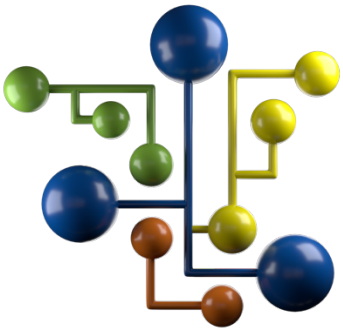

Data Science Academy

www.datascienceacademy.com.br

Processamento de Linguagem Natural

Mini-Projeto 2

Aplicação de Classificação de E-mails

No mundo da Internet hoje, uma quantidade enorme de dados é transferida entre computadores na forma de e-mails. Consequentemente, está ficando difícil classificar manualmente os e-mails importantes dos não importantes.

A classificação de e-mail foi extensivamente estudada e pesquisada no passado, mas a maior parte da pesquisa foi no campo de detecção e filtragem de spam.

Este mini-projeto se concentra na classificação de e-mails dos funcionários da Enron, empresa que recentemente esteve envolvida em escândalo de corrupção nos EUA. São mais de 500.000 e-mails que foram catalogados em um dataset disponível publicamente no link abaixo:

https://www.kaggle.com/wcukierski/enron-email-dataset

Com base na análise de texto dos e-mails, vamos criar uma heurística (regra) para definir se o e-mail é acionável (direciona o usuário a tomar uma ação) ou não acionável (apenas comunicação institucional, por exemplo).

Os scripts estão ao final do capítulo junto com o arquivo LEIAME.txt com todas as instruções. Estude atentamente o script, execute, faça mudanças. Muitos dos temas deste mini-projeto serão estudados ao longo do curso.

Bons estudos.