Wrangle Report

By Ahmed Redwan Yahya

Date: January 17, 2021

The project for data wrangling was very difficult and I learned a lot about the method of data collection and the Twitter API. I am immensely grateful to my mentor because without his encouragement and help, I could not have successfully completed this project.

For this data study, I obtained data from three distinct sources. In the form of a csv file, WeRateDogs gave Udacity exclusive access to their Twitter archive for this project. This archive contains simple tweet data for all 5000+ of their tweets as they were (tweet ID, timestamp, document, etc.). To analyze the pictures of dogs and correctly classify their breeds, each tweet image was run through a convolutionary neural network. Using the Requests Python library as a TSV format, the convolutional neural network predictions were programmatically downloaded. And finally, using the tweet IDs from the WeRateDogs folder, I queried the Twitter API for the JSON data of each tweet using the Python Tweepy library, I saved the whole collection of JSON data of each tweet, which I would later use to analyze the retweet and favorite (i.e. "like") counts of the tweet.

My biggest challenge was the data collection process for this project, especially querying the Twitter API. My biggest challenge was the Twitter API syntax and I spent 10 days visiting and revisiting any website I could find that provided Twitter API information in my attempts to work through the issue. I found that the Twitter API support documentation in general is not v I can't recall how many videos I watched on YouTube to try to gain details that would assist me with the project.Ultimately, it will take the guidance of my mentor to help me find out the solution to the problem, and I am so grateful to him for his assistance.

I copied the files for the evaluation and data cleaning processes once I had successfully collected all the data. I analyzed the data frames in pursuit of consistency and tidiness problems and then set out to address them. I started the process of cleaning by resolving missing data and mislabeled data,This was mainly contained in the Twitter archive of WeRateDogs. I then transformed columns into a correct data format, mainly changing the timestamp data into datetime objects, tweet-id from a number into a string, and column ratings into float objects. In the Predication columns of the Image Prediction dataframe, I also discussed quality problems. Utilizing the pandas library str.replace() and str.title() functions, I removed the underscore between the words and capitalized the letter in each word to make a more cohesive table. The final step in the data cleaning process was to inner join all three datasets into a final document containing all relevant information. For this task I used the pandas library using the pd.merge() function.

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.