Wrangle Report

By Saurabh Tyagi

Date: 13/nov/2018

The data wrangling project was very challenging and I learned a great deal about the data gathering process and the Twitter API..

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.

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 mislabelled information, which was predominantly found in the WeRateDogs Twitter archive. 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 and the rating columns into float objects. I also addressed quality issues in the Predication columns of the Image Prediction dataframe. 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 am extremely pleased with the new skills I acquired.