FACULDADE IBRATEC

Representação de Texto

Prof. Marcelo Pita

Pós em Inteligência Artificial

Processamento de Linguagem Natural

Texto natural

“Historicamente causadores de inúmeras vítimas, os acidentes de trânsito vêm ocorrendo com frequência cada vez menor, no Brasil. Essa redução se deve, principalmente, à implantação da Lei Seca ao longo de todo o território nacional, diminuindo a quantidade de motoristas que dirigem após terem ingerido bebida alcoólica . A maior fiscalização, aliada à imposição de rígidos limites e à conscientização da população, permitiu que tal alteração fosse possível.”

Texto natural

Não é uma boa representação para algoritmos de aprendizado de máquina.

POR QUÊ?

“Historicamente causadore
trânsito vêm ocorrendo cor
. Essa
redução se deve, principaln
ngo de
todo o território nacional, d
que
dirigem após terem ingerido bebida alcoólica . A maior fiscalização, aliada à imposição de rígidos limites e à conscientização da população, permitiu que tal alteração fosse possível.”

N-gramas

Algumas sequências de palavras têm maior probabilidade de ocorrer do que aleatoriamente.

- **Exemplos:** “São João”, “caldo de cana”, “processamento de linguagem natural”

Estes grupos de palavras são conhecidos como **n-gramas**.

Unigramas (palavras individuais), brigramas, trigramas, etc.

N-gramas

<Prática>

Modelos Vetoriais

Representam documentos como vetores em uma matriz documento-termo.

	Palavra 1	Palavra 2	Palavras 3	...	Paalavra *V*
Documento 1					
Documento 2					
...					
Documento *N*					

Modelos Vetoriais

Representam documentos como vetores em uma matriz documento-termo.

	Palavra 1	Palavra 2	Palavras 3	...
Documento 1				
Documento 2				
...				
Documento *N*				

REPRESENTAÇÃO DO DOCUMENTO

Modelos Vetoriais

Representam documentos como vetores em uma matriz documento-termo.

	Palavra 1	Palavra 2	Palavras 3	...	Paalavra *V*
Documento 1					
Documento 2					
...					
Documento *N*					

REPRESENTAÇÃO DA PALAVRA

Modelos Vetoriais

Booleano

- Palavras ocorrem ou não em um documento (0 ou 1)

	Palavra 1	Palavra 2	Palavras 3	...	Paalavra *V*
Documento 1	0	1	1	...	0
Documento 2	0	0	0	...	1
...	...	...	...	...	...
Documento *N*	0	0	1	...	1

Modelos Vetoriais

Matriz de frequência de termos (TF)

- Frequência das palavras nos documentos é registrada

	Palavra 1	Palavra 2	Palavras 3	...	Paalavra *V*
Documento 1	0	3	1	...	0
Documento 2	0	0	0	...	15
...	...	...	...	...	...
Documento *N*	0	0	2	...	5

Modelos Vetoriais

Matriz TF-IDF (term frequency inverse document frequency)

- **TF**: frequência dos termos nos documentos.
- **IDF**: frequência inversa nos documentos (raridade da palavra).

$$\mathbf{tf}(t,d) = \frac{f_d(t)}{\max_{w \in d} f_d(w)}$$

$$\mathbf{idf}(t,D) = \ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

$$\mathbf{tfidf}(t,d,D) = \mathbf{tf}(t,d) \cdot \mathbf{idf}(t,D)$$

$$\mathbf{tfidf}'(t,d,D) = \frac{\mathbf{idf}(t,D)}{|D|} + \mathbf{tfidf}(t,d,D)$$

$f_d(t) :=$ frequency of term t in document d

$D :=$ corpus of documents

Modelos Vetoriais (resumo)

Bag of Words (BoW)

$$\mathbf{DT} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Term Frequency (TF)

$$\mathbf{DT} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 5 & 1 & 0 \end{bmatrix} \end{array}$$

