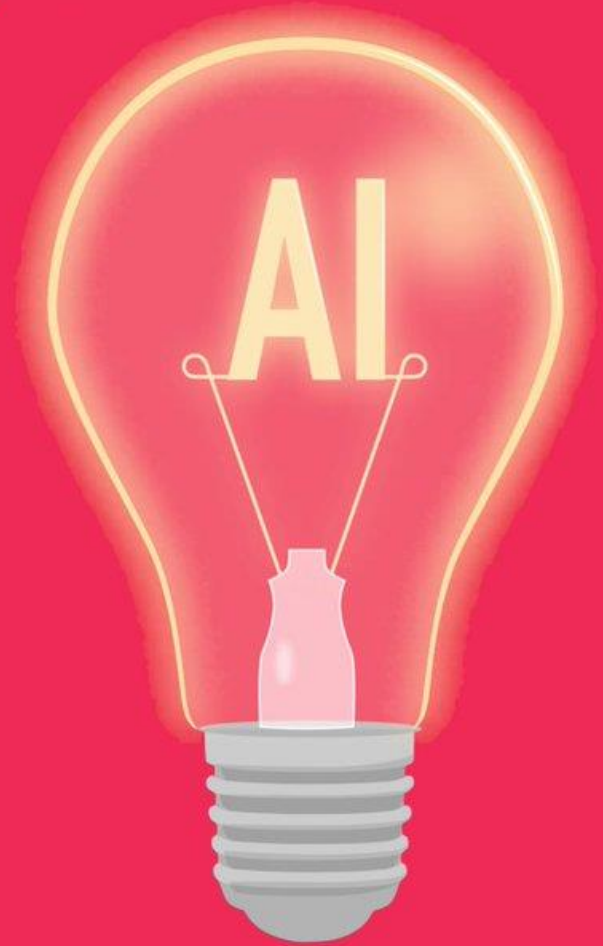


HANDS-ON PROCESSAMENTO DE LINGUAGEM NATURAL

Prof. Felipe Galiza

FIAP

Data is the new oil and artificial intelligence is the new electricity



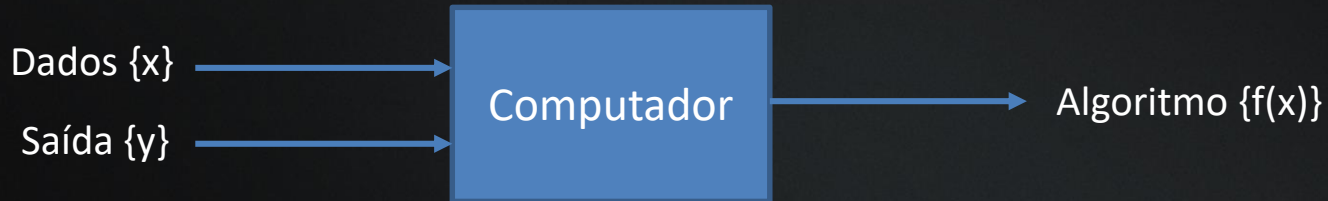
O QUE É INTELIGÊNCIA ARTIFICIAL?

Inteligência Artificial (AI) é a habilidade de uma máquina ou programa de computador aprender com dados.

