Wrangle Report

By Shreyas Shukla

Date: 11 August 2019

The data wrangling project proved to be way more challenging than I expected. This was mostly because I had next to none knowledge about HTML,JSON etc. However, I learned a great deal about the data gathering process and the API.

Data needed to be gathered 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 which contains basic tweet data (tweet ID, timestamp, text, etc.). Each tweet image was run through a convolutional neural network to analyze the images of dogs and correctly identify their breeds. 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.

The data gathering process for this project was perhaps the greatest challenge for me, particularly querying the Twitter API. I spent many days googling how to gather data. Since, companies usually don’t provide their API data in public domain, it was really a challenging task to learn it. However, from bits and pieces of knowledge from here and there, I was able to crack it and got my dataset ready.

Now, once I got the 3 datasets, next step was to evaluate the dataframes for quality and tidiness issues and then set about fixing them. I began by addressing missing data and mislabeled information, which was predominantly found in the WeRateDogs Twitter archive. I then converted columns to a proper data format, like changing the timestamp data into datetime objects, tweet_id from a number into a string etc. Similarly, I conducted cleaning operations in the other two dataframes too, merged them into a single dataframe (i.e ‘whole’) and again performed cleaning whenever required. Also, ‘rating_numerator’ had some highly varying values which could have affected our analysis. Hence, I substituted them with appropriate values using various methods. Once done with the cleaning, I tried to gain insights from the dataset ‘whole’ using various python libraries and plots. Many values/columns were deliberately left unclean, as it would have negatively affected our analysis.

Overall, despite the complexity of the project, I’m extremely pleased with the new skills I acquired.