Wrangle Report

By Savahnna L Cunningham
Date: November 26, 2017

The data wrangling project was very challenging and I learned a great deal about the data gathering process and the Twitter API. I'm extremely thankful for my Mentor, George, as I could not have completed this project successfully without his guidance and support.

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 archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 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 (i.e. "like") counts.

The data gathering process for this project was my greatest challenge, particularly querying the Twitter API. The Twitter API syntax was my greatest challenge and in my efforts to work through the problem I spent 10 days visiting and revisiting every website I could find that offered information on the Twitter API. I discovered that the support documentation for the Twitter API in general is not very good, especially for people who are trying to learn how an API works for the first time. I can't remember how many YouTube videos I watched to try and learn information that would help me with the project. Ultimately, it would take my Mentor's guidance to help me figure out the solution to the problem and I am so very grateful to him for his help.

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 addressing missing data and mislabeled information, which was predominantly found in the WeRateDogs Twitter archive. I then converted the Timestamp column from a string to a DateTime object using the apply() function to create 3 new columns of date and time information to ensure the table folllowed the Tidiness rules. The last step of the cleaning process for the WeRateDogs Twitter archive was to drop the columns that did not have any pertinent information and reorganized the remaining columns for an easier read. The cleaning process for the Tweet Query dataset was identical to that of the WeRateDogs archive, with the addition of changing the primary key to the tweet_id to ensure the dataframe matched the others to kept inline with the Tidiness rules.

The final step in the data cleaning process was addressing the quality issues in the Image Prediction dataframe. The predication columns had words labeled with both capital and lowercase letters, in addition to containing an underscore between the words that needed to be removed. These challenges were tackled easily with the use of the pandas library using the str.replace() and str.title() functions.

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.