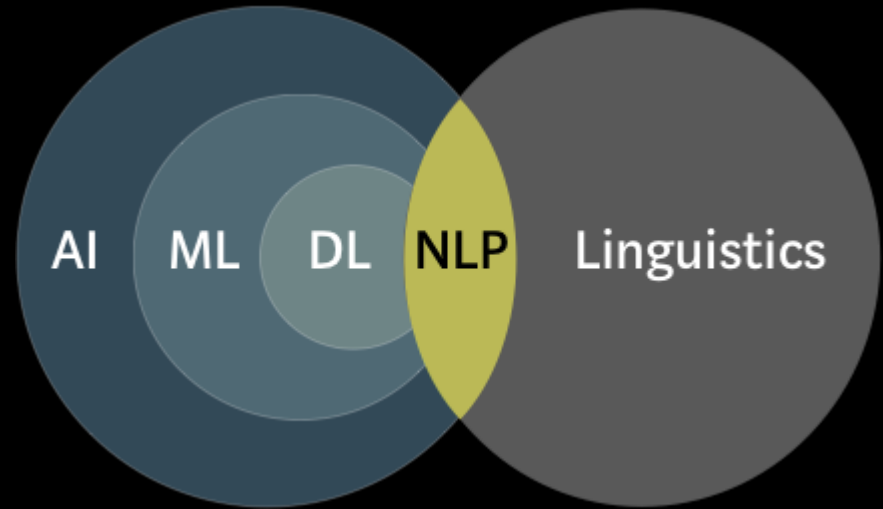
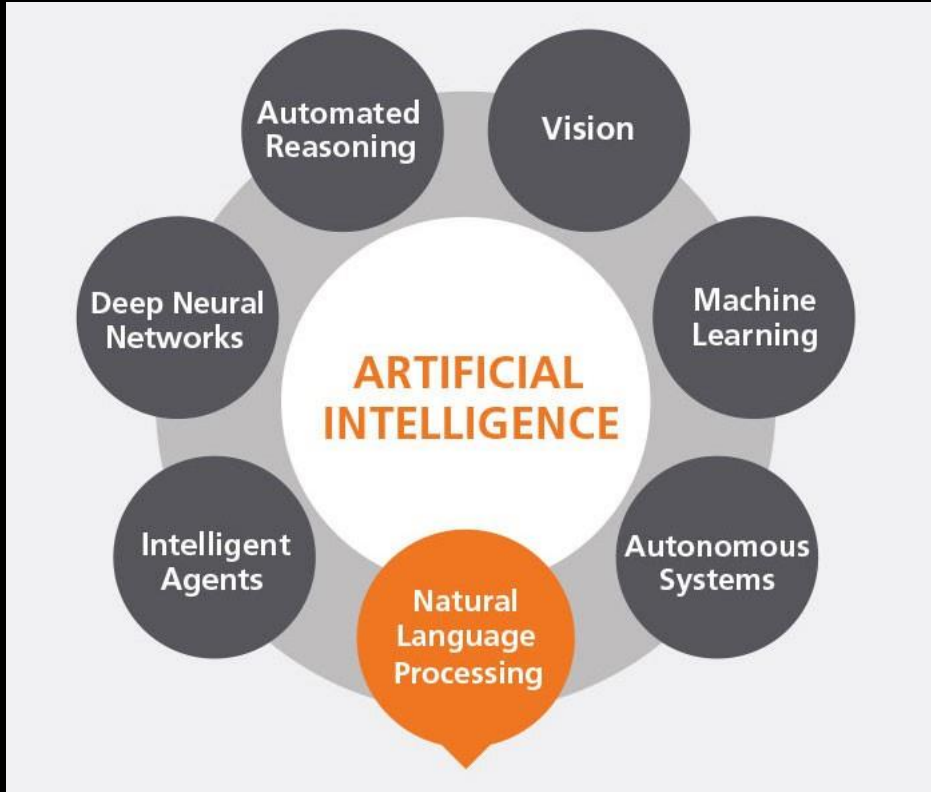
Programação Tradicional



Inteligência Artificial



O QUE É PROCESSAMENTO DE LINGUAGEM NATURAL (NLP)?



APLICAÇÕES DE

NLP

UNIVERSAL PART-OF-SPEECH (POS) TAGS

- ADJ: adjective
- ADP: adposition
- ADV: adverb
- AUX: auxiliary
- CCONJ: coordinating conjunction
- DET: determiner
- INTJ: interjection
- NOUN: noun
- NUM: numeral
- PART: particle
- PRON: pronoun
- PROPN: proper noun
- PUNCT: punctuation
- SCONJ: subordinating conjunction
- SYM: symbol
- VERB: verb
- X: other

PART-OF-SPEECH (POS) TAGGING

Text to parse

A TCS esta localizada em Alphaville.