Term Frequency–Inverse Document Frequency (TF-IDF)

$$idf_{j\,D} = log\left(\frac{|D|}{1 + |\{d \in D : j \in d\}|}\right)$$

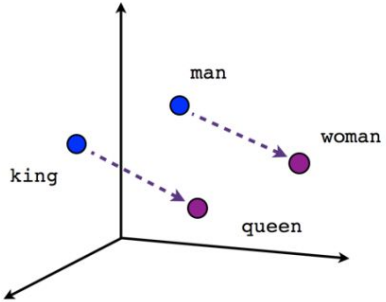
$$\widehat{\mathbf{DT}} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{bmatrix} 0 & 0.83 & 0.17 & 0 & 0 & 0 \\ 0 & 0 & 1.0 & 0 & 0 & 0 \\ 0.33 & 0 & 0 & 0 & 0 & 0.67 \\ 0 & 0 & 0 & 0.83 & 0.17 & 0 \end{bmatrix} \end{array}$$

Modelos Vetoriais

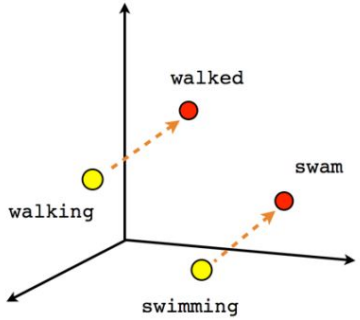
<Prática>

Vetores de Palavras

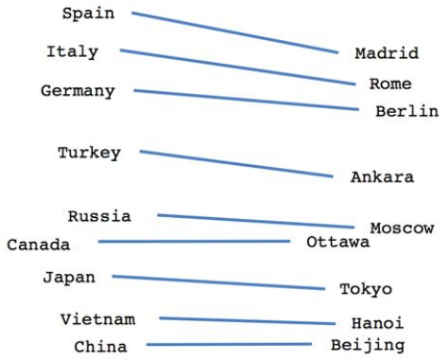
Vetores de palavras são representações vetoriais de palavras concebidas para capturar semântica e serem consistentes com álgebra de vetores.



