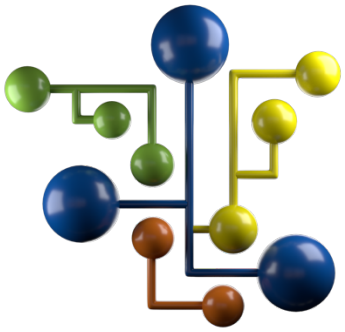

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Aspectos Linguísticos e Estatísticos da Normalização

O aspecto linguístico da normalização inclui o conceito de normalização do texto. A normalização de texto é o processo de transformar o texto em uma única forma canônica. Vamos dar um exemplo para entender a normalização do texto corretamente. Se você estiver desenvolvendo um aplicativo de pesquisa e quiser que o usuário insira John, John se tornará uma string de pesquisa e todas as strings que contiverem a palavra John também deverão ser exibidas. Se você estiver preparando dados para pesquisar, as pessoas preferem usar o formato *stemmed*; mesmo que você pesquise voando ou voe, em última análise, essas são formas derivadas da palavra voar. Portanto, o sistema de pesquisa usa o formato *stemmed* e outros formulários derivados são removidos. Esse é o aspecto linguístico da normalização.

O aspecto estatístico da normalização é usado para fazer o dimensionamento de recursos. Se você tiver um conjunto de dados em que os intervalos de um atributo de dados são muito altos e os intervalos dos outros atributos de dados forem muito pequenos, geralmente precisamos aplicar técnicas estatísticas para reunir todos os atributos ou recursos de dados em um intervalo numérico comum. Há muitas maneiras de realizar essa transformação, mas aqui vamos ilustrar o método mais comum e fácil de fazer isso chamado escalonamento min-max (Min-Max Scale). Vamos ver equações e exemplos matemáticos para entender o conceito. O dimensionamento min-max converte os recursos no intervalo de [0,1]. A fórmula geral é:

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Suponha que você tenha valores de recursos como [1, 8, 22, 25]; Se você aplicar a fórmula anterior e calcular o valor para cada um dos elementos, você obterá o recurso com um intervalo de [0,1]. Para o primeiro elemento, z = 1 - 1/25 -1 = 0, para o segundo elemento, z = 8 -1 / 25-1 = 0,2917 e assim por diante. A biblioteca scikit-learn tem uma API que você pode usar para dimensionar min-max no conjunto de dados.