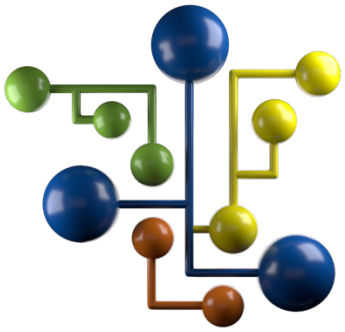

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Compreendendo o Bag of Words

O modelo BOW (Bag of Words) facilita nossa vida porque simplifica a representação usada em PLN. Neste modelo, os dados estão na forma de texto e são representados como o pacote ou *multiset* de suas palavras, desconsiderando a gramática e a ordem das palavras e apenas mantendo as palavras. Aqui, o texto é uma frase ou documento. Vamos dar um exemplo para você entender melhor o BOW. Vamos pegar o seguinte conjunto de documentos:

Documento de texto 1: John gosta de assistir ao futebol. Chris gosta de futebol também.

Documento de texto 2: John também gosta de assistir a filmes.

Com base nesses dois documentos de texto, você pode gerar a seguinte lista:

Lista de palavras = ["John", "gosta", "de", "assistir", "futebol", "Chris", "também", "filmes"]

Esta lista é chamada BOW (Bag of Words). Aqui, não estamos considerando a gramática das sentenças. Nós também não estamos incomodados com a ordem das palavras.

BOW é frequentemente usado para gerar recursos; depois de gerar a BOW, podemos derivar a frequência do termo (Term-Frequency) de cada palavra no documento, que pode ser posteriormente alimentado em um algoritmo de aprendizado de máquina.

Para os documentos anteriores, você pode gerar a seguinte lista de frequências:

Contagem de frequência para o Documento 1: [1, 2, 2, 1, 2, 1, 1, 0, 0]
Contagem de frequência para o Documento 2: [1, 1, 1, 1, 0, 0, 0, 1, 1]

Então, como geramos a lista de contagens de frequência? Para gerar a contagem de frequência do Documento 1, considere a lista de palavras e verifique quantas vezes cada uma das palavras listadas aparece no Documento 1.

Aqui, primeiro pegaremos a palavra ***John***, que aparece no Documento 1 uma única vez; a contagem de frequência para o Documento 1 é 1. Contagem de frequência para o Documento 1: [1].

Para a segunda entrada, a palavra ***gosta*** aparece duas vezes no Documento 1, portanto, a contagem de frequência é 2. Contagem de frequência para o Documento 1: [1, 2].

Agora, vamos pegar a terceira palavra da nossa lista e a palavra é ***de***. Essa palavra aparece no Documento 1 duas vezes, então fazemos a terceira entrada na contagem de frequência como 2. Contagem de frequência para o Documento 1: [1, 2, 2]. Geramos a contagem de frequência para o Documento 1 e Documento 2 da mesma maneira. Vamos aprender mais sobre frequência quando estudarmos TF-IDF, ainda neste capítulo.

O script cap05-03-BOW.py em anexo demonstra um exemplos simples de BOW.