



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.

The image features a light gray background with several horizontal, overlapping brushstrokes in shades of purple and blue. These strokes have a textured, painterly appearance with visible bristles and varying opacity. A white rectangular frame is centered on the page, containing the word 'INTRODUÇÃO' in white, bold, uppercase letters. The frame is slightly offset from the edges of the brushstrokes.

INTRODUÇÃO



“

*(...) the meaning of a word is
its use in the language.*

Wittgenstein, Ludwig



EMBEDDING

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

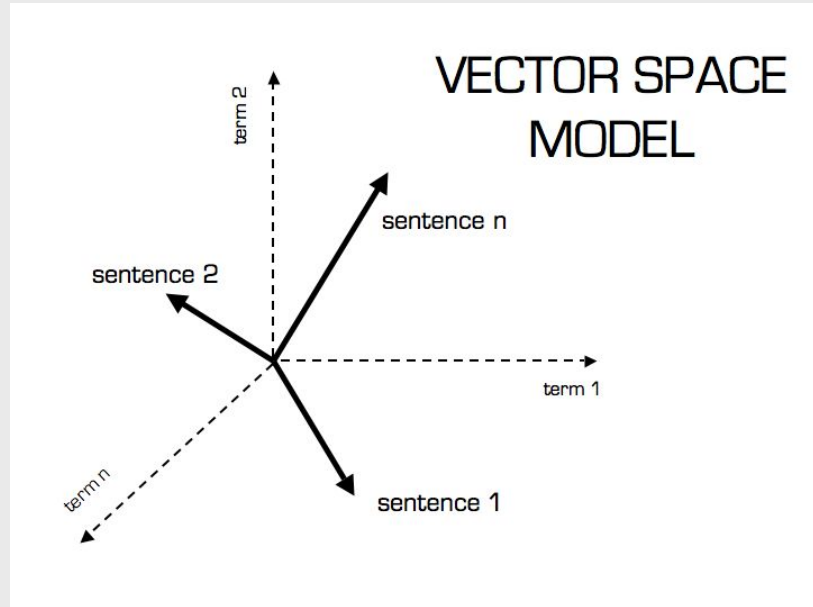


EMBEDDING

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.

