


WORD EMBEDDING

HELLO!

Ricardo Vitor Costa Neto

Github: https://github.com/Ricardovcn

Tópicos Abordados

1. **Word Embedding**

   Introdução teórica básica

2. **Pré-processamento de Textos**

   Essencial no processamento de linguagem natural.

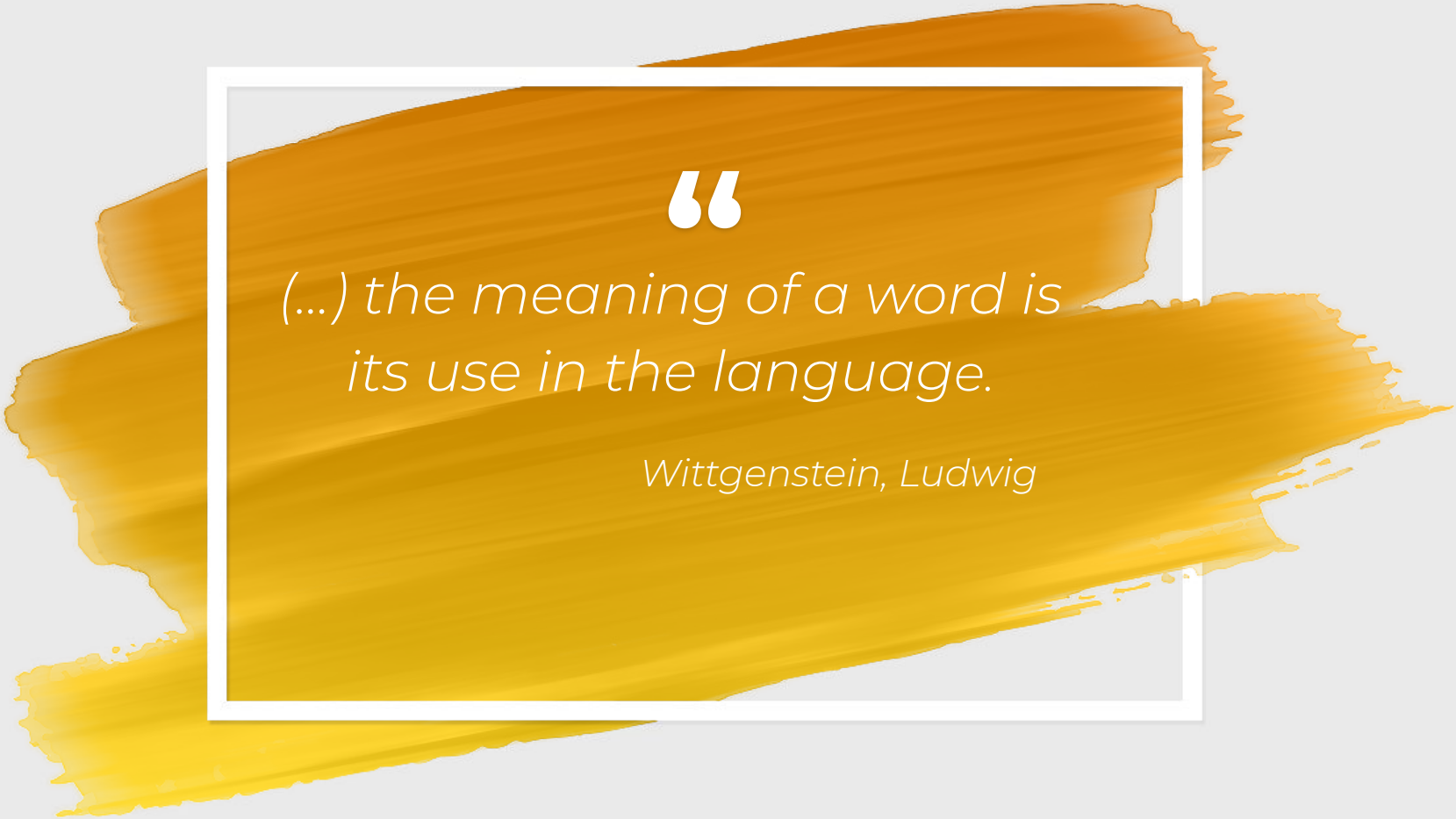
3. **Word2vec**

   Entendimento e utilização do modelo word2vec para fazer word embedding de textos.

