

SHIFT
FIAP

Ana Raquel



Carreira

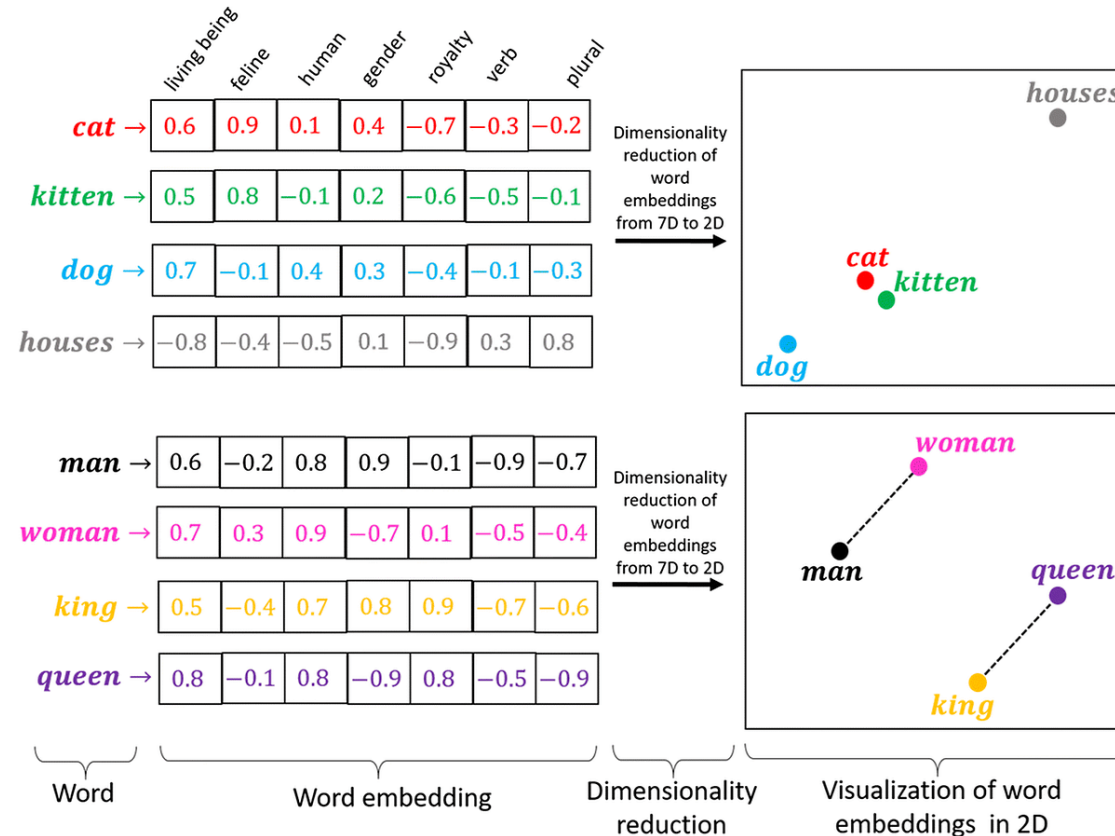
- Tecnólogo em banco de dados pela faculdade FIAP.
- MBA em inteligência artificial pela FIAP.
- Mais de 8 anos de experiência como profissional na área de dados tendo atuado em diversos projetos de Banco de Dados, BI, Analytics e Data Science.
- Cientista de dados na FIAP e professora de Machine Learning , Deep Learning, Processamento de Linguagem Natural e Data Viz na FIAP.



Word Embeddings **Como compreender** **contextos?**

O que é a técnica de word embeddings?

Podemos definir a técnica de word embeddings como a transformação de palavras em arrays (vetores).