Male-Female



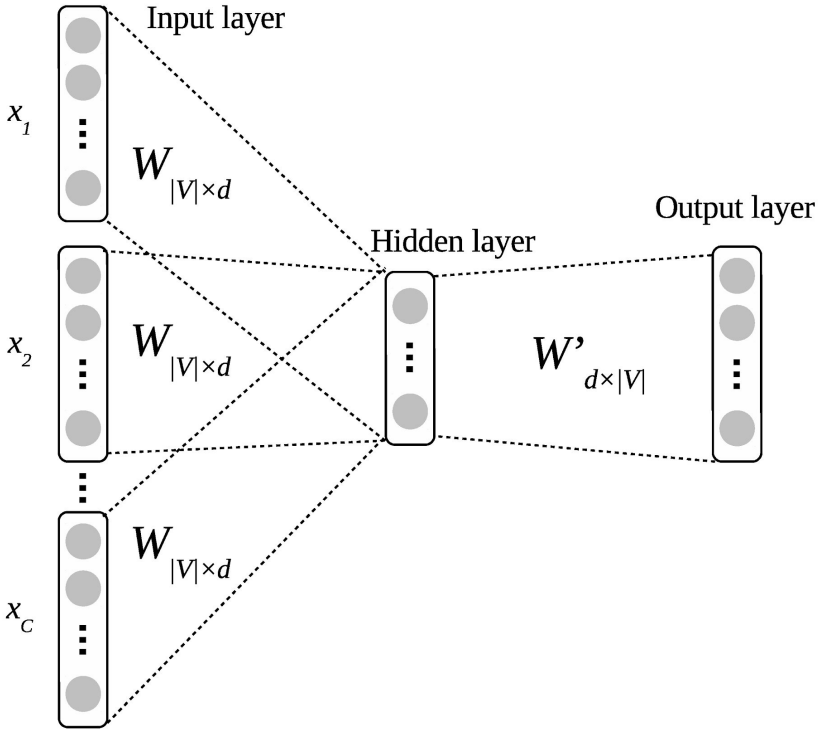
Verb tense



Country-Capital

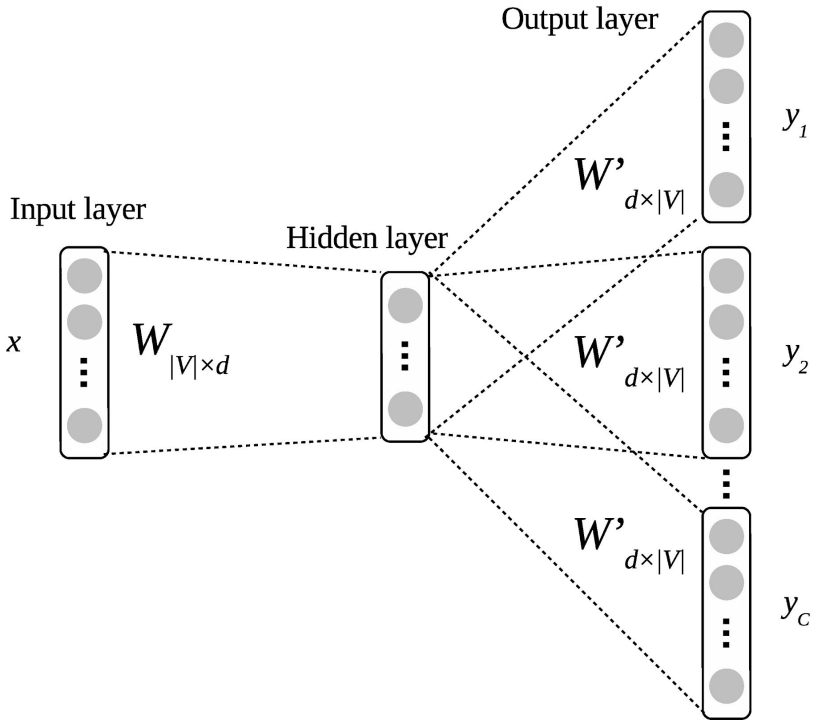
Vetores de Palavras

Continuous Bag of Words (CBOW)



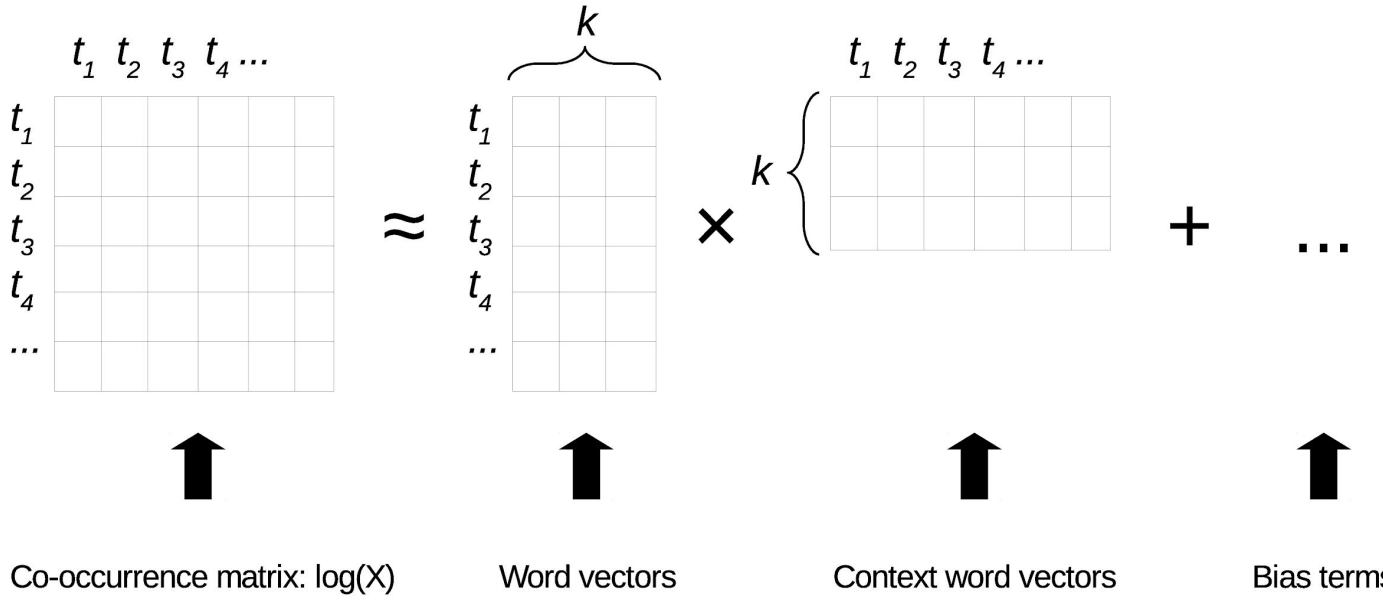
Vetores de Palavras

Skip-Gram



Vetores de Palavras

Global Vectors (GloVe)



