Relatório de Wrangle

Por Marcos Vinicius Gracioli Malta de Oliveira
Data: 22 de Agosto de 2018

O projeto de data wrangling foi uma experiência muito desafiadora pois eu aprendi muito sobre o processo de coleta, avaliação e limpeza de dados. Também aprendi como baixar arquivos programaticamente da internet e o uso da API do Twitter.

Eu juntei dados de três fontes diferentes para essa análise de dados. O perfil WeRateDogs deu acesso exclusivo à Udacity ao seu dados por meio de um arquivo do Twitter. Este arquivo contém dados básicos de tweets (tweet_ID, timestamp, texto, etc.) para todos os 5000+ dos seus tweets como estavam lá até dia 1 de agosto de 2017.

Usando os IDs de tweets do arquivo WeRateDogs, o objetivo era acessar por meio da API do Twitter o perfil e coletar os dados JSON de cada tweet usando a biblioteca Tweepy do Python e depois armazenar cada conjunto de dados JSON de cada tweet, que mais tarde, eu usaria para analisar as contagens retweet e favoritas do tweet (ou seja, "like").

O processo de coleta de dados para este projeto via api do twitter foi ao mesmo tempo interessante e frustrante já que o Twitter não liberou minhas chaves de acesso até a conclusão deste trabalho. Entrei em contato com os mentores da Udacity que prontamente me enviaram a rotina de acesso a api e o arquivo consolidado em formato JSON (“tweet_json.txt”).

Depois de reunir todos os dados, copiei os arquivos para os processos de avaliação e limpeza de dados. Eu avaliei os dataframes procurando por questões de qualidade e arrumação e, em seguida. Eu comecei o processo de limpeza, abordando os dados em falta e as informações erradas, no arquivo do Twitter. Em seguida, converti as colunas em um formato de dados adequado, alterando dados de timestamp em objetos datetime entre outros processos.

Também abordei problemas de qualidade nas colunas de “image-predictions.tsv” utilizando as bibliotecas pandas str.replace () e str.title (), então eu removi o sublinhado entre as palavras e capitalizei a letra inicial em cada palavra.

A etapa final no processo de limpeza de dados foi para juntar todos os três conjuntos de dados em um documento final contendo todas as informações relevantes. Para esta tarefa eu usei a biblioteca pandas usando a função pd.merge ().

Depois armazenei os dataframes limpos em formato csv e criei um banco de dados sqllite contendo uma tabela com os registros limpos. Consulto este banco de dados para gerar as análises finais e a visualização solicitada no projeto.