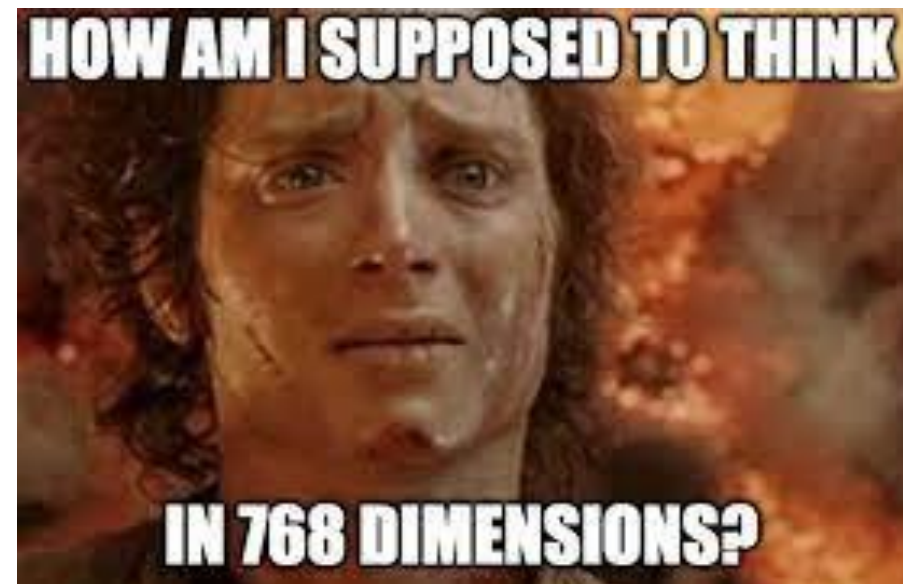
**Mas... as técnicas
TF-IDF e BOW já não
fazem isso?**

