Wrangle report

January 29, 2019

Data wrangling project was quite challenging and I learned a lot l about the data gathering process and the Twitter API. I'm extremely thankful for my Mentors and Knowledge Hub's Community of Udacity, as I could not have completed this project successfully without their 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. 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 counts. So basically there where 3 dataframes i.e tweet_archive_enhanced.csv,image_prediction.csv,tweet_df. tweet_archive_enhanced was provided by Udacity.image_prediction we have to download it using webscrapping technique I have used Request Python library to download it.

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. 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.