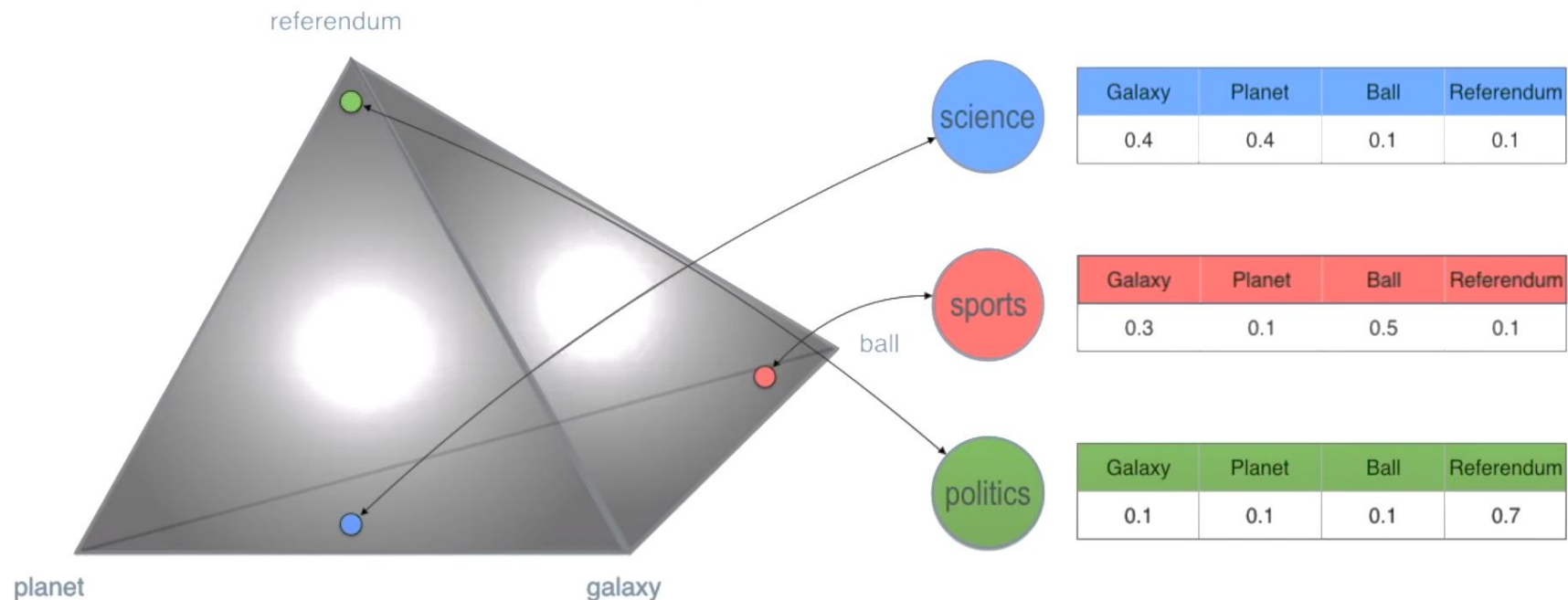
Mais tópicos? Mais dimensões





1. Latent Dirichlet Allocation (LDA)

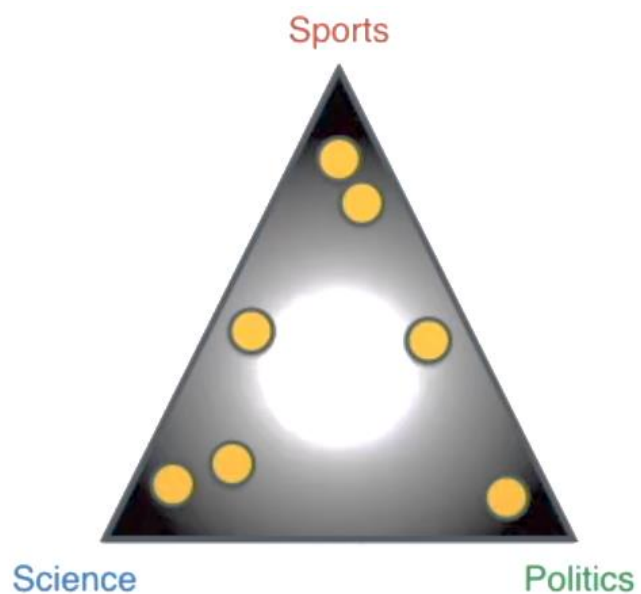
Mais tópicos? Mais dimensões



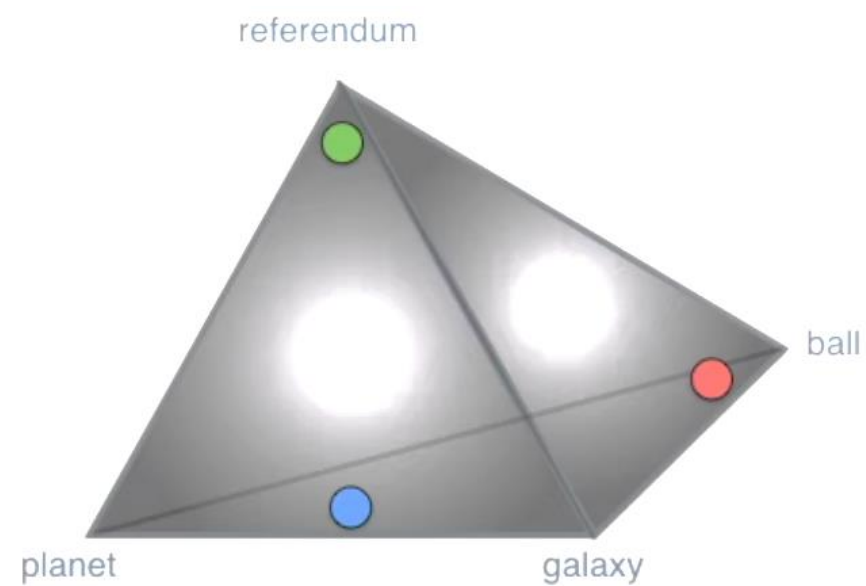


1. Latent Dirichlet Allocation (LDA)

Duas Distribuições De Dirichlet



Documentos - Tópicos

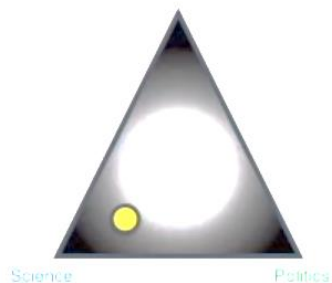


Tópicos - Palavras

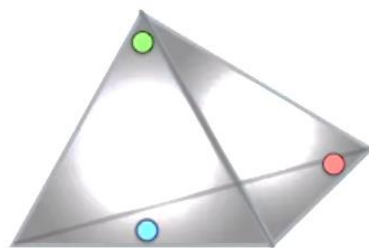


1. Latent Dirichlet Allocation (LDA)

$$\prod_{j=1}^M P(\theta_j; \alpha)$$



$$\prod_{i=1}^K P(\varphi_i; \beta)$$

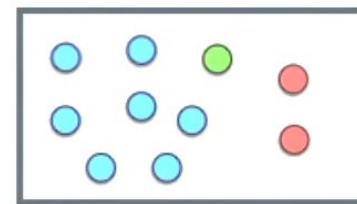


Galaxy	Planet	Ball	Referendum
0.4	0.4	0.1	0.1

