Wrangle Report of We Rate Dogs

by Souvik Das

Introduction

The purpose of this project is to clean, gather and assess the data in hand by using the data wrangling techniques taught in the Udacity Nanodegree program. In this project, we are analysing the tweet archive of the twitter account We Rate Dogs.(@dog_rates)

Gathering

1. Twitter archive file: Downloaded this file manually from resources section: twitter_archive_enhanced.csv

2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Twitter API & JSON: Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas dataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Assessing

The assessing step is achieved by putting the data gathered in Google Docs and finding errors observable to human eye.

Next these are assessed programmatically by using .info(),.head(),.value_counts() methods to detect the data types and the range of the values present.

Cleaning

The following observations are made and changes are done programmatically:

1. Keep original ratings (no retweets) that have images

2. Delete columns that won't be used for analysis

3.Separate timestamp into day - month - year (3 columns)

4. Correct numerators with decimals

5.Correct denominators other than 10:

- Manually (few examples assessed by individual print text).
- Programmatically (Tweets with denominator not equal to 10 are usually multiple dogs).

6. Drop 66 jpg_url duplicated

7. Create 1 column for image prediction and 1 column for confidence level

8. Delete columns that won't be used for analysis

9. Keep original tweets only

10. Change tweet_id to type int64 in order to merge with the other 2 tables

11. Erroneous datatypes (doggo, floofer, pupper and puppo columns)

12.All tables should be part of one dataset

Conclusion

The data wrangling steps were reiterated as I found new types of data quality problems. It was a new and enjoyable experience to find and eliminate the faulty data. After obtaining clean data, it was important to aggregate into one dataset for further analysis.

Through this project I have learned gathering data using external APIs, handling JSON objects , cleaning and assessing data to organize into single dataframe and making word cloud through gathered data.