Vetores de Palavras

<Prática>

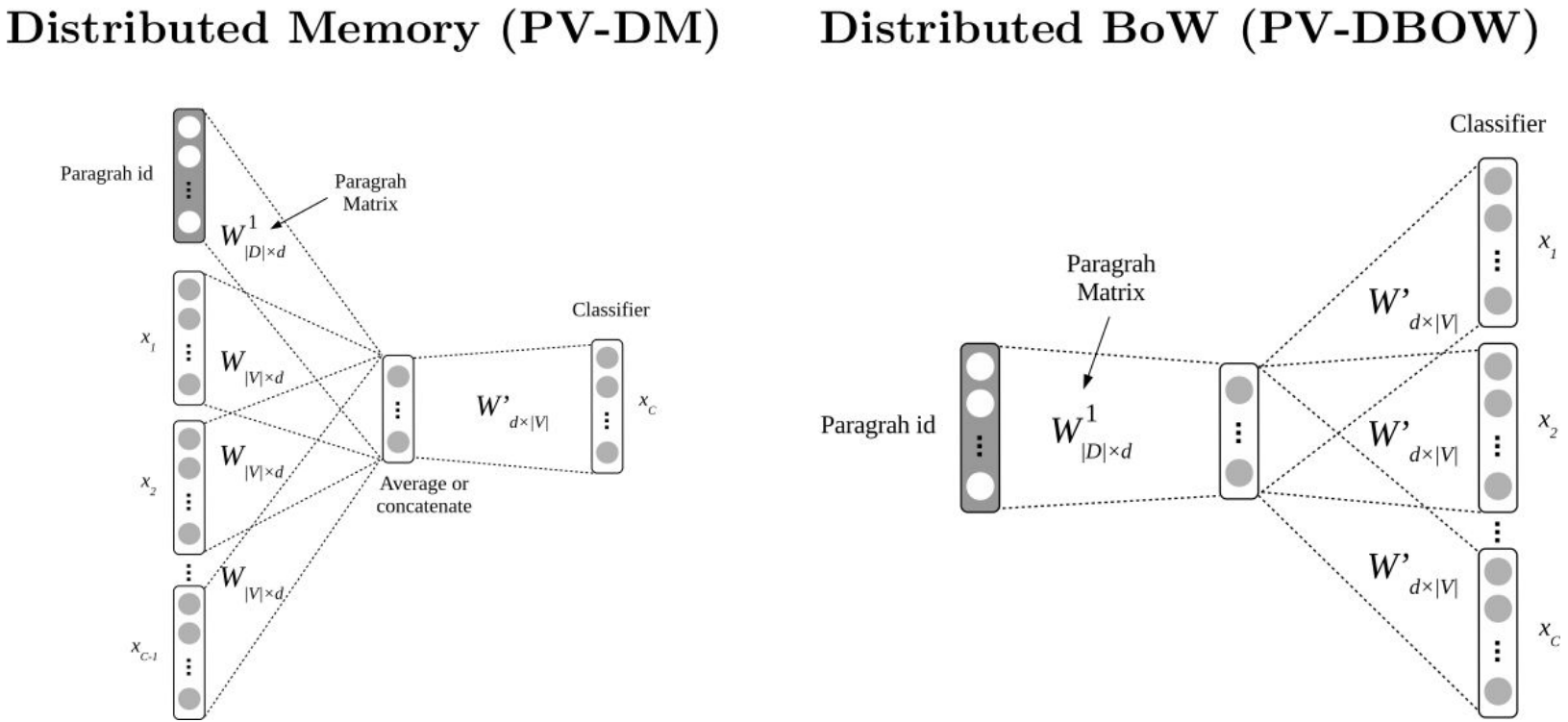
Vetores de Palavras

Média de vetores de palavras

Representação de documento que consiste na média dos vetores das palavras que compõem os documentos.