Galaxy	Planet	Ball	Referendum
0.1	0.1	0.1	0.7

Galaxy	Planet	Ball	Referendum
0.3	0.1	0.5	0.1

$$\prod_{t=1}^N P(Z_{j,t} | \theta_j)$$



$$P(W_{j,t} | \varphi_{Z_{j,t}})$$

galaxy galaxy planet
galaxy planet ball
galaxy planet planet
referendum

planet ball referendum
referendum referendum
galaxy referendum
referendum referendum

galaxy ball ball galaxy
ball ball galaxy
planet referendum

Topics

science
science
sports
science
science
politics
sports
sports
science

planet
galaxy
ball
planet
galaxy
referendum
galaxy
ball
referendum

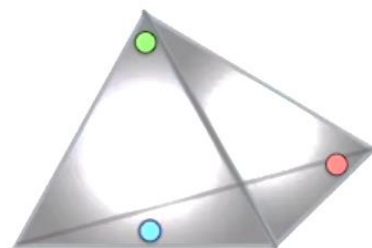
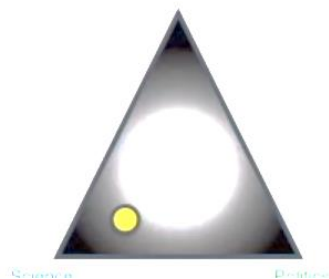