4. **Word Mover's Distance**

   Técnica utilizada para comparar sentenças ou documentos.

INTRODUÇÃO


“
(...) the meaning of a word is its use in the language.
Wittgenstein, Ludwig

PROCESSAMENTO DE LINGUAGEM NATURAL

Subárea da Inteligência Artificial que estuda a capacidade e as limitações de uma máquina em entender a linguagem dos seres humanos.

PROCESSAMENTO DE LINGUAGEM NATURAL

“Entender” um texto significa reconhecer o contexto, fazer análise sintática e semântica, criar resumos, extrair informação, interpretar os sentidos, analisar sentimentos e até aprender conceitos com os textos processados.

EMBEDDING

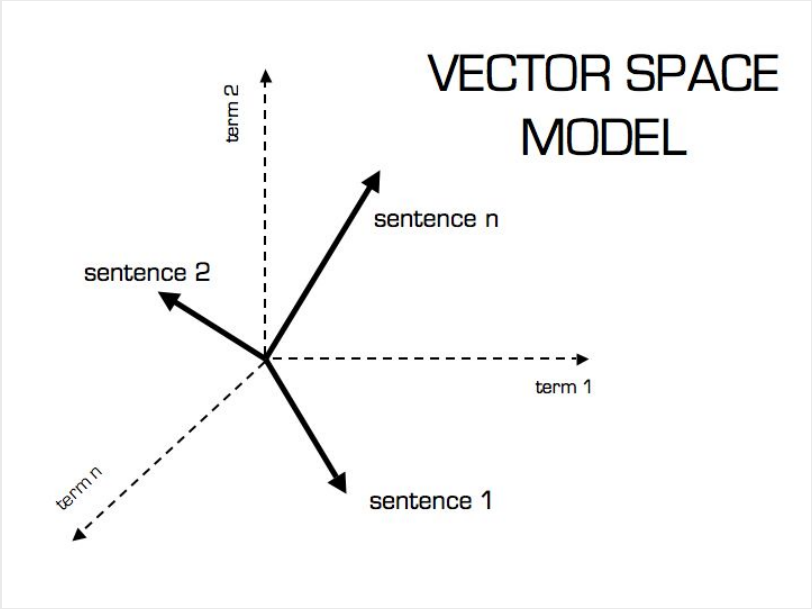
Conceito matemático que consiste em converter um objeto matemático em outro objeto matemático, preservando sua estrutura.

EMBEDDING

Transformar um objeto com alta dimensionalidade em um objeto com uma menor dimensionalidade, preservando sua estrutura.

ESPAÇO VETORIAL

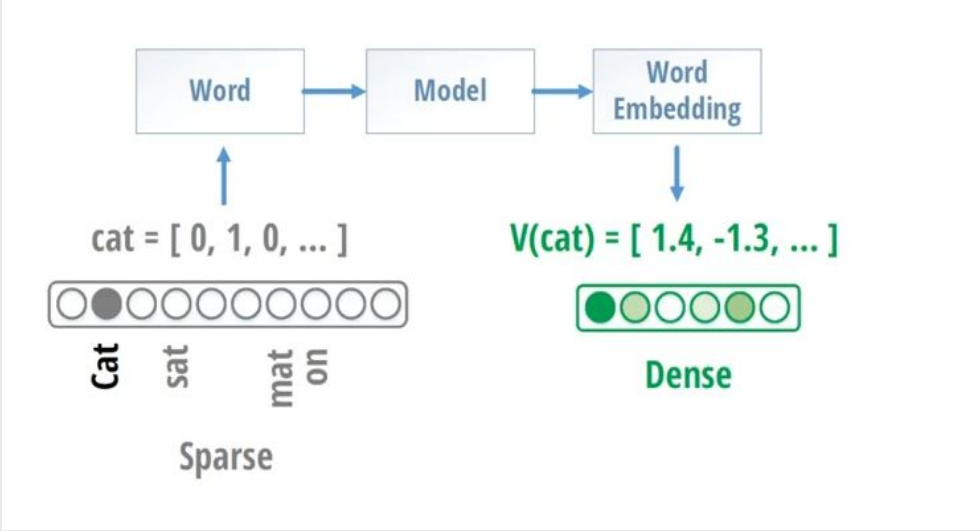
Modelo matemático que consiste na representação vetorial de variáveis categóricas.



