Wrangle Report

By: Debashish Singh
Date: March 28 2019

The data wrangling project was very challenging and I learnt a great deal about the data gathering process and twitter API. I' am extremely thankful to the team of mentors and the udacity team for guiding me throughout the process.

I gathered data from three different sources for this data analysis. WeRateDogs gave Udacity exclusive access to their twitter archive for this project in the form of a csv file. This file contains basic tweet data of their tweets as they stood on August 2017. Each tweet image was run through a convolutional neural network to analyze the images of dogs and correctly identify their breeds. The convolutional neural network predictions were programmatically downloaded using the Requests Python library as a tsv file. And finally, using the tweet IDs from the WeRateDogs archive I queried the Twitter API for each tweet's JSON data using the Python's Tweepy library. I stored each tweet's entire set of JSON data, which I would later use to analyze the tweet's retweet and favorite counts.

The data gathering process for this project was my greatest challenge, particularly querying the Twitter API. I also learnt about the requests library, json and wordcloud throughout the process. The Twitter API syntax was my greatest challenge and in my efforts to work through the problem. The Twitter API syntax was my greatest challenge and in my efforts to work through the problem

Once I had successfully gathered all the data, I copied the files for the assessment and data cleaning processes. I evaluated the dataframes looking for quality and tidiness issues and then set about fixing them. I began the cleaning process by tidying the data and merging all the different information into a single dataframe called twittwer_master. I also created a column for the dog breed extracting the data from 4 different columns. I then converted columns to a proper data format, primarily changing the timestamp data into datetime objects, tweet_id from a number into a string.

As there were many outliers which were distorting the visualizations I standardized the ratings columns (Both numerators and denominators) and removed the outliers. I removed the underscore between the words and capitalized the letter in each word to make a more cohesive table.

In summary, this project was my biggest challenge to date, specifically using the Twitter API to gather the JSON data. Overall, this project was completed successfully and I'm extremely pleased with the new skills I acquired.