1. Latent Dirichlet Allocation (LDA)

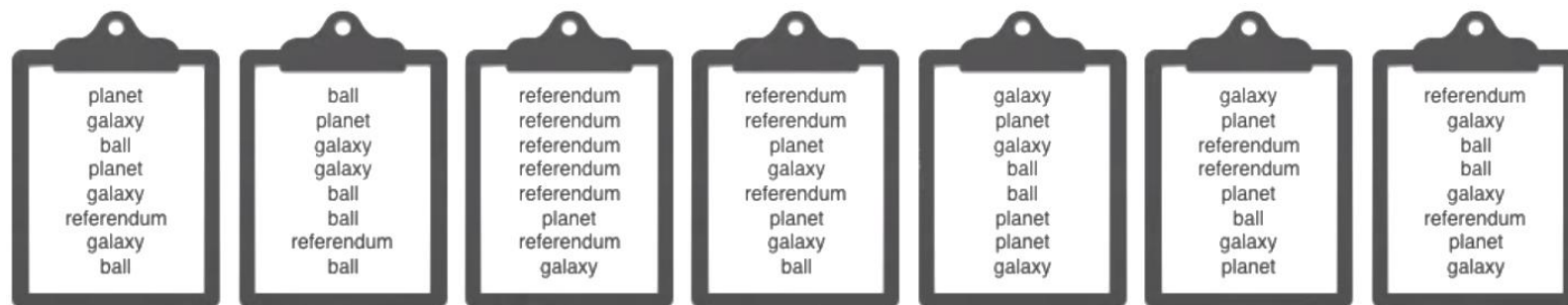
$$\prod_{j=1}^M P(\theta_j; \alpha)$$

$$\prod_{i=1}^K P(\varphi_i; \beta)$$

$$\prod_{t=1}^N P(Z_{j,t} | \theta_j) \quad P(W_{j,t} | \varphi_{Z_{j,t}})$$



