Project 4 **Udacity | School of Data Science**
Data Analyst Nanodegree
Nov 13, 2018 Class

Jan 30th, 2019
Yi Su

Wrangle Report

Data were collected from three different sources. First data was collected from the "twitter-archive-enhanced.csv" file which was given by Udacity. Second data was extracted programmatically from a URL. Python's request library was used to extract data from URL. The third data was extracted from Twitter API using python's tweepy library. Then I checked for missing values, columns of data that I found unhelpful for future analysis, and any columns that did not abide to a tidy format. I decided to reduce the size of the data frame greatly. I extracted the favourites and retweet counts for each tweet.

The data gathering for this project was my greatest challenge, especially getting the tweet_json.txt from Twitter. The Twitter API syntax was the hardest part in this project. I spent 2 days to browsing websites to find the solutions. I realized that the support documentation for the Twitter API in general is hard to understand for the people who are learning how an API works for the first time.

Once I had gathered all the data, I copied all the files for the assessment and data cleaning processes. I evaluated the dataframe looking for quality and tidiness issues. I began the cleaning process by addressing missing data and mislabeled information. Then I 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. Furthermore, I also addressed quality issues in the Predication columns of the Image Prediction dataframe. Furthermore, I removed the underscore between the words and capitalized the letter in each word. Finally, I merge all three datasets into a final document.

The data wrangling project was very challenging and I learned a lot about the data gathering process and the Twitter API. I'm very pleased with the new skills I got.