wrangle_report

September 9, 2021

0.1 Report: Data Wrangling Project

This Report is a part of a Data Analysis NanoDegree Project offered by Udacity. The project aims to gather data from Twitter and combine it with a third party data frame to create analysis about the tweets and the predicted dog's breed.

I have gathered the files **image_predictions.tsv** and **twitter_archive_enhanced-2.csv** using the requests package. Although the image_predictions.tsv file has almost all the information from the WeRateDog user, there is some missed variable, which I have gathered using the tweepy package. The very same .tsv file had to be parsed to be programatically accessed.It was originally downloaded from https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imagepredictions/image-predictions.tsv To get additional data and learn using APIs programmatically in python we created a twitter developer account to get the credentials for our API calls. With the tweet_id contained in the twitter archive file I queried the API to get the entire stored JSON data for those tweets. This process takes about 33 minutes. I stored the results in the file: tweet_json.txt.

The assessing part of the project was done visually first, just visualizing the files in the notebook.The dog's names issue was solved by evaluating if it starts with a capital letter it was a name if not it was an ordinary word and I have converted to "None". Most of the issues involving non-usual values to rating_numerator and rating_denominator were solved using a new tailored regular expression to gather the ratings from text column.Finally, I have solved the tidiness issues combining the tables twitter_archive_enhanced-2.csv and image_predictions.tsv in one called twitter_archive_master.csv. I have also merged 4 columns (doggo, pupper, puppo, and floofer) into one, which I have bundled and named as 'dog_kind'.

I have documented 8 issues but this final file version is not totally free of issues, because I faced the Data Wrangle as an iterative process, what I did so far was the first iteration.

For this reason, the twitter_archive_master.csv file is the final file version with a minored number of issues, and ready for a Data Analysis.