Megadados 2019-2

Wikipedia WordCloud

Pipeline de Engenharia de Dados utilizando TF-IDF.

Introdução e Objetivos

O objetivo desse trabalho é gerar uma WordCloud das palavras mais relevantes à outra palavra da Wikipedia. Isso pode ser realizado a partir de uma pipeline utilizando nossos conhecimentos de engenharia de dados. Essa pipeline ainda possui uma sub-pipeline que realiza o cálculo do TF-IDF das palavras relacionadas e com esses valores em mãos é possível gerar a WordCloud resultante.

Metodologia

Na recuperação de informações TF-IDF é uma estatística numérica cujo objetivo é refletir a importância de uma palavra para um documento em uma coleção ou corpus. É frequentemente usado como um fator de ponderação em pesquisas de recuperação de informações, mineração de texto e modelagem de usuários. O valor TF-IDF aumenta proporcionalmente ao número de vezes que uma palavra aparece no documento e é compensado pelo número de documentos no corpus que contém a palavra, o que ajuda a ajustar o fato de que algumas palavras aparecem com mais frequência em geral.

O valor TF-IDF de uma palavra aumenta proporcionalmente à medida que aumenta o número de ocorrências dela em um documento, no entanto, esse valor é equilibrado pela frequência da palavra no corpus. Isso auxilia a distinguir o fato de a ocorrência de algumas palavras serem geralmente mais comuns que outras.

Depois de obtidos esses valores são filtrados as top 25 palavras e feitas permutações entre elas para obter um dicionário de pares. Com esse dicionário pode-se contar os pares e obter os mais notáveis, assim possibilitando a geração de uma WordCloud.

Resultados e Conclusões

Utilizando Clusters e Buckets da AWS S3 foi criada a infraestrutura necessária para o projeto ser realizado, porém a pipeline foi lenta demais para que as queries pudessem ser feitas com eficiência.