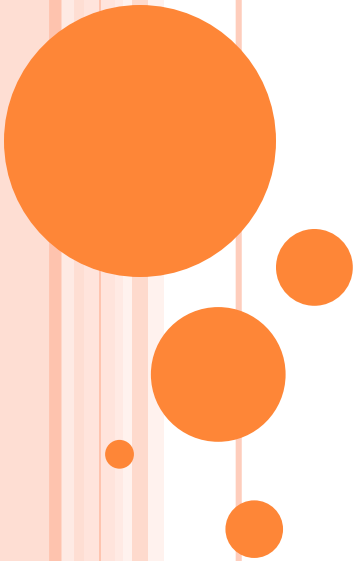


Python para PLN

Introdução à *Word Embeddings*

Roney Lira de Sales Santos roneysantos@usp.br
Prof. Thiago A. S. Pardo

Introdução à *Word Embeddings*

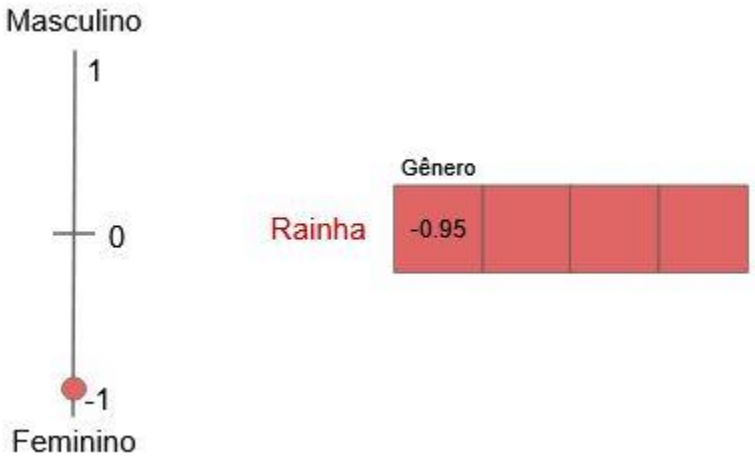
- Hipótese distribucional
  - Palavras tem significados parecidos quando são usadas em contextos parecidos.

- Modelos de linguagem
  - Predizem a próxima palavra, dado um conjunto de palavras
  - Exemplo: "O gato corre atrás do _____"
    - Qual a próxima palavra? "rato"? "cachorro"? "carro"?
  - Os modelos de linguagem são usados para tarefas como processamento de voz, autocorreção de ortografia, etc.

INTRODUÇÃO À *WORD EMBEDDINGS*

- *WORD EMBEDDING*: representação vetorial de uma palavra.
  - texto -> números

- Exemplo: Definição da palavra "rainha" com uma escala "Gênero", que vai de -1 a 1: quanto mais perto de -1, mais feminina:



Fonte: https://medium.com/turing-talks/word-embedding-fazendo-o-computador-entender-o-significado-das-palavras-92fe22745057

Introdução à *Word Embeddings*

- Porém, só com a informação sobre gênero não é possível representar bem a palavra...

- Podem ser adicionadas várias outras dimensões, ou quadradinhos, com a escala que a palavra tem mais a ver.

  - No exemplo anterior, imagine "rainha" e "rei" em escalas de "Realeza", "Fruta" e "Violência", por exemplo:

	Rainha	Rei
Gênero	-0.95	0.789
Realeza	0.89	0.96
	...	...
Fruta	0.015	-0.05
Violência	0.56	0.8

Fonte: https://medium.com/turing-talks/word-embedding-fazendo-o-computador-entender-o-significado-das-palavras-92fe22745057

Introdução à *Word Embeddings*

- E como esses valores são atribuídos?
- A partir de aprendizado de máquina!
  - Usa-se algum algoritmo para gerar, a partir do seu contexto.
- E o tamanho ideal do vetor, ou seja, a quantidade de dimensões?
  - Depende do seu corpus/*dataset* de treinamento: quanto menor, menos dimensões
  - Geralmente é um valor entre 100 e 1000.
- Um algoritmo muito utilizado para obter as *word embeddings* é chamado **Word2Vec**.

Fonte: https://medium.com/turing-talks/word-embedding-fazendo-o-computador-entender-o-significado-das-palavras-92fe22745057

Similaridade do Cosseno

- O cálculo da similaridade é feito por meio da medida do cosseno

$$scos(\vec{f},\vec{v}) = \frac{\vec{f}\cdot\vec{v}}{\left|\vec{f}\right|\left|\vec{v}\right|} = \frac{\sum_{i=1}^{n} f_i v_i}{\sqrt{\sum_{i=1}^{n} f_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}$$

  - Intervalo [0-1], onde 0 representa vetores completamente diferentes e 1 representa vetores completamente similares.