WORD EMBEDDING

Modelos e técnicas de NPL onde as palavras ou sentenças de um documento são representadas em um espaço vetorial denso e de dimensionalidade reduzida.

WORD EMBEDDING

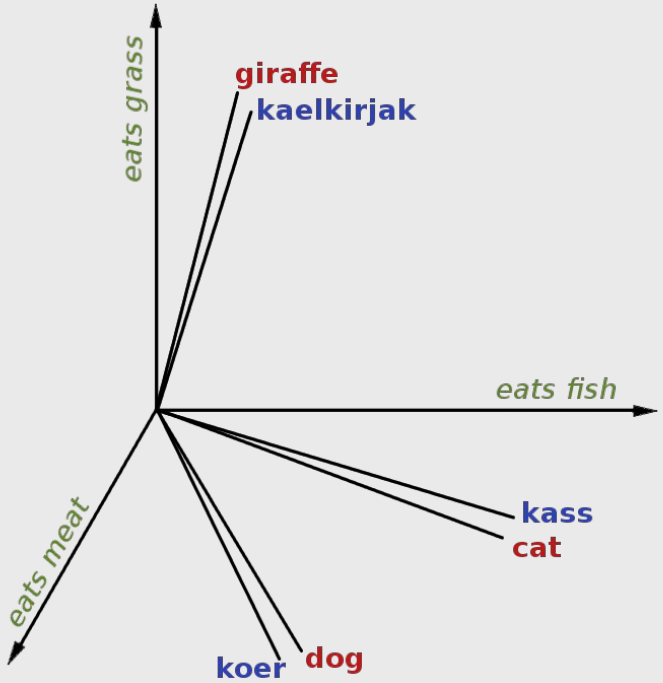


One Hot Encoding

The cat sat on the mat.

cat	**1**	**0**	**0**	**0**
sat	**0**	**1**	**0**	**0**
mat	**0**	**0**	**1**	**0**
on	**0**	**0**	**0**	**1**