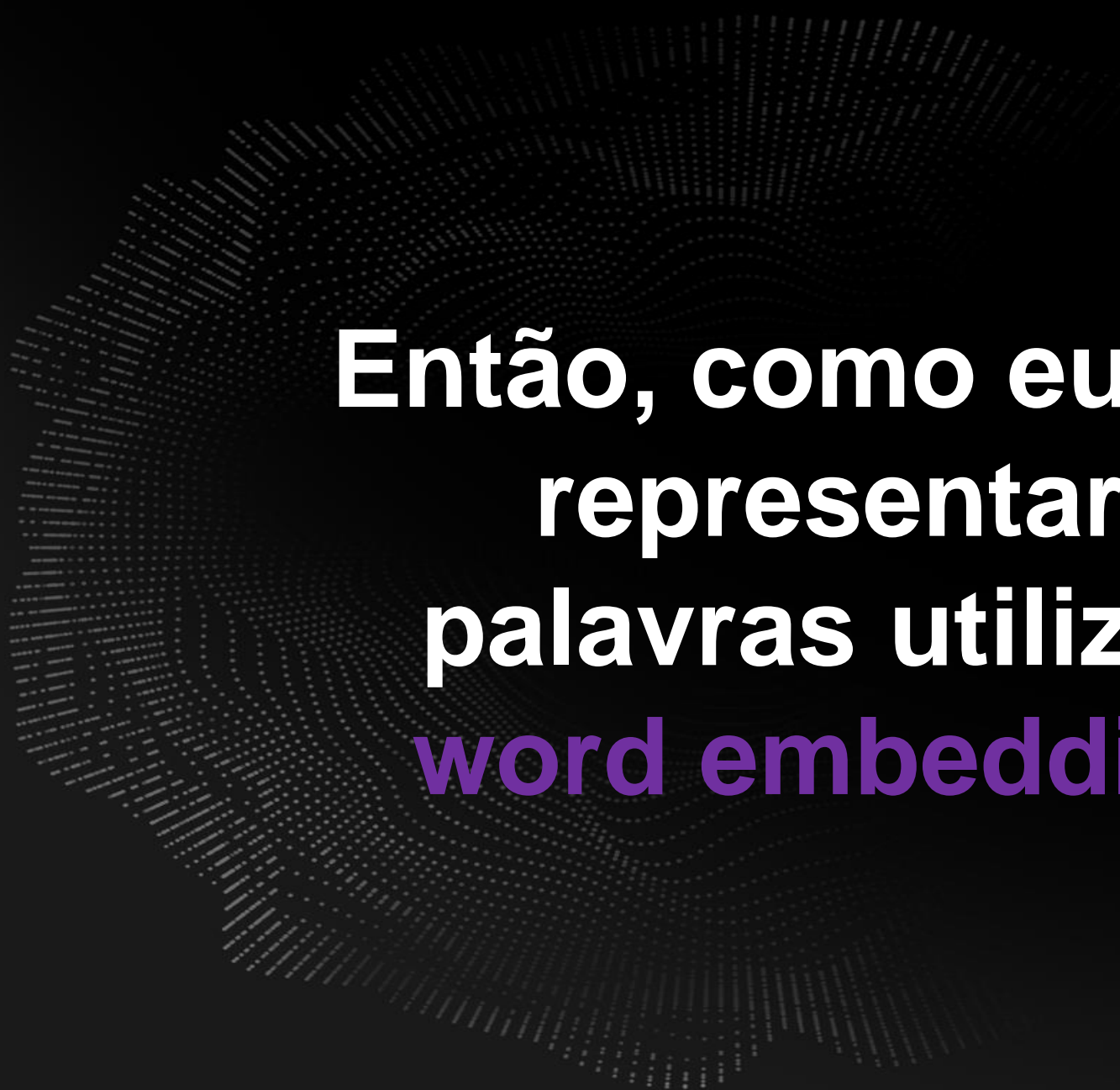


Sim! Mas, posso listar algumas limitações...

Os vetores obtidos terão o **mesmo tamanho do vocabulário**, o que se torna um problema quando temos **vocabulários muito grandes**.
Problema da dimensionalidade dos dados!

Não captura contexto e significado, isto é, extraem pouca informação semântica e sintática dos textos.





**Então, como eu posso
representar as
palavras utilizando
word embeddings?**

A técnica de word embeddings

Como já foi mencionado antes, essa técnica consiste em **transformar palavras em vetores**, **permitindo** que o computador processe o **significado semântico** das palavras.

Observação: significado semântico é o significado de uma palavra associado a uma palavra, frase ou sentença que é derivado do seu uso em um determinado contexto.

Por exemplo:

Palavras

- Banco
- Dados
- Flor

Esses vetores numéricos representam a semântica da palavra com **base em sua relação com outras palavras em um corpus de treinamento**. Palavras semanticamente semelhantes têm vetores de embedding semelhantes.

A técnica de word embeddings

Maria é uma flor de menina!

| | Flor |
|---------|------|
| Gênero | 0,95 |
| Cama | -98 |
| Garrafa | -97 |
| Jardim | 0,89 |
| Menina | 0,98 |
| Laranja | 0,45 |

Flor está altamente correlacionada com Gênero.

Flor tem baixíssima correlação com Cama.

Flor tem baixíssima correlação com Garrafa.

Flor está altamente correlacionada com Jardim.

Flor está altamente correlacionada com Menina

Flor tem uma correlação média com Laranja.

A técnica de word embeddings

Como já foi mencionado antes, essa técnica consiste em **transformar palavras em vetores**, **permitindo** que o computador processe o **significado semântico** das palavras.

Perceba que essa técnica supera as **duas dificuldades que tínhamos citado anteriormente**, pois **nós que escolhemos quantas dimensões** o embedding terá.

Qual tamanho devo escolher?

Depende muito da sua base de dados! Quanto menos, menor será o tamanho das possibilidades de palavras.

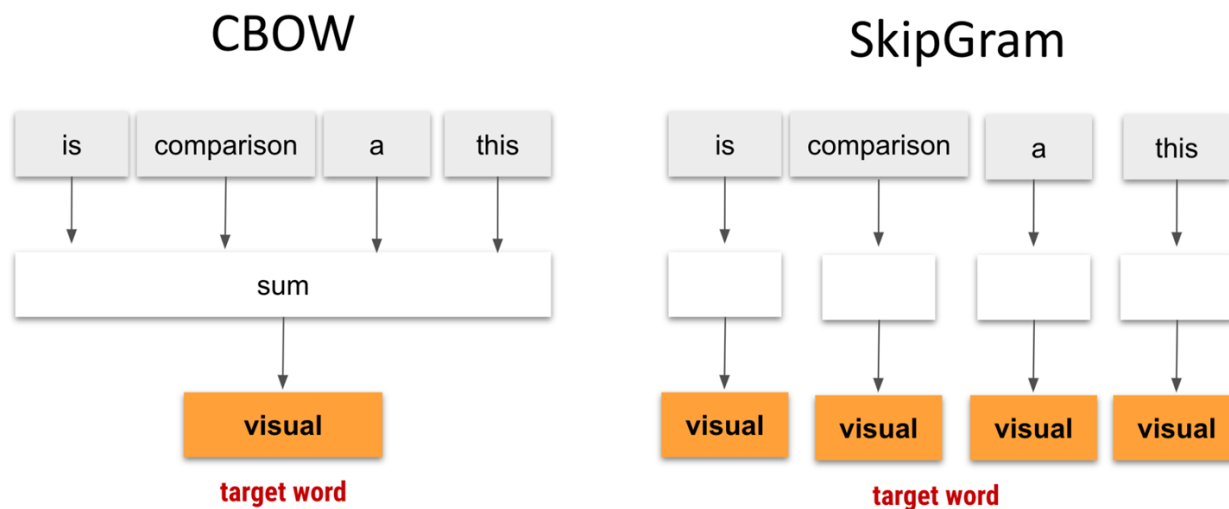
Mas como eu consigo aprender a construir esses vetores?