Comparação dos documentos gerados
com o originais

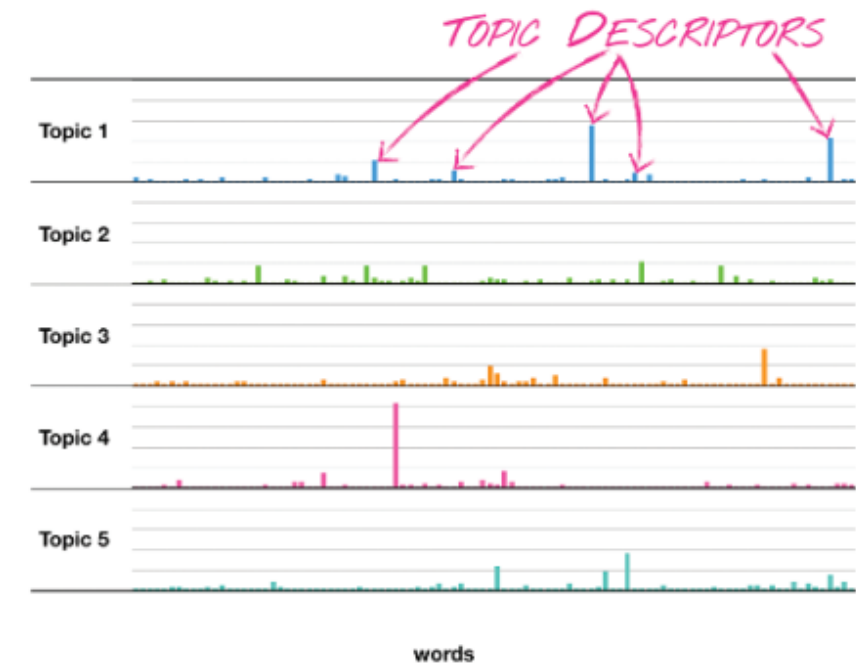
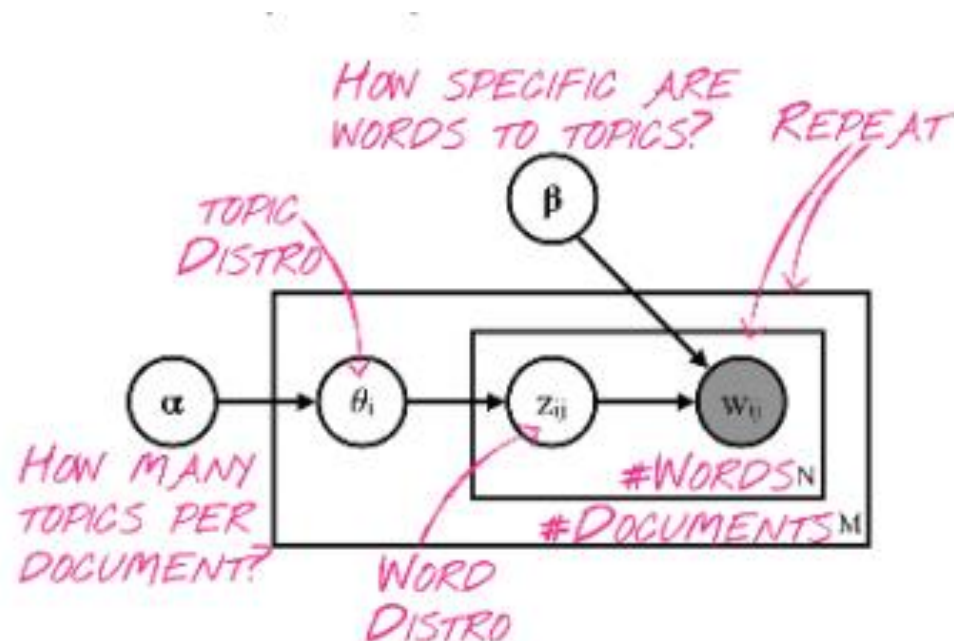
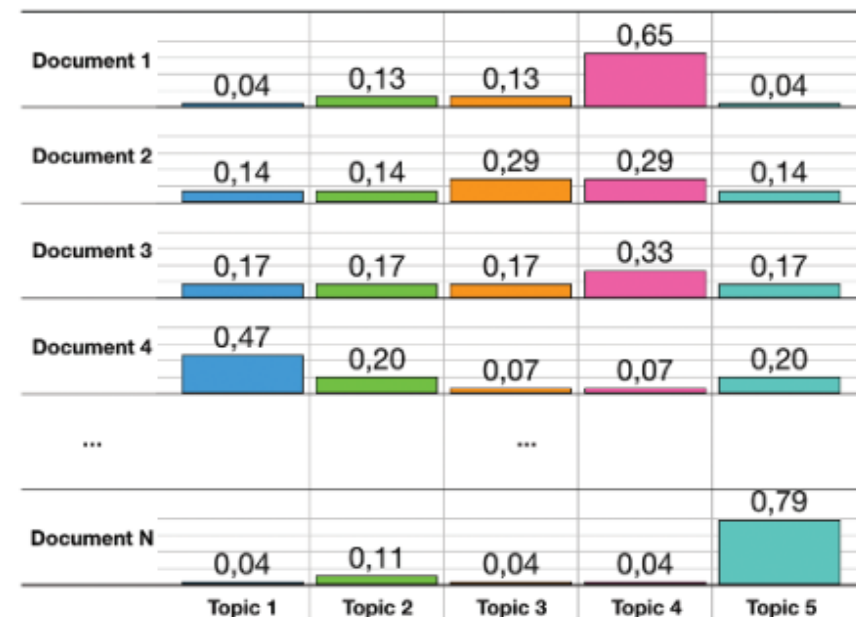




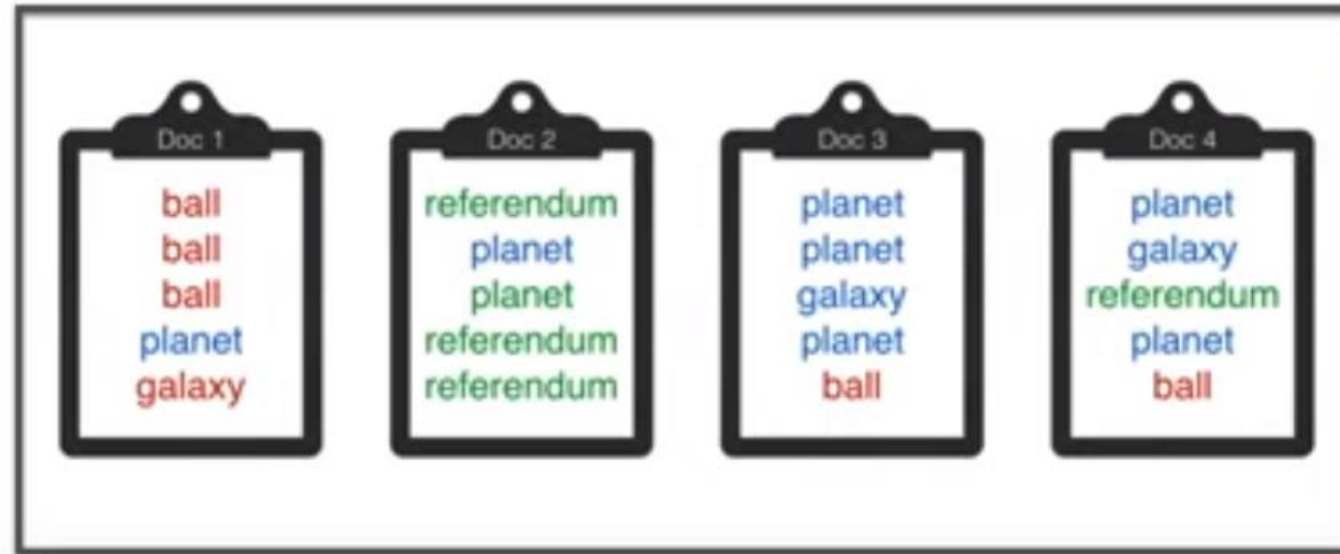
1. Latent Dirichlet Allocation (LDA)