WORD EMBEDDING

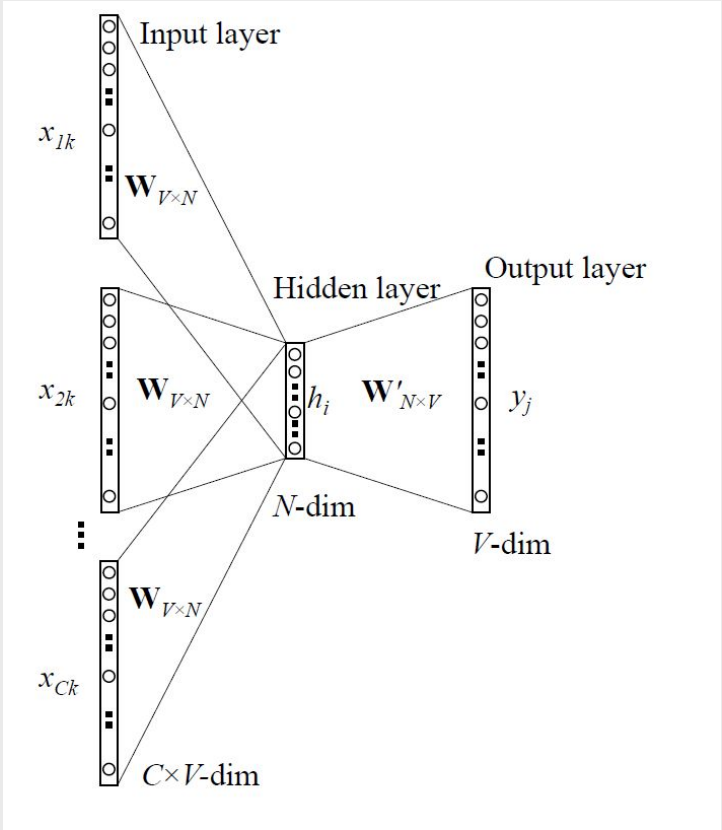


WORD2VEC

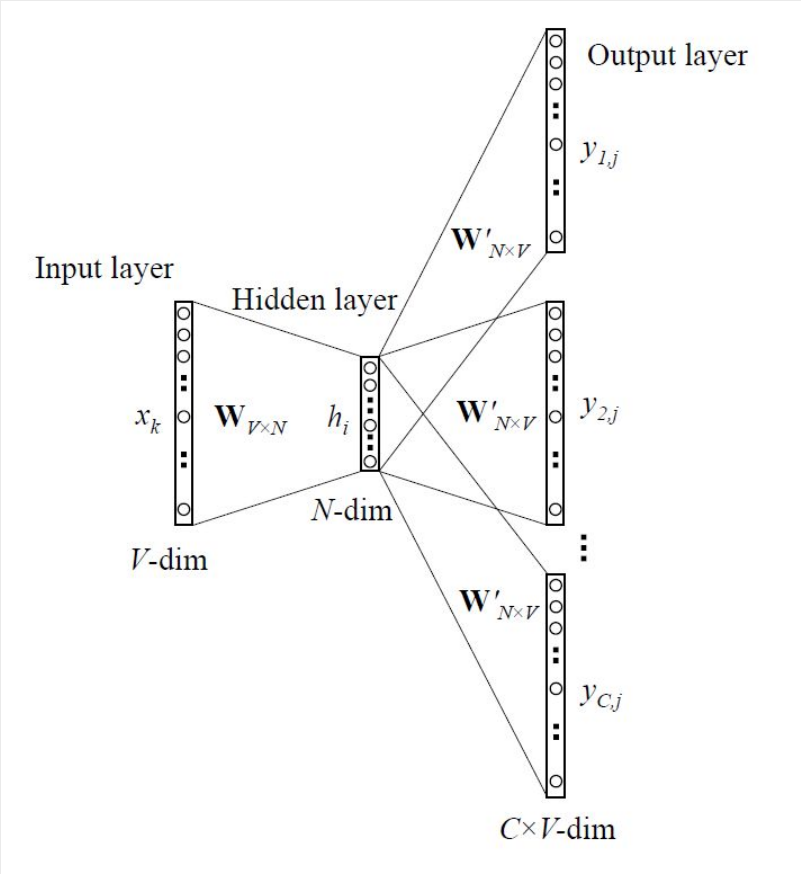
WORD2VEC

Word2vec é um par de modelos de aprendizado não supervisionado para criação de uma representação vetorial de palavras presentes em textos que usam linguagem natural.

