

# PROCESSAMENTO DE LINGUAGEM NATURAL

## Representação vetorial de textos



# TÓPICOS

1. Representação vetorial
2. *One-hot*
3. Matrizes de Frequência



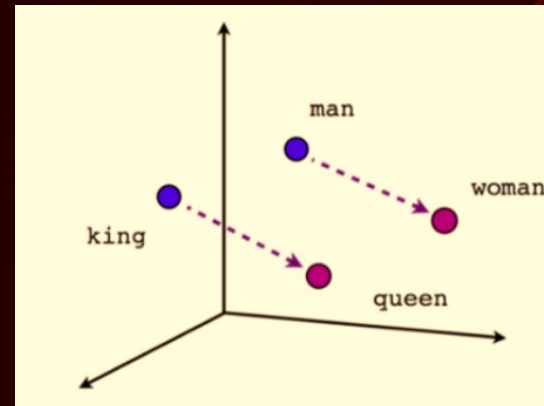
# LINGUAGENS

- O que é a linguagem?

“Sistema de **símbolos de um vocabulário** que, quando colocados numa determinada **ordem** e expressos num determinado **contexto**, emitem um **significado**.”

# REPRESENTAÇÃO DISCRETA X CONTÍNUA

- **Representação discreta**
  - palavras, *tokens*, listas de *tokens*
  - é difícil de fazer manutenção (p. ex. WordNet)
  - é difícil de usar para calcular similaridade entre palavras
- **Representação contínua**
  - vetores numéricos
  - permite aproximações
  - cálculo de similaridade facilitado





# VETORES

- Estruturas compostas por várias dimensões
- Cada dimensão “representa um pouco” do conteúdo

## Dimensões

1. Início com caractere maiúsculo (1: sim, 0: não)
2. Quantidade de caracteres
3. Quantidade de *tokens*
4. Quantidade de vogais
5. Quantidade de consoantes

## cachorro

0	8	1	3	5
---	---	---	---	---

## Lua de mel

1	10	3	4	4
---	----	---	---	---

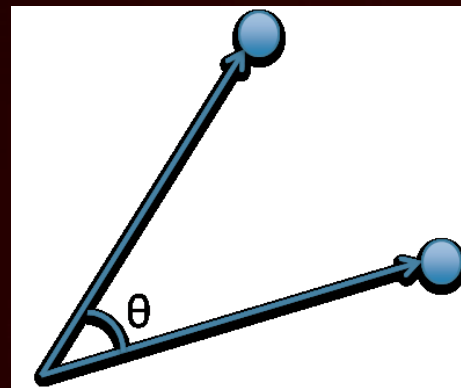
The background is a dark reddish-brown color. It features a complex network of thin, light-colored lines that resemble a circuit board or a neural network. These lines are interconnected by small circles, some of which are highlighted with a slight glow. On the right side of the image, there is a faint, stylized outline of a human brain, composed of the same circuit-like lines. In the center, there is a solid red rectangular box containing the text "ONE-HOT" in white, bold, italicized capital letters.

***ONE-HOT***

# ESPAÇO VETORIAL


- Conjunto de vetores de mesma dimensão
  - Estudado pela Álgebra Linear (Python: *numpy*)
- Cálculo de similaridade (distância) nesse espaço
  - Similaridade de cosseno
  - *sklearn.metrics.pairwise* tem *cosine\_similarity*

$$\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$




# REPRESENTAÇÃO VETORIAL

- *One-hot encoding*
  - A representação tem o tamanho do vocabulário
  - No vetor de uma palavra, a sua posição no vocabulário recebe o valor 1, as demais, 0

	1	0	0	0	...	0	0
Index:	0	1	2	3	...	99998	99999

	0	1	0	0	...	0	0
Index:	0	1	2	3	...	99998	99999



# REPRESENTAÇÃO VETORIAL

- ***One-hot encoding***

- A representação tem o tamanho do vocabulário
- No vetor de uma sentença, as posições de suas palavras recebem valor 1, as demais, 0

[Colab - Vetores](#)

	o	menino	foi	para	a	escola	de	ônibus
O menino foi para a escola	1	1	1	1	1	1	0	0

# MATRIZES DE FREQUÊNCIA

The background is a dark reddish-brown color. It features a complex network of thin, light-colored lines that resemble a circuit board or a neural network. These lines are interconnected by small circles, some of which are highlighted with a slight glow. On the right side of the image, there is a faint, stylized outline of a human brain, composed of the same circuit-like lines, suggesting a connection between technology and the human mind.

# REPRESENTAÇÃO VETORIAL

- **Hipótese Distributiva**
  - Formulada pela primeira vez por Joss (1950), Harris (1954) e Firth (1957)
  - Assume que palavras semelhantes têm contextos similares

“Diga-me com quem andas e eu te direi quem tu és”



# REPRESENTAÇÃO VETORIAL

[Colab - Vetores](#)

- **Matriz de Frequência Termo-Documento**
  - Associa frequência de co-ocorrência de termos em documentos (sentenças)

“e” ocorre  
1 vez no  
documento  
1

	doc1	doc2	doc3	doc4	doc5	...	doc N
e	1	0	0	0	0	...	1
agora	1	0	0	0	0	...	1
josé	1	0	0	0	0	...	1
a	0	1	1	0	1	...	0
...							

representação do  
documento N

representação  
de “a”



# REPRESENTAÇÃO VETORIAL

- **Matriz de Frequência Termo-Termo**
  - **Associa frequência de co-ocorrência entre termos**

[Colab - Vetores](#)

E agora, José?  
A festa acabou,  
a luz apagou,  
o povo sumiu,  
a noite esfriou,  
e agora, José?  
e agora, você?  
você que é sem nome,  
que zomba dos outros,  
você que faz versos,  
que ama, protesta?  
e agora, José?

	e	agora	josé	a	festa	...	protesta
e	0	4	3	0	0	...	0
agora	4	0	3	0	0	...	0
josé	3	3	0	0	0	...	0
a	0	0	0	0	1	...	0
você	1	1	0	0	0	...	0
...	...	...	...	...	...	...	...



# REPRESENTAÇÃO VETORIAL

- **Matriz de TF-IDF**

- Calcula a dimensão de uma palavra pela sua frequência X a frequência inversa de documentos que aparece

[Colab - Vetores](#)

$$w_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right)$$

- $f_{i,j}$  = # de ocorrências de  $i$  em  $j$
- $df_i$  = # de documentos contendo  $i$
- $N$  = # total de documentos

# REPRESENTAÇÃO VETORIAL

- **Problemas**
  - **Esparsidade**
    - Representações vetoriais geradas por métodos de contagem são muito esparsas, ou seja, possuem muitos zeros
  - **Não escalável**
    - À medida que o número de documentos e vocabulário cresce, a dimensão dos vetores torna-se um gargalo

# O QUE VIMOS?

- Representação vetorial
- *One-hot*
- Matrizes de Frequência



# PRÓXIMA VIDEOAULA

- **Prática: Modelos de Linguagem e Representações Vetoriais**

# REFERÊNCIAS

- **Curso de Processamento de Linguagem Natural**
  - **Profa. Helena Caseli (UFSCar)**
- **Curso de Processamento de Linguagem Natural**
  - **Prof. Thiago Pardo (ICMC-USP)**
- **Curso de Linguística Computacional**
  - **Prof. Thiago Castro Ferreira (UFMG)**