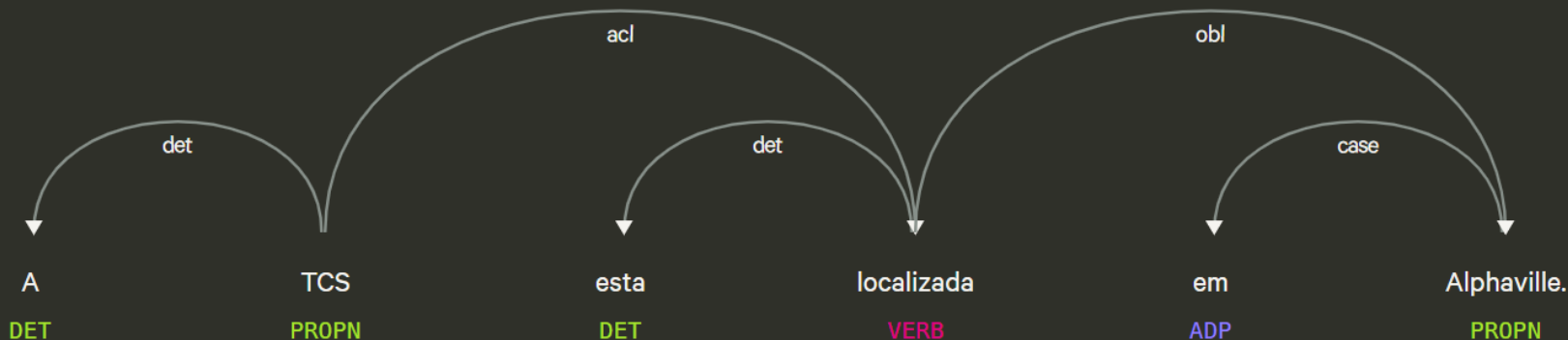


Model ⓘ

Portuguese - pt_core_news_sm (v2.0.0) ▾

☒ Merge Punctuation

☒ Merge Phrases



Source: <https://explosion.ai/demos/displacy>

RECONHECIMENTO DE ENTIDADES NOMEADAS (NER)

displaCy Named Entity Visualizer

Hoje o colaborador Felipe Galiza realizará uma apresentação na unidade da TCS que esta localizada em Alphaville



Model ⓘ

Portuguese - pt_core_news_sm (v2.0.0) ▼

Entity labels (select all)

