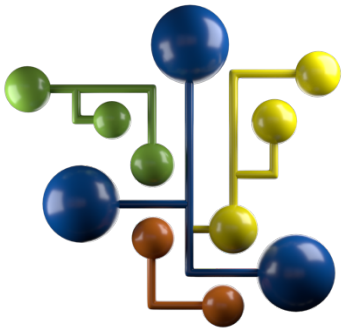

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Vetorização

A vetorização é um aspecto importante da extração de características no domínio de PLN. Transformar o texto em um formato vetorial é uma tarefa importante. As técnicas de vetorização tentam mapear todas as palavras possíveis para um inteiro específico. Existem muitas APIs disponíveis que facilitam sua vida. O scikit-learn possui o DictVectorizer para converter texto em um formulário de codificação simples. A outra API é o CountVectorizer, que converte a coleção de documentos de texto em uma matriz de contagens de tokens. Mas recentemente, o scikit-learn disponibilizou o TfidfVectorizer que aplica as operações com um único algoritmo. Também podemos usar word2vec para converter dados de texto para o formato vetorial (o que será estudado em detalhes mais a frente aqui no curso).

Consulte o link abaixo para mais detalhes sobre essas funções:

http://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction

Agora, vamos compreender o que é o conceito de Hot-Encoding para um aplicativo PLN. Essa codificação única é considerada parte da vetorização.