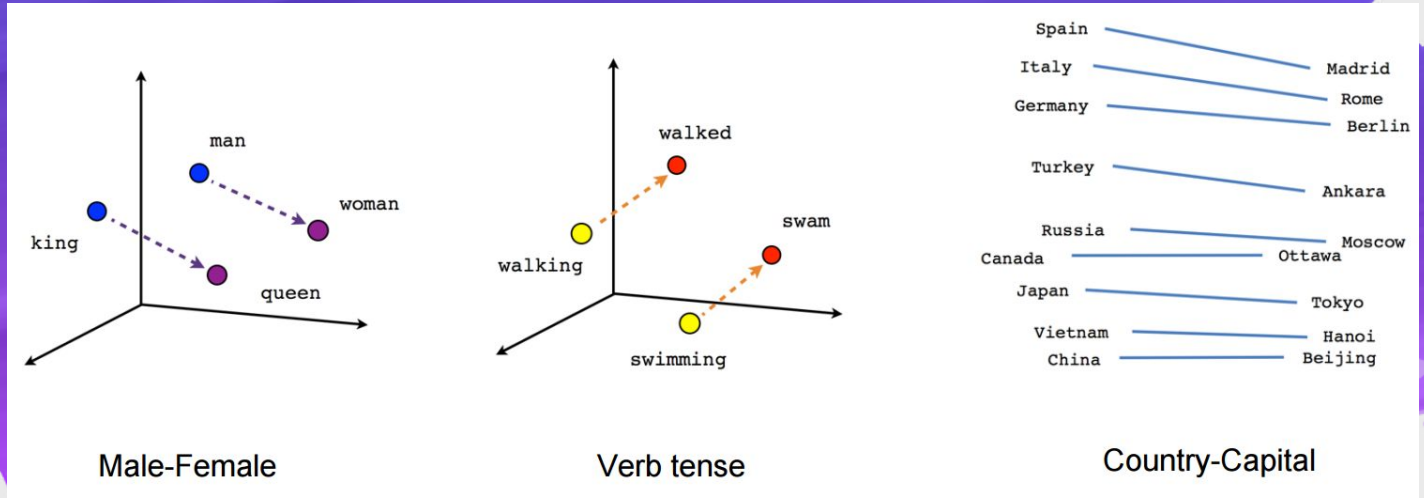
CBOW


Input layer
x_{1k}
$\mathbf{W}_{V\times N}$
Hidden layer
Output layer
x_{2k}
$\mathbf{W}_{V\times N}$
h_i
$\mathbf{W}'_{N\times V}$
y_j
N-dim
V-dim
$\mathbf{W}_{V\times N}$
x_{Ck}
$C\times V$-dim