☒ PER

☒ ORG

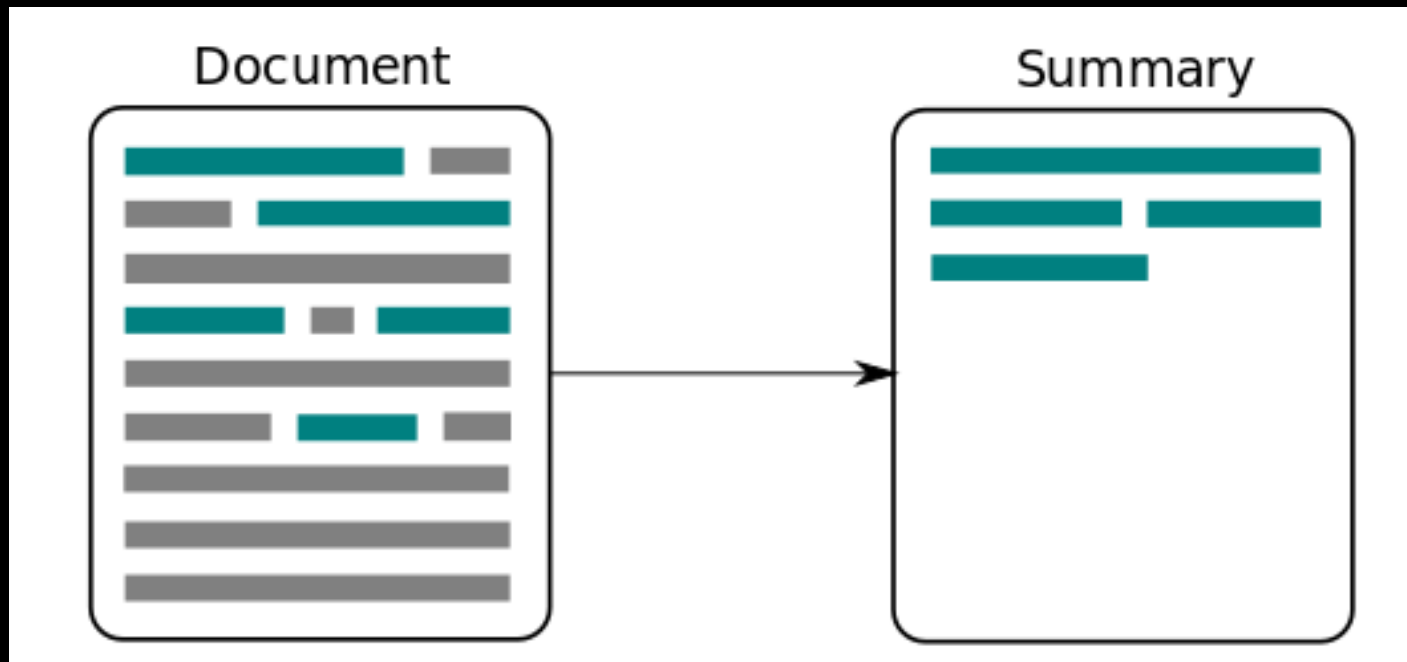
☒ LOC

☒ MISC

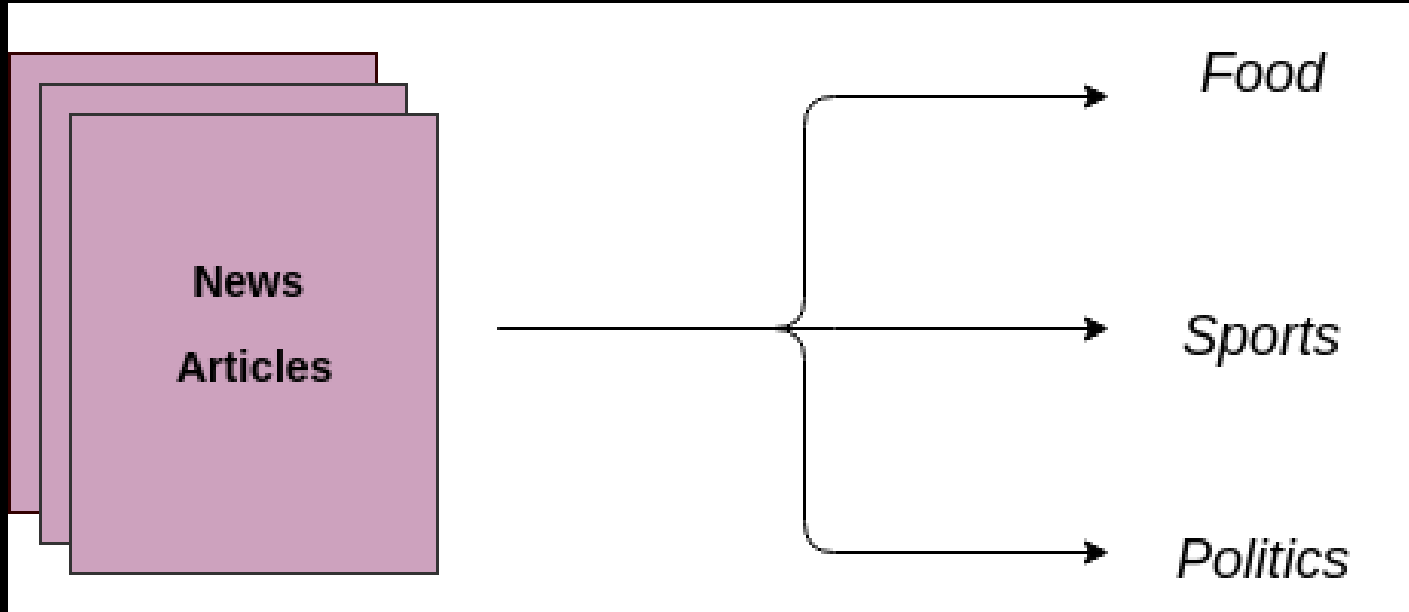
Hoje o colaborador **Felipe Galiza** **PER** realizará uma apresentação na unidade da **TCS** **ORG** que esta localizada em **Alphaville** **LOC**