Word2Vec

Introdução ao Word2Vec

Word2vec é um **algoritmo para obter word embeddings treinando uma rede neural de apenas uma camada de neurônios** (Não se preocupe, você irá aprender redes neurais no próximo módulo!) O modelo word2vec é considerado “semi supervisionado”.



By: Kavita Ganesan

This is a visual comparison

Arquitetura CBOW

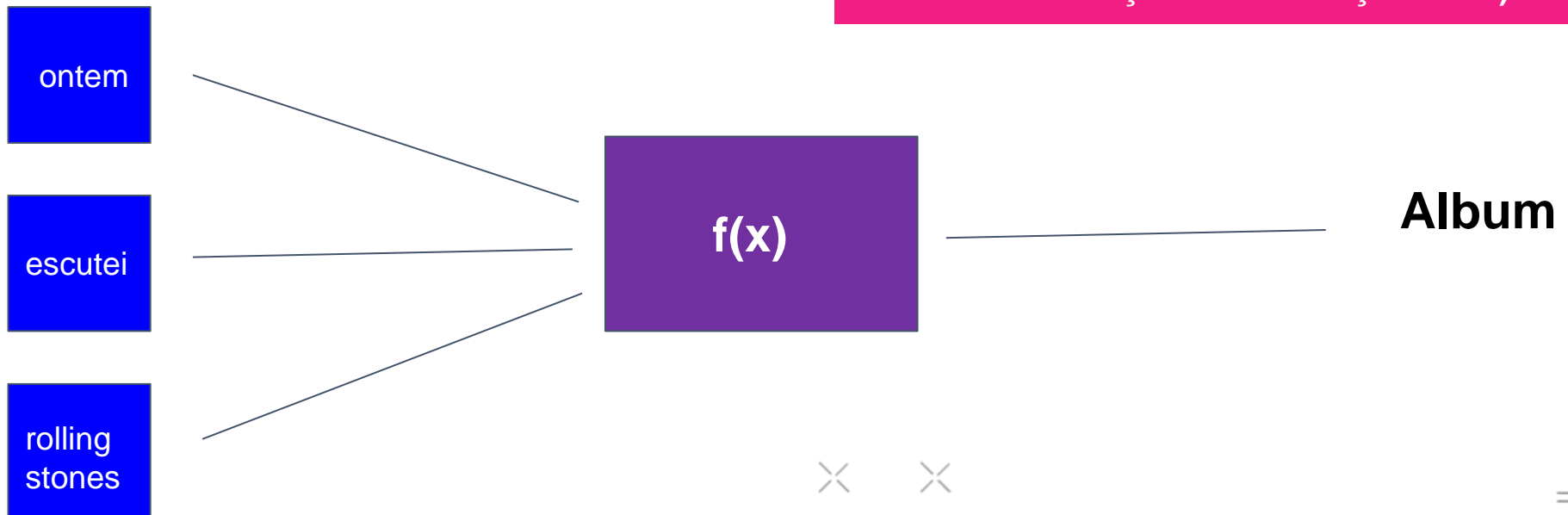
A arquitetura CBOW (Continuous Bag of Words) é um modelo de processamento de linguagem natural usado para **prever uma palavra de destino com base em um conjunto de palavras de contexto** em torno dela.

