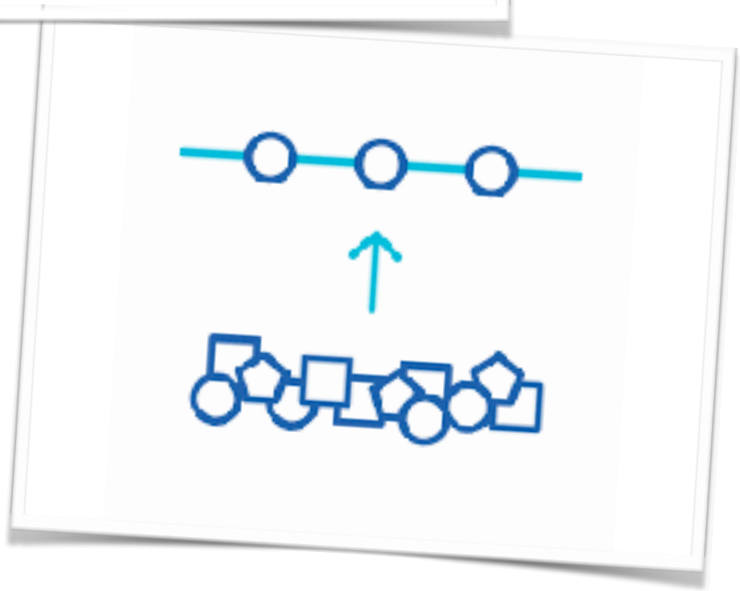
Filipe Silveira
1 de junho de 2018

Wrangle Report

Descrição sobre os esforços de wrangling do projeto @WeRateDogs



Filipe Silveira
1 de junho de 2018

Wrangle Report

Descrição sobre os esforços de wrangling do projeto @WeRateDogs

Foi disponibilizado, para a realização desse projeto, dois arquivos, uma chamado “twitter_archive_enhanced.csv" e o outro de “image_predictions.tvs”. O primeiro contém os dados dos tweets da conta @werategods e o segundo contém previsões de um algoritmo de I.A. da Udacity sobre as raças dos cachorros das imagens dos tweets do primeiro arquivo. Outra fonte de dado utilizada foi a API do Twitter, que foi utilizada para pegar alguns dados extras dos tweets do primeiro arquivo.

A primeira atividade que fiz foi importar os dois arquivos no jupyter notebook utilizando a biblioteca padas para fazer uma inspeção visual e de sua estrutura. Para isso utilizei os comandos head() e info(). Só com esses dois comandos já deu para perceber alguns problemas de qualidade e arrumação que os dados tinham.

Após essa visualização inicial, resolvi primeiro criar um função python, usando a biblioteca tweepy, para consultar os dados de cada tweet e salvar no arquivo “tweet_json.txt”. Alguns dados sobre alguns tweets não puderam ser encontrados. Provavelmente foram apagados.

Após os dados recolhidos da API do Twitter, comecei a busca maior pelos problema de qualidade e arrumação que a base tinha. Não tive muitos problemas para encontrar a quantidade especificada pelo projeto. Os problemas que eu encontrei estão listados abaixo:

- **Qualidade**:
  - **twitter-archive-enhanced:**
    - É uma especificação do projeto que só trabalhemos com o tweets da conta @WeRateDogs. Quando as colunas retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp estão preenchidas, significa que é um retweet. Foram encontrados 181 retweets
    - Reply não podem ser considerados também. Foram encontrados 78 replys.
    - Colunas in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
    - timestamp não está como um objeto datetime
    - Conteúdo HTML que não interessa na coluna source, que são os clients onde os tweets foram feitos
    - Coluna 'expanded_urls' com 59 registros faltantes. Padrão da URL: https://twitter.com/dog_rates/status/{tweet_id}/photo/1
    - Falta coluna rating com o resultado do cálculo = rating_numerator/rating_denominator
    - Colunas doggo, floofer, pupper, puppo, name com valores String 'None' ao invés de nulo(NaN)
  - **image-predictions:**
    - As colunas p, p_conf, p*_dog não deixa claro o que elas representam

- 22 previsões faltantes

- **Arrumação**:

  - **twitter-archive-enhanced:**

    - As colunas doggo, floofer, pupper, puppo são variáveis de um atributo que pode ser chamado 'classificação'

    - As contagens de Retweets e Favoritos estão no lugar errado ('tweet_json.txt')

  - **image-predictions:**

    - Colunas p1, p2, p3 são categorias e possuem valores extras que devem ser separados.

Feito a análise e encontrado os problemas, chega a hora da limpeza da base. Tive minha maior dificuldade foram nos problemas de arrumação, visto minha falta de experiência com a linguagem python e a biblioteca pandas. Em destaque a dificuldade, cito o problema de arrumação na base image-predicitions, foram horas de tentativas e pesquisa até da certo.