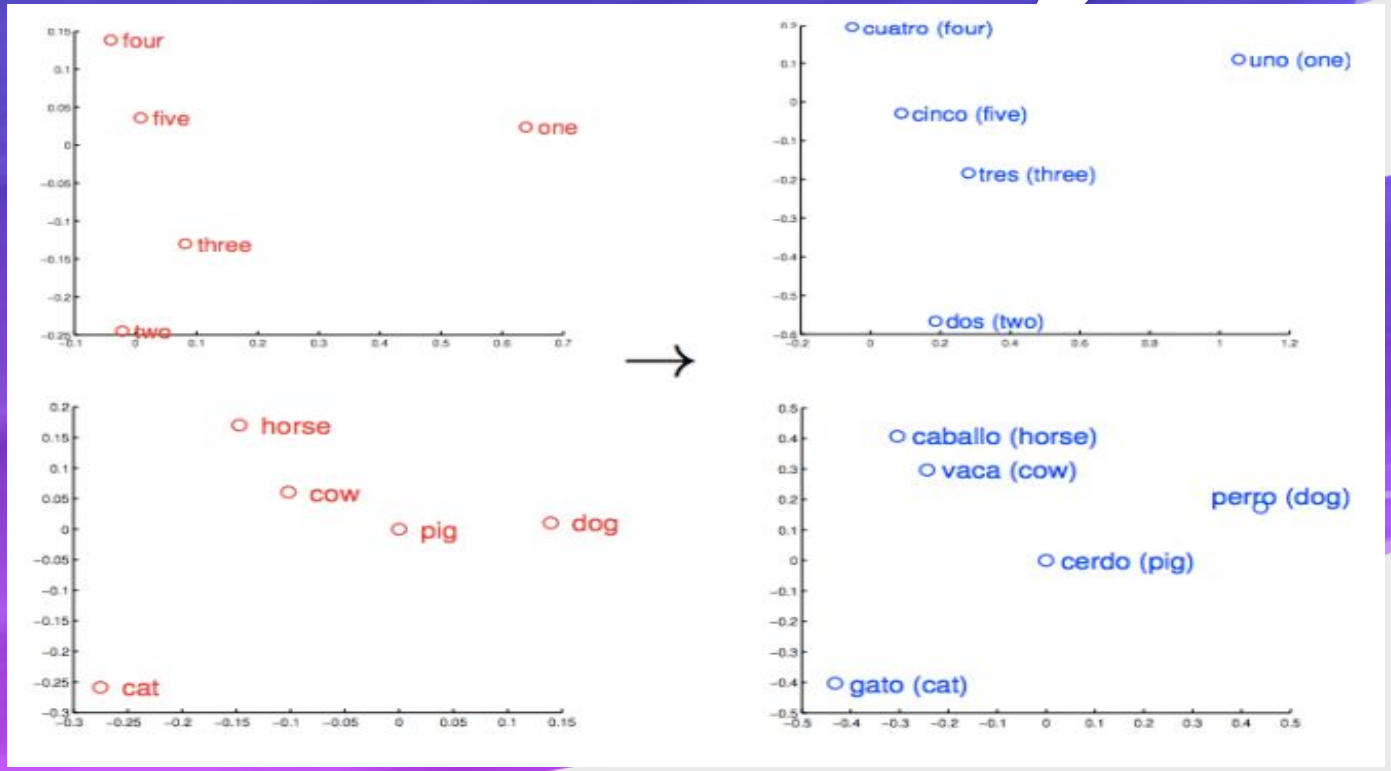
SKIP-GRAM


Output layer
$y_{1,j}$
$\mathbf{W}'_{N\times V}$
Input layer
Hidden layer
x_k
$\mathbf{W}_{V\times N}$
h_i
$\mathbf{W}'_{N\times V}$
$y_{2,j}$
N-dim
V-dim
$\mathbf{W}'_{N\times V}$
$y_{C,j}$
$C\times V$-dim

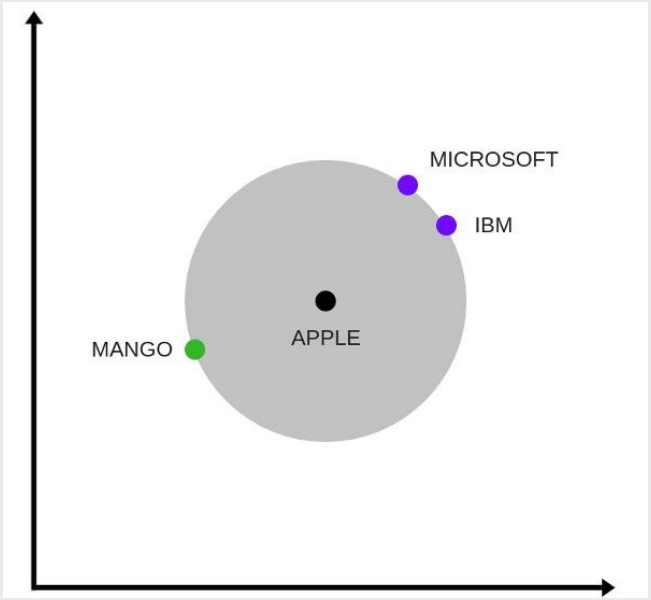
Word Embedding



Word Embedding



DESVANTAGENS DO WORD2VEC



WORD MOVER'S DISTANCE

WORD MOVER'S DISTANCE

É uma té

THANKS!

Any questions?

You can find me at @Ricardovcn &
ricardovitorcn@gmail.com

Referências

Material em Português:

https://www.youtube.com/watch?v=EVMIR6siWbI

Material em Inglês:

https://radimrehurek.com/gensim/models/word2vec.html

https://www.kaggle.com/pierremegret/gensim-word2vec-tutorial

http://kavita-ganesan.com/gensim-word2vec-tutorial-starter-code/#.W467ScBjM2x

https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

http://proceedings.mlr.press/v37/kusnerb15.pdf

https://towardsdatascience.com/word-distance-between-word-embeddings-cc3e9cf1d632