Representação gráfica dos parâmetros do modelo de tópicos $\Theta = P(\text{tópico}|\text{documento})$ (acima) e $Z = P(\text{palavra}|\text{tópico})$ (abaixo).

Os descritores de tópicos são as k palavras mais prováveis com a maior probabilidade para cada tópico em Z .



1. Latent Dirichlet Allocation (LDA)



How to train LDA?
Gibbs sampling

https://www.youtube.com/watch?v=BaM1uiCpj_E

[# LDA paper by Blei *et. al.* 2003.](#)



2. Latent Semantic Analysis (LSA)

Essa abordagem é conhecida como modelo de espaço vetorial

(Deerwester, Dumais, Furnas, Landauer & Harshman, 1990)

Matriz de contagens: Para cada palavra conta-se quantas vezes ela ocorreu no documento - geralmente ponderadas por TF-IDF (Frequência do Termo - Frequência Inversa do Documento), para realçar similaridades latentes.

Transformação da matriz inicial mantendo apenas os k maiores valores próprios da Decomposição em Valores Singulares (SVD - Singular Value Decomposition).



2. Latent Semantic Analysis (LSA)

Matriz Palavra-Documento

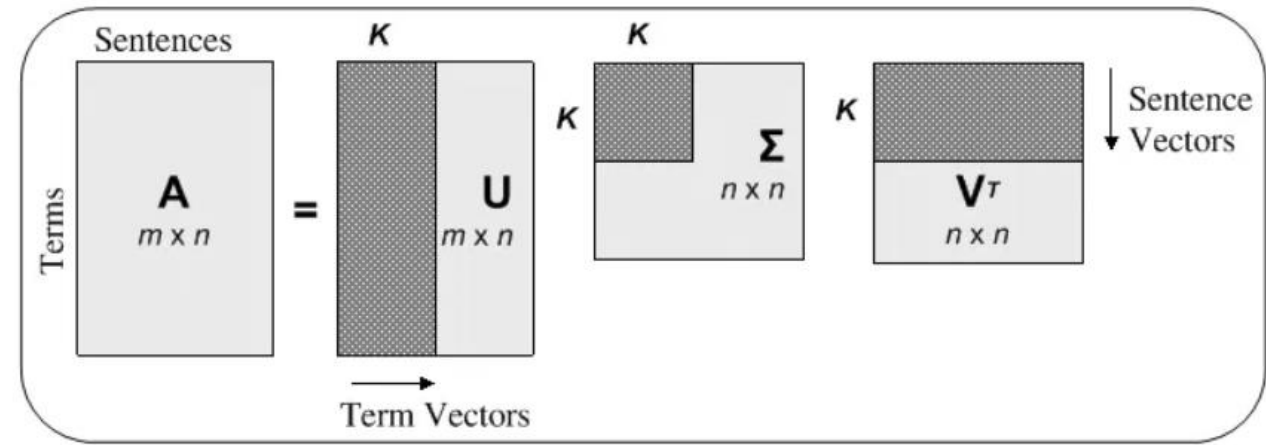
- Representação matemática dos dados de texto
- Linhas correspondem a termos e colunas a documentos
- Entradas da matriz medem a importância de uma palavra num documento

Single Vector Decomposition

- A **LSA é a aplicação da SVD** a uma matriz que representa a relação entre palavras e documentos (matriz palavras-Documento).
- Podemos pensar que cada termo define uma dimensão de entrada e cada documento designa uma amostra.
- SVD é um método para fatorização de matrizes. Usando este método, uma matriz pode ser decomposta em três matrizes



