Processamento Natural de Linguagem

Aula 4

Modelos clássicos NLP

Semantix®

All about data

Quem sou eu?

Filipe Theodoro

Cientista de dados da Semantix

Instrutor do Semantix Academy

Contatos

filipe.theodoro@semantix.com.br

linkedin.com/in/filipe-theodoro

Semantix®
All about data

Ementa

- Bag of Words (BoW)
- TF-IDF
- LSA
- LDA
- BM25
- Exercício de fixação

Bag of Words (BoW)

Semantix®
All about data

Bag of Words (BoW)

- Saco de palavras.
- Transforma tokens em uma matriz de ocorrências.

Tokens
[‘eu’, ‘tenho’, ‘um’, ‘gato’, ‘de’, ‘estimação’]
[‘meu’, ‘gato’, ‘odeia’, ‘cachorro’, ‘de’, ‘rua’]
[‘meu’, ‘gato’, ‘e’, ‘meu’, ‘cachorro’, ‘foram’, ‘na’, ‘rua’]

	eu	tenho	um	gato	de	estimação	meu	odeia	cachorro	rua	e	foram	na
DOC 1	1	1	1	1	1	1	0	0	0	0	0	0	0
DOC 2	0	0	0	1	1	0	1	1	1	1	0	0	0
DOC 3	0	0	0	1	0	0	2	0	1	1	1	1	1

TF-IDF

Semantix®
All about data

TF-IDF

- TF – frequência do termo
- IDF – inverso da frequência nos documentos
- Determina a importância da palavra em relação a um corpus

$$TF\_IDF(token) = TF(token, documento) x\ IDF(token)$$

$$IDF(token) = \log\left(\frac{1+N}{1+DF(token)}\right) + 1$$

LSA

Semantix®
All about data

LSA

- Latente semantics analysis (Analise Latente Semântica)
- Modelagem de tópicos
- Relação da ocorrência das palavras e documentos
- *Método não supervisionado*
- *Utiliza SVD (Decomposição em valores singulares), método de álgebra linear para reduzir a dimensão de uma matriz.*
- *É possível identificar palavras sinônimas.*

LDA

Semantix®
All about data

LDA

- Alocação latente de Dirichlet (Latent Dirichlet allocation)
- Modelo estatistico generativo
- Cada documento é uma mistura de um pequeno número de tópicos
- A presença de cada palavra é atribuível a um dos tópicos do documento.

BM25

Semantix®
All about data

BM25

- Função de rankeamento utilizado em Recuperação de Informação
- Estima a relevância dos documentos para uma entrada

$$SCORE(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)}$$

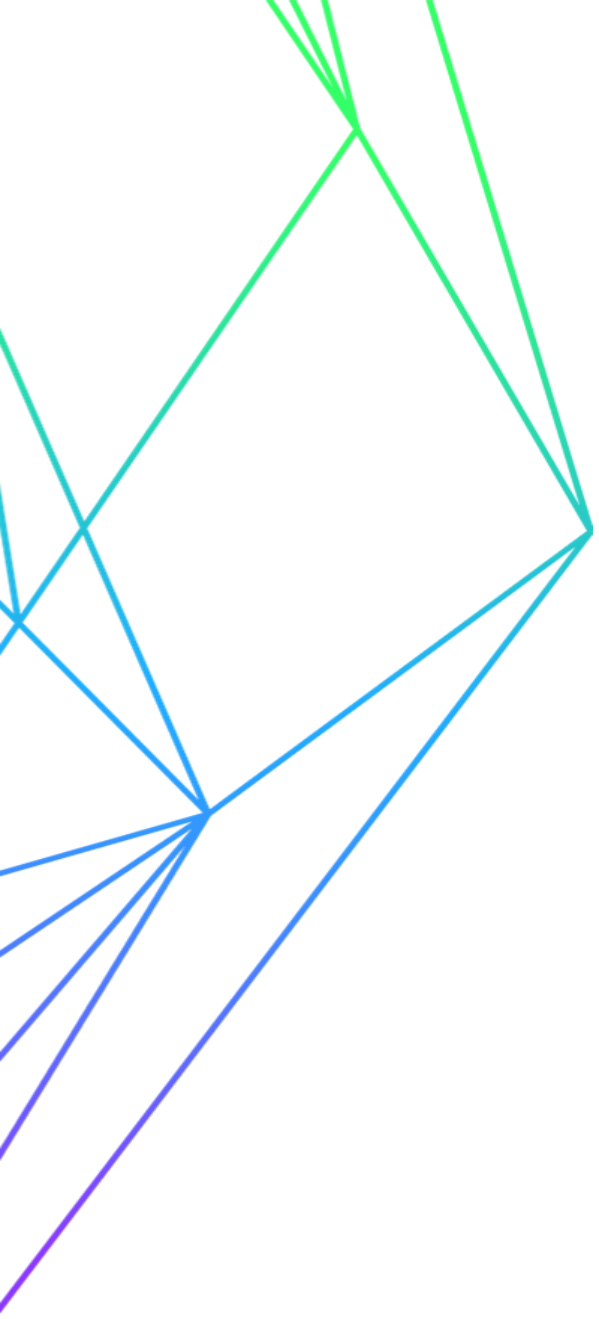
$$IDF(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

Exercício de fixação

Semantix®
All about data

Bibliografia

- SVD - https://www.ufsj.edu.br/portal2-repositorio/File/nepomuceno/mn/08MN_SL6.pdf
- LDA - https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/
- BM25 - https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/okapi_trec3.pdf



Semantix®

All about data

filipe.theodoro@semantix.com.br

www.semantix.com.br