Marcelo de Souza Pena - RA: 11039314

Título: Estimando a atratividade de vídeos de divulgação científica no YouTube com aprendizado de máquina

Resumo: neste projeto a princípio usarei um software para minerar informações gerais (nome do vídeo, nome do canal, número de inscritos, duração e link do vídeo) de uma playlist no YouTube criada por mim mesmo. São dez canais selecionados, todos produtores de conteúdo de divulgação científica com número de inscritos entre 100 mil e 400 mil, isso pois minha intenção é avaliar a relação entre o conteúdo e o número de visualização e utilizar canais com milhões de inscritos, como o Manual do Mundo, acabaria por adicionar um viés aos dados, de forma que mesmo o vídeo menos visto do canal citado possivelmente teria mais visualização que os mais vistos de um canal menor, com cerca de 100 mil inscritos. Focar nesta faixa evitaria este viés. De cada canal selecionado são minerados 30 vídeos: os 10 mais vistos, os 10 mais recentes e os 10 mais antigos. Além as informações coletadas via software, farei uma classificação manual para acrescentar tema, formato, data e número de visualizações de cada vídeo. Então vem o pré-processamento onde alterarei o formato da data para o número de dias desde publicação, converterei a duração dos vídeos para segundo e retirarei as colunas que não serão necessárias no aprendizado. Em seguida adequarei os dados para treinamento de modelos de regressão e de classificação.

Justificativa/motivação: atuo como divulgador científico produzindo e compilando material em formato de texto, mas planejo expandir para formato de vídeo também, por isso aprender com a experiência de outros divulgadores pode me ajudar a ser bem sucedido nessa empreitada, além de auxiliar os próprios divulgadores que compõem o dataset a melhorarem seus números adequando seu conteúdo.

Objetivo: criação de um modelo que estime o número de visualizações de um vídeo de divulgação científica no YouTube após determinado tempo de sua publicação.

Metodologia: a mineração será feita com o ScrapeStorm, gerando um arquivo csv que editarei manualmente para acrescentar as informações que não são possíveis de minerar, parte do pré-processamento é feito no próprio csv, parte utilizando Python 3 no Jupyter, então aplicarei as técnicas vistas na disciplina para a criação dos modelos e visualização dos resultados utilizando as bibliotecas Scikit Learn, Matplotlib, dentre outras.