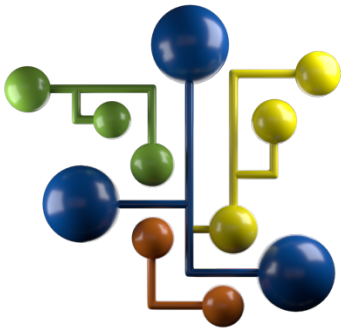

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Recursos e Ferramentas Semânticas

Tentar obter o significado exato de uma linguagem natural ainda é uma tarefa desafiadora no domínio da PLN, embora tenhamos algumas técnicas que foram desenvolvidas recentemente e recursos que podemos usar para obter a semântica da linguagem natural. Atualmente, as principais técnicas e recursos são:

Algoritmo de análise semântica latente que usa frequência de termo - frequência de documento inverso (tf-idf) e o conceito de álgebra linear, como similaridade de cosseno e distância euclidiana, para encontrar palavras com significados semelhantes. Essas técnicas fazem parte da semântica distribucional.

O outro é word2vec. Este é um algoritmo recente que foi desenvolvido pelo Google e pode nos ajudar a encontrar a semântica de palavras e palavras que têm significados semelhantes. Vamos explorar o word2vec e outras técnicas como Engenharia Avançada de Recursos e Algoritmos PLN mais a frente. Além do Word2vec, outro recurso poderoso é o WordNet, que é o maior corpus disponível para nós e é marcado por humanos. Também contém tags de sentido para cada palavra. Esses bancos de dados são realmente úteis para descobrir a semântica de uma determinada palavra. Você pode dar uma olhada no WordNet no seguinte link: https://wordnet.princeton.edu.

Vimos a maioria dos conceitos relacionados ao domínio da PLN e também vimos como podemos derivar recursos usando esses conceitos e ferramentas disponíveis. Agora é hora de pular para a próxima seção, que nos dará informações sobre os recursos estatísticos.