Relatório de Data Wranling do projeto WeRateDogs

A primeira fase do projeto foi o gathering, onde extraímos e obtemos os dados que precisávamos para trabalhar nos problemas. As formas de extrações foram as seguintes:

- twitter-archive-enhanced.csv: arquivo CSV fornecido pela Udacity, apenas importamos ele com a biblioteca Pandas, o arquivo tinha separador por vírgula;

- image-predictions.tsv: Recebemos uma URL de um arquivo, fizemos a requsição do mesmo com a biblioteca requests do Python, o arquivo tinha separação em tab;

- tweet_json.txt: O último foi preciso utilizar a biblioteca tweepy e extrair dado diretamente do Twitter da WeRateDogs, o arquivo foi salvo em TXT e posteriormente importado para o projeto com o Pandas;

Com os arquivos em mãos partimos para a fase de análise visual e de programação com o intuito de identificar os problemas.

Para auxiliar nesta tarefa foram utilizados métodos como describe, info e a própria impressão do dataset, no total foram identificados onze problemas de qualidade, quatro problemas de arrumação e dados faltantes.

Antes de iniciar a fase de limpeza criamos os arquivos a serem limpos e deixamos os originais intactos, e assim prosseguimos para a fase de resolução dos problemas.

Inicialmente resolvemos o problema de divergência de dados, alguns tweets foram deletados e assim os dados que foram extraídos via tweepy não correspondiam com os fornecidos pela Udacity, então foram igualados com os dados do Twitter sendo usados como referência, no caso apagamos tweets não mais presentes no Twitter do dataset principal.

Agora com o dataset alinhado aos dados providos do Twitter, iniciamos pelos problemas de arrumação, que eram os seguintes:

- A coluna de retweet_count de twitter_json deve estar e weratedogs;
- A coluna de favorite_count de twitter_json deve estar em weratedogs;
- tweet_id de imagepredictions está ao contrário dos demais datasets;
- Os status de cachorro de weratedogs poderiam estar em uma coluna;

Com os dados devidamente organizados, foi iniciada a correção dos problemas de qualidade que eram:

- Alguns twitters foram deletados, o que vai causar inconsistência com os dados presentes no dataset da Udacity e dos extraidos via API;

- Alguns cachorros tem mais que um status;

- rating_numerator com números menores que 10 (todos verificados no Twitter eram 10) ;

- rating_numerator com números maiores que 20 (todos verificados no Twitter eram entre 10~16);

- rating_denominator com números menores que 10 (todos verificados no Twitter eram 10+) ;

- rating_denominator com números maiores que 10 (todos verificados no Twitter eram 10+) ;

- name com nomes como None, a, an, the ✔
- expanded_urls com links diferentes do twitter (https://www.gofundme.com/mingusneedsus);

- timestamp com um +0000 desnecessário no fim;

- O dataset apresenta alguns retweets, verificar *retweeted_status_id*;

- timestamp está como string;

Importante citar que todos os problemas de qualidade identificados estavam em apenas um dataset, que era o provido pela Udacity em CSV, nomeado de twitter-archive-enhanced.csv e no projeto como weratedogs_clean.

Após cada problema resolvido com código foram feitos testes para certificar que o problema realmente foi resolvido, todos os testes estão presentes no Notebook e assim com esta segurança foi finalizada a etapa de wrangling e iniciada a de análise.