Source: <https://explosion.ai/demos/displacy-ent>

RESUMO AUTOMÁTICO DE TEXTOS



CLASSIFICAÇÃO AUTOMÁTICA DE TEXTOS



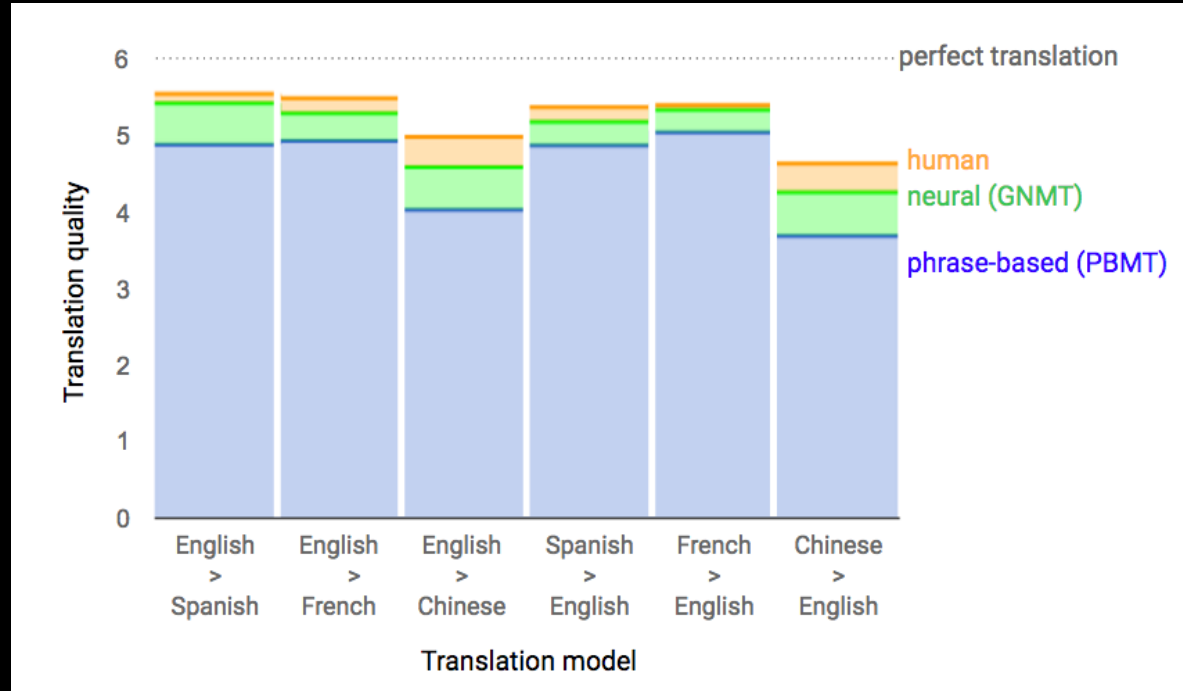
TRADUÇÃO AUTOMÁTICA

<i>Input sentence:</i>	<i>Translation (PBMT):</i>	<i>Translation (GNMT):</i>	<i>Translation (human):</i>
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

Example from Google®'s machine translation system (2016)

Source: <https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html>

TRADUÇÃO AUTOMÁTICA



Source: <https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html>

GERAÇÃO AUTOMÁTICA DE LEGENDAS



man in black shirt is playing guitar.



construction worker in orange safety vest is working on road.



two young girls are playing with lego toy.



boy is doing backflip on wakeboard.

Legendas geradas automaticamente.

Source: <https://cs.stanford.edu/people/karpathy/cvpr2015.pdf>

O QUE VAMOS FAZER HOJE?