Ontem eu escutei o _____ dos Rolling Stones.

Frase de entrada

Ontem escutei _____ Rolling Stones

Aplicação de pré-processamento de texto (stop words, remoção de acentuação e etc.)



Arquitetura Skip-Gram

A arquitetura Skip-gram é um modelo de processamento de linguagem natural usado para **prever um contexto, dada uma palavra de entrada.**

Album.

Palavra de entrada



Aplicação de pré-processamento de texto (stop words, remoção de acentuação e etc.)

Album

$f(x)$

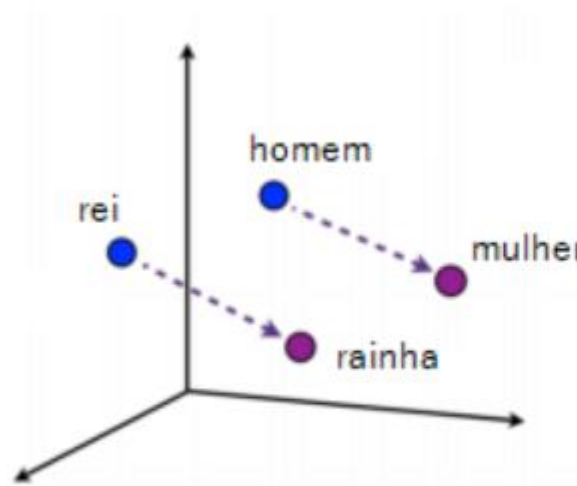
ontem

escutei

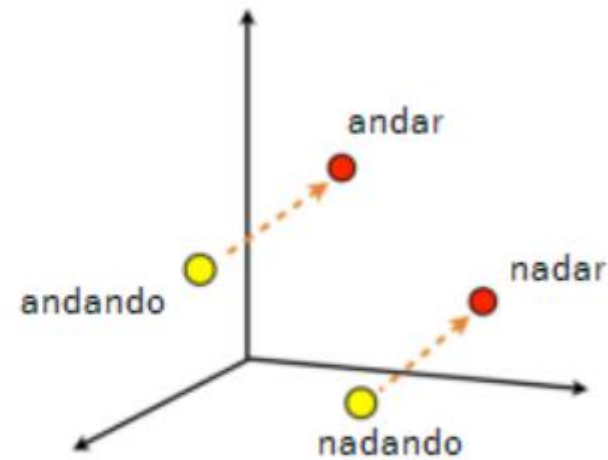
rolling
stones

Introdução ao Word2Vec

Com a rede treinada, podemos extrair os embeddings da matriz de pesos da camada oculta da rede neural. As relações entre as palavras é dada por semelhança entre as palavras dentro de cada vetor.



