

Uso de Conceitos Básicos de Machine Learning Na Classificação de E-mails Como Spam

Lucas Lima da Cunha

lucas.lcunha@hotmail.com

Introdução

Mensagens de spams são consideradas um grande problema na internet, tanto por conta do gasto em processamento e armazenamento de um e-mail inútil, quanto pelos riscos que muitos apresentam a privacidade e segurança dos receptores, sendo carregados de conteúdo maliciosos.

Cientes deste problema empresas como a Google empenham grande investimento e esforço em melhorar suas ferramentas de detecção de spam através de técnicas como machine learning. Segundo um artigo online, recentemente ao passarem a utilizar um framework de ML de código aberto chamado TensorFlow, desenvolvido pela própria Google, eles passaram a bloquear por volta de 100 milhões de mensagens de spam a mais por dia e alegam uma precisão de mais de 99,9%ⁱ.

Analisando essa demanda de tecnologia para resolução deste problema foi desenvolvido um estudo com as aplicações de machine learning na classificação de e-mails.

Metodologia

Foram utilizadas a plataforma Colab, da Google, para desenvolvimento dos notebooks (códigos de ML), bibliotecas básicas e um dataset (conjunto de dados) contendo mais de 5 mil mensagens de e-mails, a quantidade que cada uma das 149 palavras mais comuns aparecem na mensagem, contagem de quantas vezes aparecem palavras comuns ao todo, a data da mensagem e um campo informando se é spam ou não.

O processo de estudo deste dataset foi dividido em dois. Na primeira parte foram extraídas estatísticas destes dados e na segunda foi desenvolvido um script que aprenderia com este dataset e seria capaz de identificar padrões de spam e classificar mensagens.

- Primeira Parte

Utilizando então a plataforma Colab foi importado o dataset e dele foi possível extrair estatísticas interessantes a partir de alguns comandos, como por exemplo uma nuvem de palavras evidenciando as palavras mais comuns, os dias de cada mês com maior movimentação de mensagens, gráficos comparando a quantidade de mensagens spam e não spam por mês entre outros.

- Segunda Parte