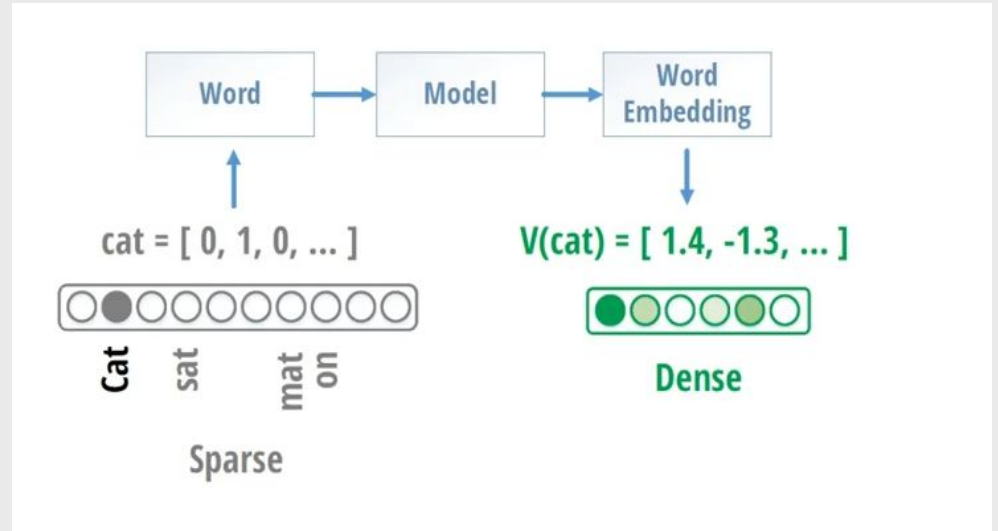




WORD EMBEDDING

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

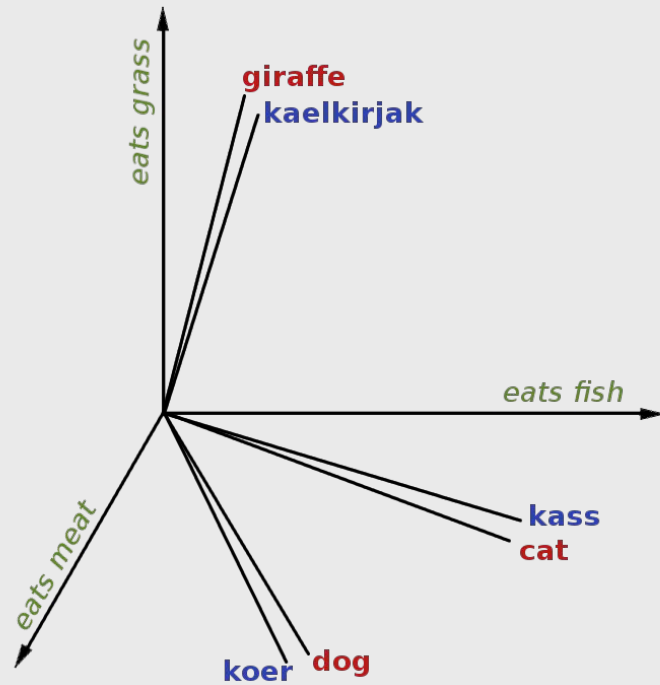


One Hot Encoding

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





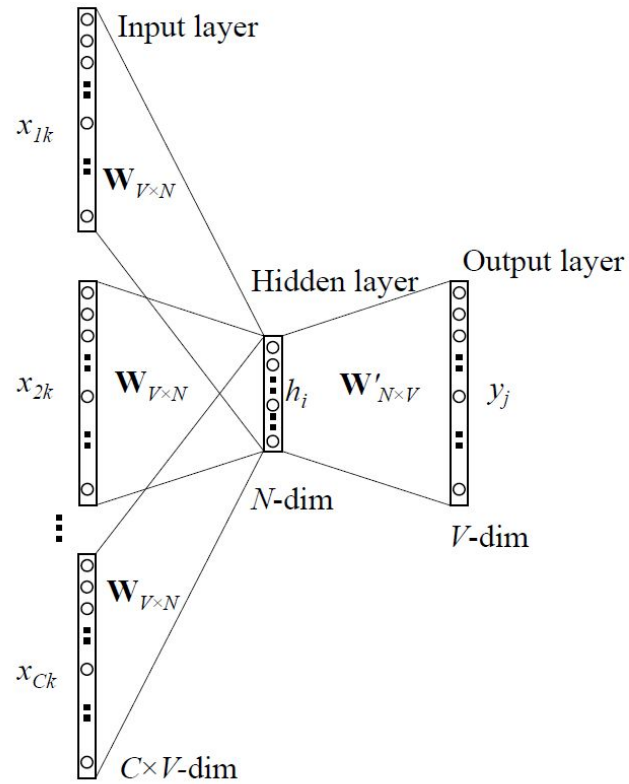
WORD2VEC



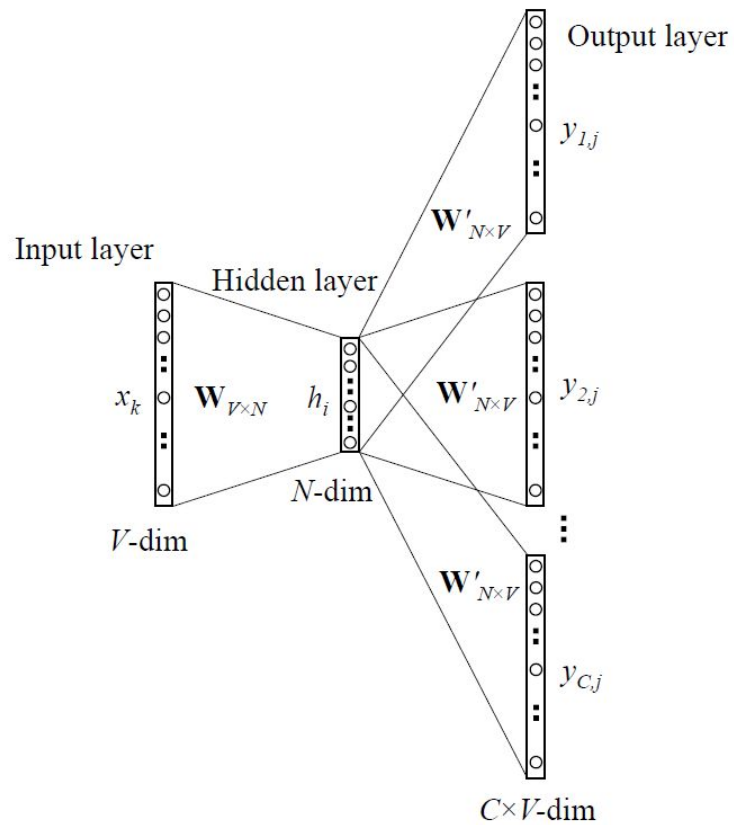
WORD2VEC

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.