$$\begin{matrix} W_1 \\ \begin{bmatrix} W_{11} \\ W_{12} \\ \\ \\ W_{1n} \end{bmatrix} \end{matrix} + \begin{matrix} W_2 \\ \begin{bmatrix} W_{21} \\ W_{22} \\ \\ \\ W_{2n} \end{bmatrix} \end{matrix} + \cdots + \begin{matrix} W_n \\ \begin{bmatrix} W_{n1} \\ W_{n2} \\ \\ \\ W_{nn} \end{bmatrix} \end{matrix} = \begin{matrix} D \\ \begin{bmatrix} \frac{W_{11}+W_{21}+\ldots+W_{n1}}{n} \\ \\ \\ \frac{W_{1n}+W_{2n}+\ldots+W_{nn}}{n} \end{bmatrix} \end{matrix}$$

Paragraph Vector



Medidas de similaridade textual

Informam o nível de proximidade, ou distância, entre dois textos.

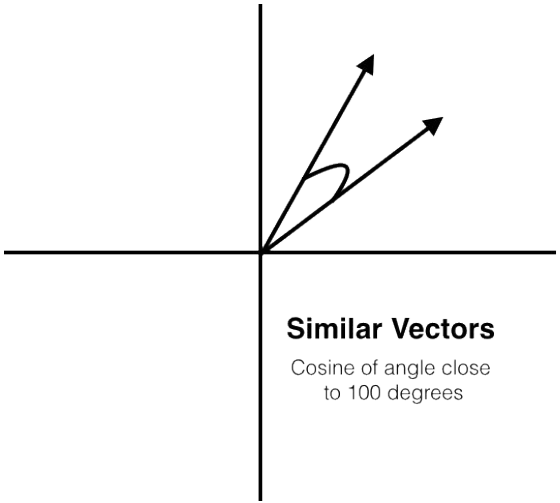
Principais modelos:

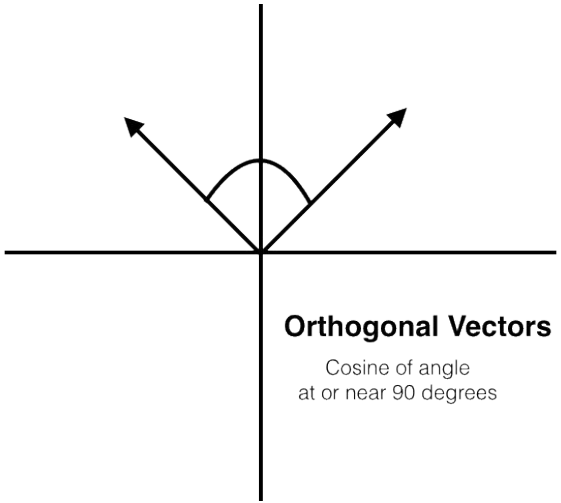
- **Similaridade do cosseno**
- Distância Euclidiana
- Similaridade de Jaccard
- Distância de Manhatan
- Coeficiente de Dice

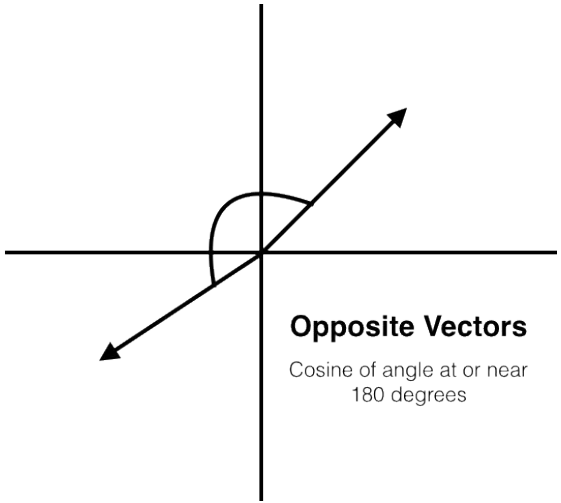
Medidas de similaridade textual

Similaridade do cosseno

$$similarity = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|_2 \|\mathbf{B}\|_2} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$







Medidas de similaridade textual

<Prática>