Wrangle Report

By Chaitanya Cherukuri
Date: May 7, 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 project. 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 urlretrieve function from urllib.request subpackage 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 analyse 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 a great challenge and, in my efforts, to work through the problem I spent more 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.

Once I had successfully gathered all the data, I started the second step in the data wrangling process which is assessing the data. The main goal for assessing the data is to look for quality and tidiness issues and then document them at the end of the section. I began assessing the data visually at first and then programmatically using several functions from python pandas library. I documented all the quality and tidiness issues that I found while assessing the data. I began the data cleaning process by addressing the missing data and mislabeled information, which was predominantly found in the WeRateDogs Twitter archive. Next, I addressed the tidiness issues which involves merging the three dataframes into a combined single dataframe and condensing four columns into single column to make the data tidy. For this task I used the pd.merge() function from the pandas library. It is always best to address tidiness issues first before moving on to address remaining quality issues since fixing the data quality issues on a tidy data is easier when compared to fixing quality issues on an untidy data. I then converted a column to a proper data format, primarily changing the timestamp data into datetime object. I also addressed a quality issue in the Predication columns of the Image Prediction dataframe. Utilizing the pandas library function str.lower(), I converted each word in the column to lower case to make a more cohesive table. Moreover, I also addressed several quality isuues in the Twitter Archive dataframe. One of it is removing the retweets from the final dataframe by finding the retweets present in the data using a regular expression and str.match() function and next addressing the inaccurate values present in the rating_numerator column by changing the inaccurate values with the correct values which were obtained by searching for a regular expression pattern in the text using re.search() function. Finally, I addressed the last issue by replacing incorrect names present in the name column. 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.