2. Latent Semantic Analysis (LSA)



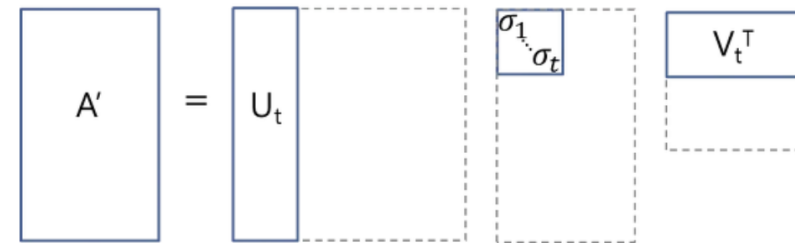
Decomposição em Valores Singulares (SVD - Singular Value Decomposition)

- Decomposição da matriz documento-palavra em três matrizes:
 - U (documento e tópico), mostra a associação dos documentos com os conceitos
 - S (valores singulares), contém valores singulares que indicam a importância de cada tópico nos documentos
 - V (palavra e tópicos), mostra a associação dos termos com os conceitos

Equação: M (matriz termo-documento) = $U \Sigma V^T$



2. Latent Semantic Analysis (LSA)



Decomposição em Valores Singulares (SVD - Singular Value Decomposition)

$$A_{n \times m} = U_{n \times r} S_{r \times r} V_{m \times r}^T$$

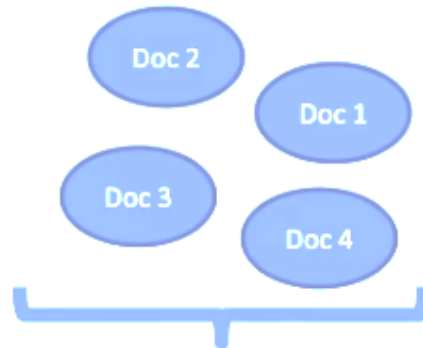
- **A é a matriz documento-palavra.** Ela contém as frequências das n palavras únicas (colunas) nos m documentos (linhas). Esta matriz é aproximadamente a multiplicação de três matrizes: **U**, **S** e **V transposta**. Aqui, a dimensão r denota o número de tópicos.
- **U é a matriz documento-tópico.** Para criar esta matriz, a SVD encontra os valores próprios de $A^T A$ e coloca-os em colunas. Esta operação deixa U com colunas vetores singulares.
- **S é uma matriz diagonal** com valores singulares que são raízes quadradas dos valores próprios (*eigenvalues*) de $A^T A$ ou $A A^T$. Os valores singulares são organizados em ordem decrescente.
- Por fim, **V é a matriz de incorporação de palavras.** Neste caso, a SVD encontra os vetores próprios de $A A^T$ e coloca-os em colunas, deixando V com linhas de vetores singulares



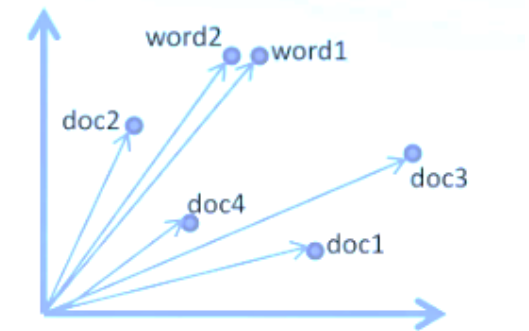
2. Latent Semantic Analysis (LSA)

LSA é um método para descobrir conceitos ocultos em documentos

LSA pressupõe que palavras próximas em significado ocorrerão em trechos similares de texto