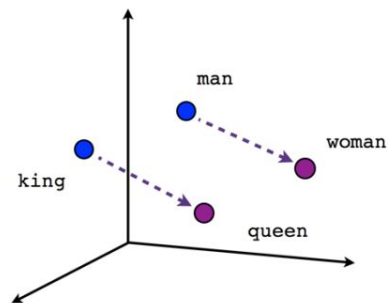
CBOW



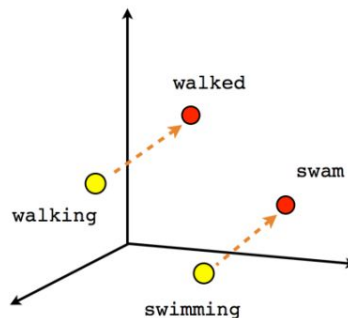
SKIP-GRAM



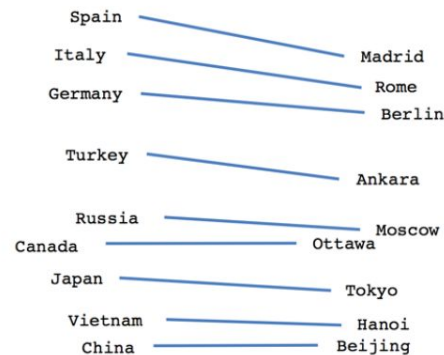
Word Embedding



Male-Female

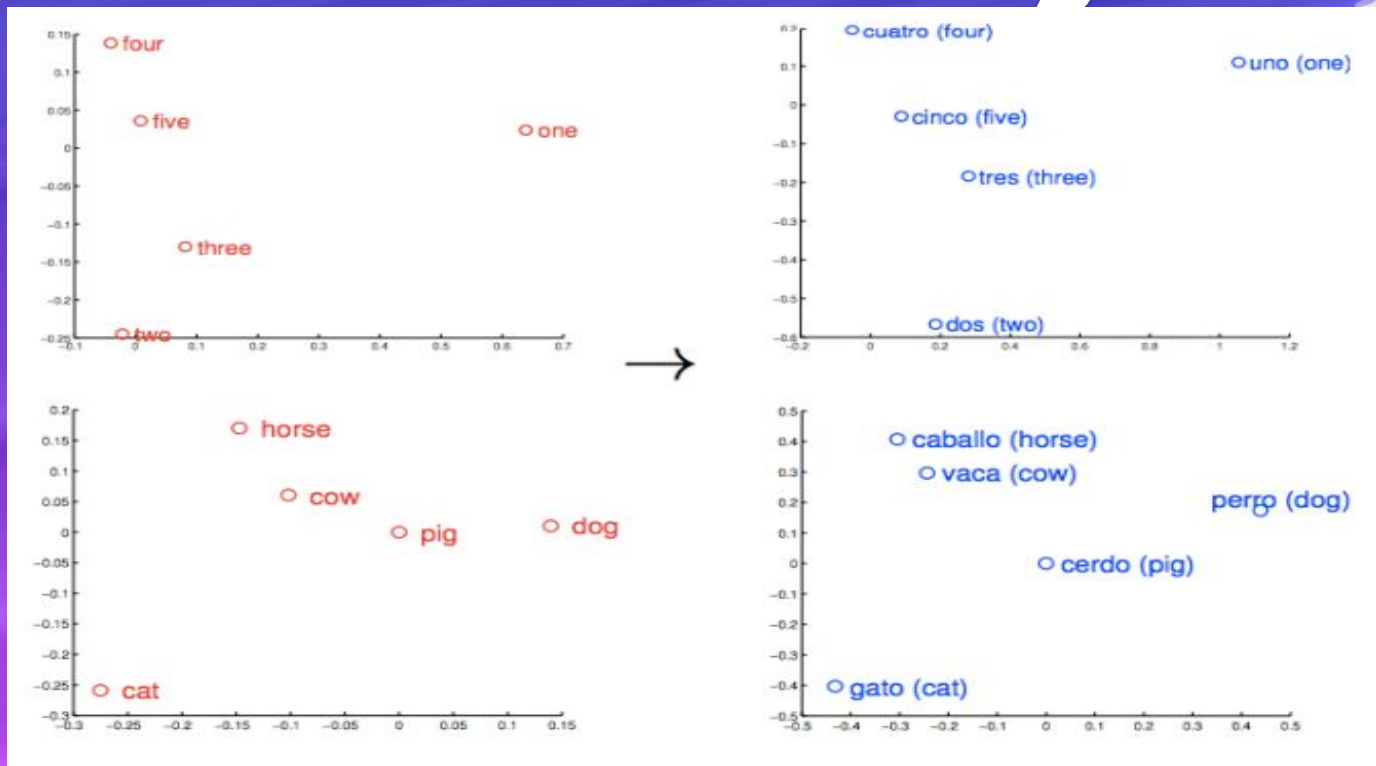


Verb tense

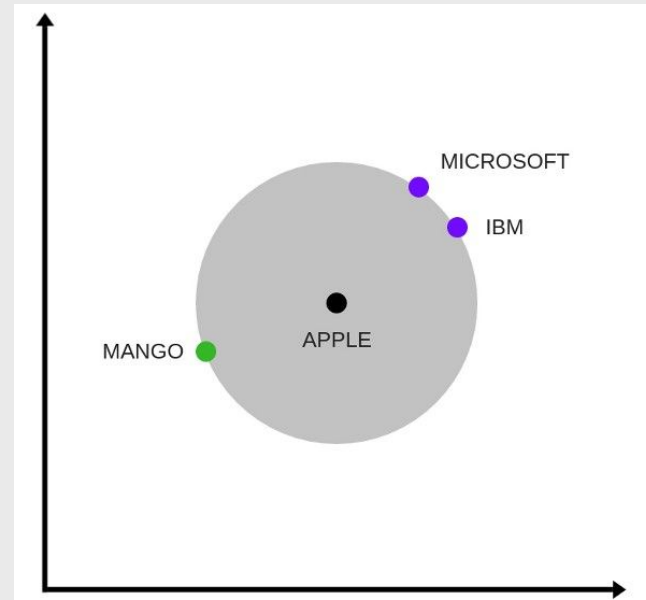


Country-Capital

Word Embedding




DESVANTAGENS DO WORD2VEC






WORD MOVER'S DISTANCE

A large, expressive purple brushstroke graphic that sweeps across the left side of the slide, partially framing the title.

WORD MOVER'S DISTANCE


É uma técnica usada para comparar sentenças/documentos a partir do resultado de um modelo de word embedding, ou seja, comparar 2 conjuntos de vetores e dizer a distância entre eles.

A large, vibrant green brushstroke graphic that sweeps across the left side of the slide, partially overlapping a white rectangular frame.

WORD MOVER'S DISTANCE


Vamos usar um exemplo clássico com 2 sentenças:

- Obama speaks to the media in Illinois
- The president greets the press in Chicago



WORD MOVER'S DISTANCE

- O primeiro passo é processar as 2 frases com as mesmas técnicas usadas para pré-processar o dataset inicial.

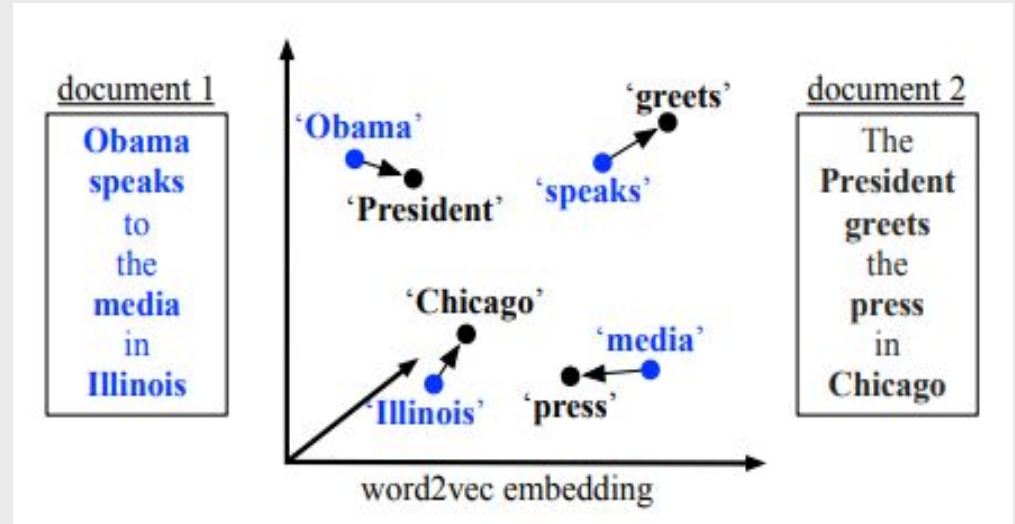


WORD MOVER'S DISTANCE

Como resultado, vamos obter as seguintes sentenças:


- obama speaks media illinois
- president greets press chicago

WORD MOVER'S DISTANCE



Fonte:

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



WORD MOVER'S DISTANCE

A distância é calculada através do somatório do menor custo possível de transferência das palavras de uma das sentenças para a outra.



THANKS!

Any questions?

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



Referências

Material em Português:

<https://www.youtube.com/watch?v=EVMIR6siWbl>

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>