PROCESSAMENTO DE LINGUAGEM NATURAL

Profª. MSc. Elene Ohata

Novembro/2023

TÓPICOS

- Part of Speech
- Named Entity Recognition
- Bag of Words
- TF-IDF
- Classificação

PART OF SPEECH (POS)

- Identificação da função (classe gramatical) de cada palavra em uma frase.
- Algumas palavras em diferentes posições podem ter outro sentido.
- A marcação de POS desempenha um papel para entender em que contexto a palavra é usada na frase. Então, é útil na análise de frases, recuperação de informações e análise de sentimentos, entre outras aplicações.

NAMED ENTITY RECOGNITION (NER)

- Reconhecimento de Entidade Nomeada
- Consiste em identificar e categorizar informações-chave (entidades) em textos.
- Uma entidade pode ser qualquer palavra ou série de palavras que se referem ao mesmo tema.
- Cada entidade detectada é classificada em uma categoria predeterminada. As entidades podem ser nomes de pessoas, organizações, locais, horários, quantidades, valores monetários, porcentagens, entre outros.

Profª. MSc. Elene Ohata

EXTRAÇÃO DE CARACTERÍSTICAS

- **Maioria dos algoritmos de aprendizado de máquina não suportam texto.**
- **Precisamos realizar uma “extração de características” do texto para transformar em atributos numéricos.**

BAG OF WORDS

João gosta de assistir filmes no cinema.

Maria também gosta de filmes.

João também gosta de ir a praia.

BAG OF WORDS

João gosta de assistir filmes no cinema.

Maria também gosta de filmes.

João também gosta de ir a praia.

João, gosta, de, assistir, filmes, no, cinema, Maria, também, ir, a, praia

BAG OF WORDS

João gosta de assistir filmes no cinema.

Maria também gosta de filmes.

João também gosta de ir a praia.

João	gosta	de	assistir	filmes	no	cinema	Maria	também	ir	a	praia
2	3	3	1	2	1	1	1	2	1	1	1

BAG OF WORDS

João gosta de assistir filmes no cinema.

Maria também gosta de filmes.

João também gosta de ir a praia.

João	gosta	de	assistir	filmes	no	cinema	Maria	também	ir	a	praia
2	3	3	1	2	1	1	1	2	1	1	1

BAG OF WORDS

João gosta de assistir filmes no cinema.

Maria também gosta de filmes.

João também gosta de ir a praia.

	João	gosta	de	filmes	também
Sent. 1	1	1	1	1	0
Sent. 2	0	1	1	1	1
Sent. 3	1	1	1	0	1

TF-IDF

- TF = Term Frequency
- IDF = Inverse Document Frequency
- TF-IDF = TF * IDF

- Diminui o “peso” de termos que aparecem com muita frequência e aumenta em termos que aparecem com menos frequência.

TF

João gosta de assistir filmes no cinema.

Maria também gosta de filmes.

João também gosta de ir a praia.

	Sent. 1	Sent. 2	Sent. 3
João	1/7	0	1/7
gosta	1/7	1/5	1/7
de	1/7	1/5	1/7
filmes	1/7	1/5	0
também	0	1/5	1/7

IDF

$$\log\left(\frac{(Number\ of\ documents)}{(Number\ of\ documents\ containing\ word)}\right)$$

	IDF
João	log(3/2)
gosta	log(3/3)
de	log(3/3)
filmes	log(3/2)
também	log(3/2)

TF-IDF

João gosta de assistir filmes no cinema.

Maria também gosta de filmes.

João também gosta de ir a praia.

	Sent. 1	Sent. 2	Sent. 3
João	1/7*log(3/2) = 0,025	0*log(3/2) = 0	1/7*log(3/2) = 0,025
gosta	1/7*log(3/3) = 0	1/5*log(3/3) = 0	1/7*log(3/3) = 0
de	1/7*log(3/3) = 0	1/5*log(3/3) = 0	1/7*log(3/3) = 0
filmes	1/7*log(3/2) = 0,025	1/5*log(3/2) = 0,035	0*log(3/2) = 0
também	0*log(3/2) = 0	1/5*log(3/2) = 0,035	1/7*log(3/2) = 0,025

CLASSIFICAÇÃO

Profª. MSc. Elene Ohata

Novembro/2023

OBRIGADA!

elene.ohata@lapisco.ifce.edu.br