Conjunto de documentos,
cada documento contém
várias palavras



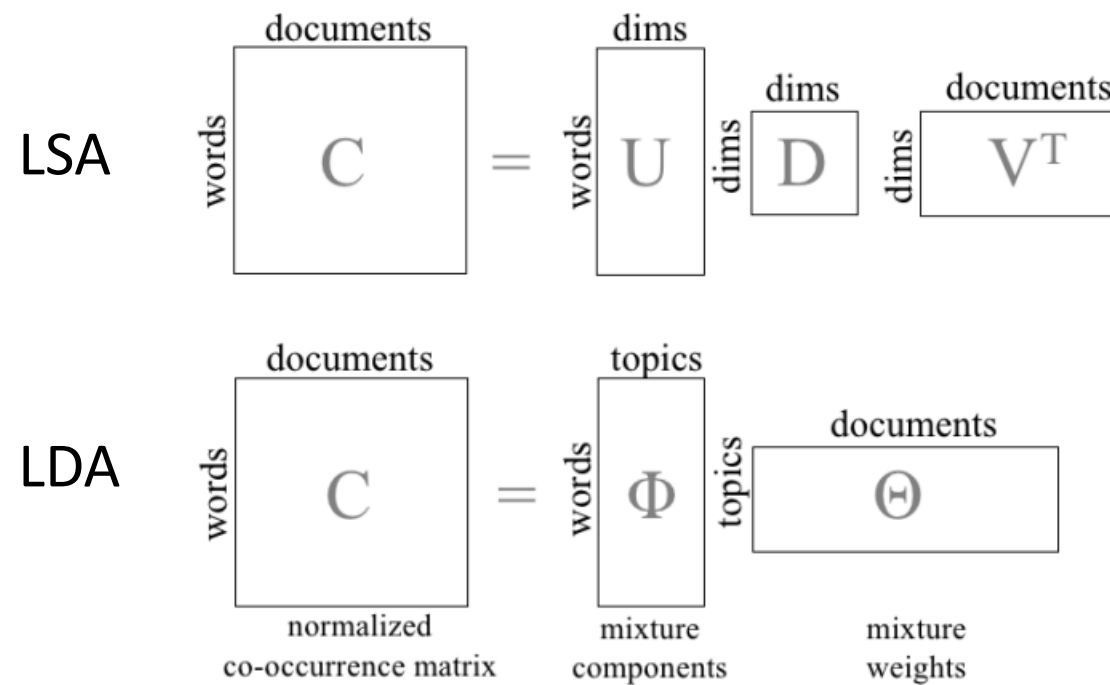
O algoritmo LSA pega em **documentos e palavras** e avalia o vetor no espaço vetorial semântico usando:

- Uma matriz documento/palavra
- decomposição em valores singulares

Espaço vetorial semântico.
Se a palavra 1 e a palavra 2 estão próximas, significa que o seu significado (latente) é relacionado.



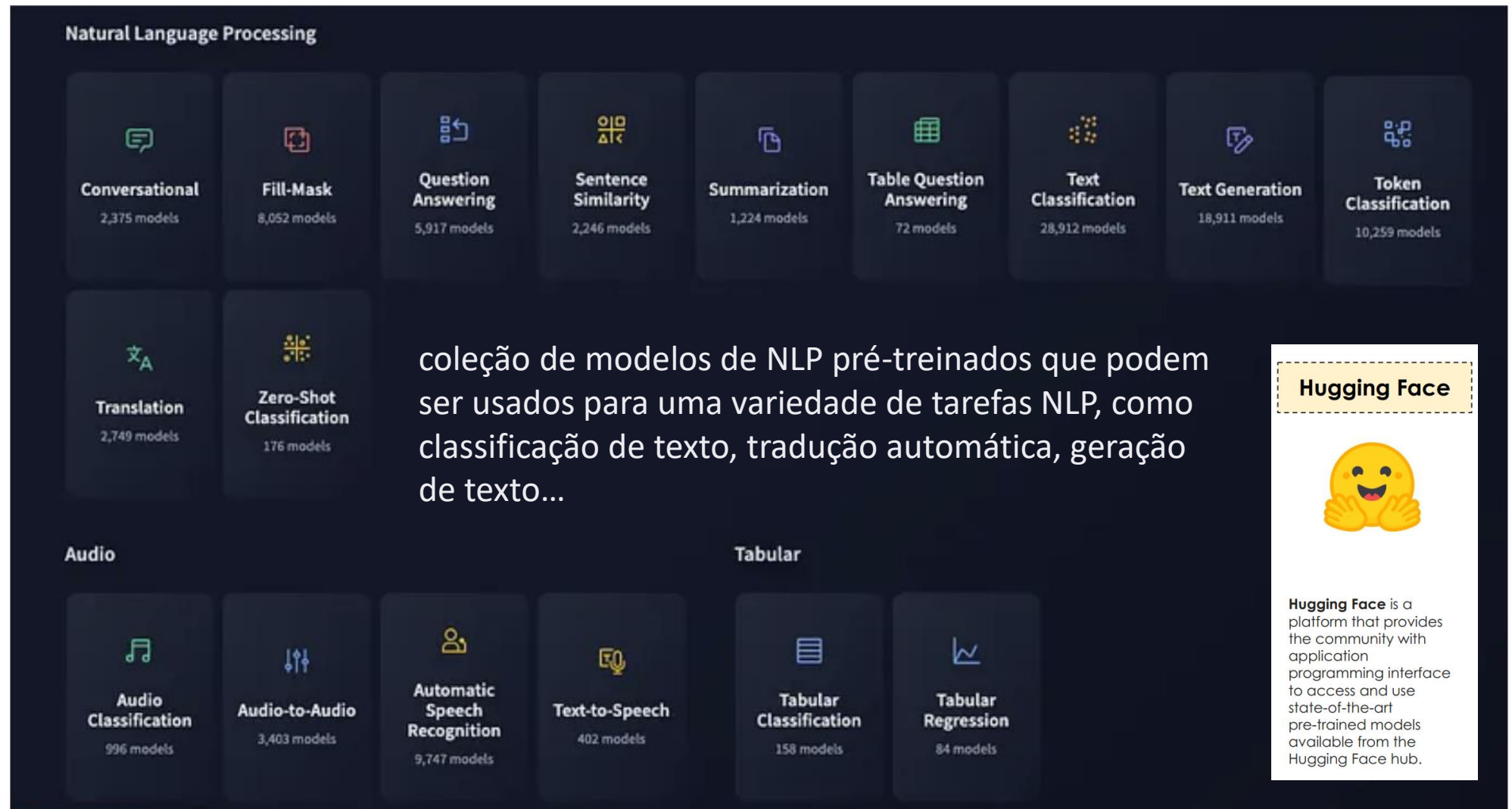
Modelação de Tópicos (*Topic Modelling*) – LDA vs LSA