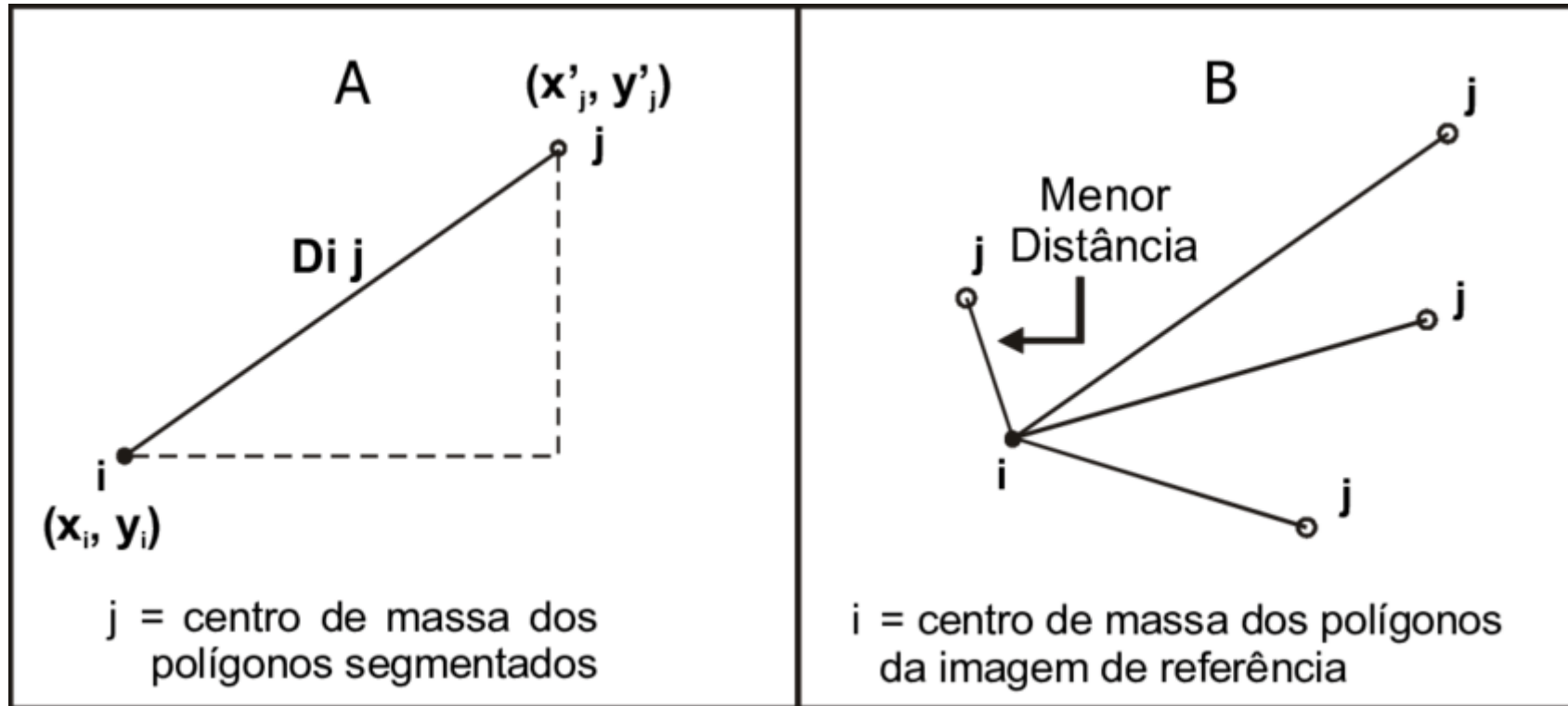
Gênero



Conjugação verbal

Como é encontrada a semelhança entre os vetores?

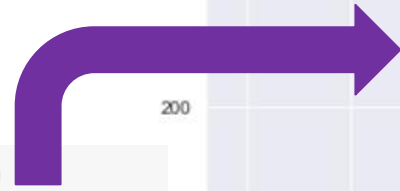
Uma das formas mais clássicas é a distância euclidiana! A distância com menor trajeto, é considerada as palavras mais próximas da palavra buscada por similaridade.



Como é na prática:

```
modelo.wv.most_similar(positive=["lisa"])
```

```
[('bart', 0.8244233131408691),  
 ('abe', 0.7787426710128784),  
 ('saxophone', 0.7556131482124329),  
 ('sweetie', 0.743913471698761),  
 ('milhouse', 0.7406502962112427),  
 ('daughter', 0.7358652949333191),  
 ('maggie', 0.7330158352851868),  
 ('braces', 0.7248966693878174),  
 ('brother', 0.7239959239959717),  
 ('mary', 0.7219533324241638)]
```



Configurando os principais hiperparâmetros do word2vec

Vamos conhecer os principais hiperparâmetros para a construção do algoritmo word2vec:

vector_size: Configurar o tamanho do vetor das palavras.

min_count: tamanho do total de palavras raras dentro do corpus.

window: número de palavras para considerar, tanto olhando para trás quanto para frente para compreender contexto.

alpha: taxa de learning rate (vamos aprender com mais detalhes sobre essa parte no próximo módulo)

min_alpha: A taxa de aprendizado cairá linearmente para min_alpha conforme o treinamento progride.

epochs: Número de iterações (epochs) no corpus.

Obrigada!

Ana Raquel



[linkedin.com/ana-raquel-fernandes-cunha](https://www.linkedin.com/ana-raquel-fernandes-cunha)

Copyright © 2023 | Ana Raquel Fernandes Cunha

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento é expressamente proibido sem consentimento formal, por escrito, do professor/autor.