CLASSIFICAÇÃO DE NOTÍCIAS EM JORNAIS

FERRAMENTAS OPEN SOURCE PARA NLP



spaCy



gensim

NLTK

INSTALAÇÃO DAS FERRAMENTAS PARA O HANDS-ON

- `conda create -n nlp python=3`
- `conda activate nlp`
- `pip install jupyter spacy scikit-learn pandas matplotlib`
- `python -m spacy download pt_core_news_sm`
- Faça o download do dataset: [News of the Brazilian Newspaper](#)

PROCESSAMENTO DE LINGUAGEM NATURAL (NLP)



PRÉ-PROCESSAMENTO DO TEXTO

- Transformar todas as letras para a forma minúscula, remoção de pontuação, remoção de quebras de linhas, etc
- Remoção de palavras que não adicionam informação relevante sobre o documento:

English Stop Words: **a, an, and, are, as, at, on, by**

Brazilian Portuguese Stop Words: **de, a, o, que, em, um, não, uma, por, na, mais**



Stemização e Lematização (Normalização Lexical)

- Reduz diversas formas e derivações de uma palavra para uma base comum:

Stemização: química, químicas, químico, químicos => químic

Lematização: Sou, és, é, somos, sois, são, éramos, fomos, fostes, seríamos => ser

Tokenização

- Divide um texto em tokens
- Esses tokens podem ser parágrafos, frases ou palavras individuais

Olá. Estamos em São Paulo. Hoje trabalharemos com NLP.

Olá.

Estamos em São Paulo

Hoje trabalharemos com NLP.

Olá

Estamos

no

São

Paulo

Hoje

Trabalharemos

com

NLP

HANDS-ON:

FUNDAMENTOS DE NLP

(TOKENIZAÇÃO, LEMATIZAÇÃO, STOP WORDS,
POS TAGGING E NER)

REPRESENTAÇÃO VETORIAL PARA DOCUMENTOS

Salton, 1983

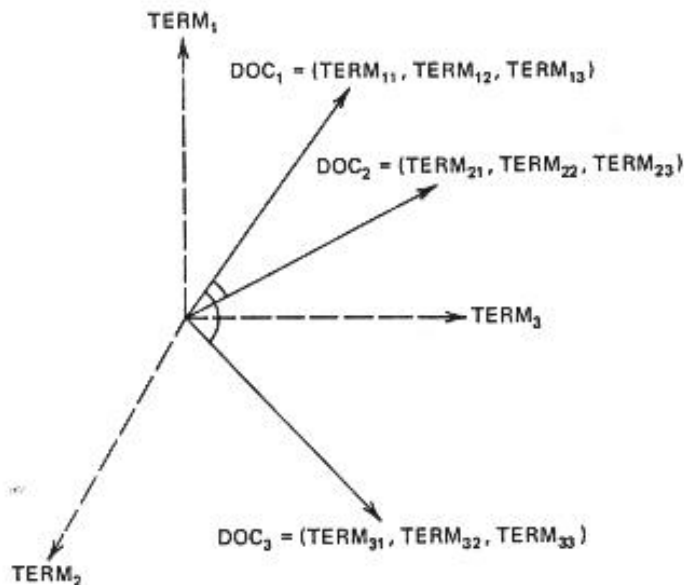



Figure 4-2 Vector representation of document space.

A vector space model for automatic indexing

Full Text:  PDF

Authors: [G. Salton](#) [Cornell Univ., Ithaca, NY](#)
[A. Wong](#) [Cornell Univ., Ithaca, NY](#)
[C. S. Yang](#) [Cornell Univ., Ithaca, NY](#)

Published in:

- Magazine
Communications of the ACM [CACM Homepage archive](#)
Volume 18 Issue 11, Nov. 1975
Pages 613-620
[ACM](#) New York, NY, USA
[table of contents](#) doi>[10.1145/361219.361220](#)

BAG OF WORDS (BoW)

O Bag of Words é um algoritmo que conta quantas vezes uma palavra aparece em um documento.

Representação BoW			
Vocabulário	Doc 1	Doc 2	Doc 3
Nós	10	5	8
Trabalhamos	23	40	5
Com	7	32	17
Tecnologia	5	15	25

Após normalização, todos os elementos de cada vetor de documento somam 1, representando a probabilidade de uma determinada palavra estar presente no documento avaliado.

TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TFIDF)

- Quanto maior for o número de documentos em que aparece uma determinada palavra, menos valiosa essa palavra é como informação relevante
- Enquanto o BoW avalia a frequência, o TF-IDF avalia a relevância das palavras que compõem um documento
- A ideia do TF-IDF é destacar apenas as palavras que forem frequentes e distintas

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

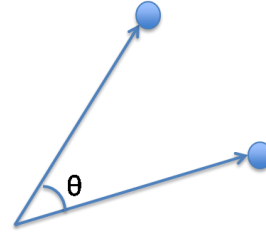
$tf_{i,j}$ = número de ocorrências da palavra i , no documento j

df_i = número de documentos contendo a palavra i

N = número total de documentos no corpus (conjunto de documentos)

SIMILARIDADE DE COSSENOS

$$\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



Similar scores
Score Vectors in same direction
Angle between them is near 0 deg.
Cosine of angle is near 1 i.e. 100%

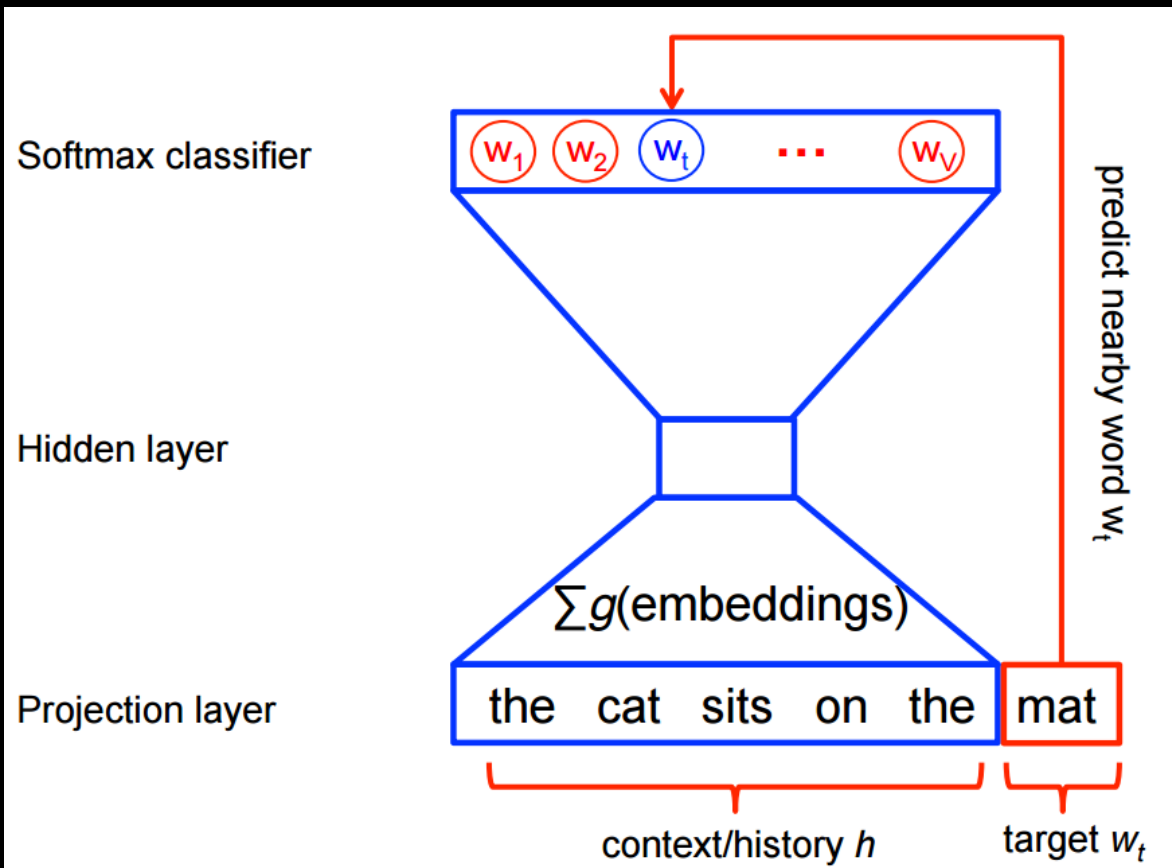
Unrelated scores
Score Vectors are nearly orthogonal
Angle between them is near 90 deg.
Cosine of angle is near 0 i.e. 0%

