Processamento de Linguagem Natural

UNIDADE 04

Representação vetorial de textos

O que é BoW?

Quando falamos em representar textos por meio de vetores numéricos, a primeira técnica que devemos entender é o *Bag of Words* (BoW). Na Unidade de Aprendizagem **Representação vetorial de textos – bag of words**, disponível na semana, vamos entender como podemos representar textos em forma de vetores numéricos e como funciona o método BoW.

Como melhorar a representação vetorial por TF-IDF?

Além do BoW, podemos representar numericamente os textos por meio do método Term Frequency – Inverse Document Frequency (TF-IDF), que é apresentado em detalhes na Unidade de Aprendizagem Representação vetorial de textos – TF-IDF.

A seguir, você encontra dois notebooks com exemplificações práticas de como aplicar os métodos BoW e TF-IDF utilizando algumas bibliotecas da linguagem Python. No vídeo tutorial, executaremos juntos esses notebooks.

Notebook – Google Colab

Explore o *notebook* a seguir, que demonstra a utilização do BoW.

https://colab.research.google.com/drive/1-dWYANo-zM1gLCxdd8YY6vUvzYvGjYS1

Notebook – Google Colab

Explore o *notebook* a seguir, que demonstra a utilização do TF-IDF.

https://colab.research.google.com/drive/1-c6g5bPW5lwHclhJFHB1muBR5hdDTrpk

Conclusão

Nesta unidade, você aprendeu a:

- Descrever como o computador realiza a interpretação de dados textuais por conversão numérica.
- Definir o conceito de vetorização de palavras.
- Analisar e diferenciar os métodos BoW e TF-IDF.
- Aplicar os métodos utilizando Python.

Referências

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing**. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 15 dez. 2020.

THE SCIKIT-LEARN DEVELOPERS. **sklearn.feature_extraction.text.TfidfVectorizer**. 2018. Disponível em: https://docs.w3cub.com/scikit_learn/modules/generated/sklearn.feature_extraction.text.tfidfvectorizer/. Acesso em: 9 mar. 2020.