Para melhorar os resultados da modelação de tópicos:

Para aprimorar os resultados é crucial considerar as características (termos) presentes no corpus, que é representado como uma matriz de termos por documento que é geralmente esparsa. Reduzir a dimensionalidade dessa matriz é crucial para melhorar os resultados.

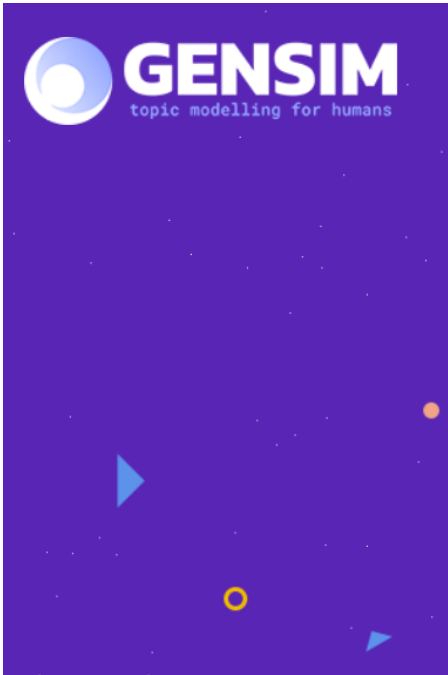
- **Filtro de Frequência:** Organize os termos pela frequência. Termos mais frequentes tendem a ser mais relevantes nos resultados. Termos de baixa frequência são considerados características fracas do corpus e podem ser descartados. Uma análise exploratória pode ajudar a definir um valor de frequência limiar para a exclusão de termos.
- **Filtro de Partes do Discurso (POS):** Este filtro considera o contexto dos termos, não apenas a frequência. Como a modelagem de tópicos busca identificar padrões recorrentes de termos, nem todos são igualmente importantes. Termos que são palavras de suporte na linguagem, como preposições ou modalidades, podem ser identificados e removidos através de suas etiquetas POS.
- **LDA ou LSA em parte:** Para identificar os termos mais importantes dos tópicos, o corpus pode ser dividido em partes de tamanho fixo. Executar o LDA/LSA várias vezes nesses lotes pode gerar resultados variados, mas os termos de tópicos mais relevantes serão aqueles que aparecem na interseção de todas as partes.



Screenshot from Hugging Face tasks
page <https://huggingface.co/tasks>



Data Science



- 1.TextBlob: <https://textblob.readthedocs.io/en/dev/>
- 2.Spacy: <https://spacy.io/>
- 3.Natural Language Toolkit (NLTK): <https://www.nltk.org/>
- 4.Genism: <https://radimrehurek.com/gensim/>
- 5.PyNLPI: <https://github.com/proycon/pynlpl>



<https://github.com/dipanjanS/text-analytics-with-python>