Opposite scores
Score Vectors in opposite direction
Angle between them is near 180 deg.
Cosine of angle is near -1 i.e. -100%

HANDS-ON: CLASSIFICAÇÃO DE TEXTOS

(TF-IDF E MACHINE LEARNING)

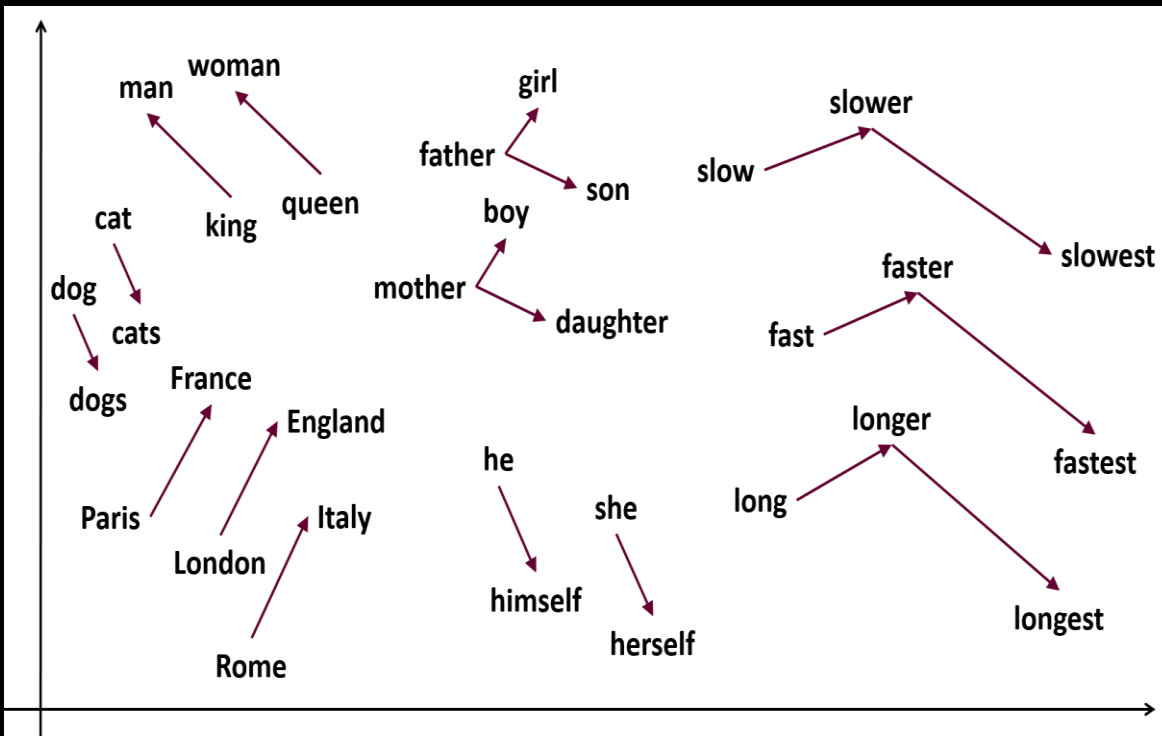
REPRESENTAÇÕES BASEADAS EM CONTEXTO: WORD2VEC



REPRESENTAÇÕES BASEADAS EM CONTEXTO: WORD2VEC

$\text{vec}(\text{"man"}) - \text{vec}(\text{"king"}) = \text{vec}(\text{"queen"}) - \text{vec}(\text{"woman"})$

$\text{vec}(\text{"man"}) - \text{vec}(\text{"king"}) + \text{vec}(\text{"woman"}) = \text{vec}(\text{"queen"})$



FIAP