WeRateDogs Twitter Archive - Wrangle Report
By-Piyush Sinha

In this report I outline the wrangling efforts to assemble and clean the data required for analysis of the WeRateDogs Twitter Archive.

Data Gathering

I gathered data from 3 sources, stored in separate files:

1. WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
2. The image predictions file, programmatically downloaded from the Udacity servers.
3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library. The `favourite_count` and `retweet_count` were extracted programmatically from this file.

I loaded the 3 raw data files into separate tables: twitter_df,image_pred_df,tweets_json_df.

ASSESSING DATA

Once the three tables were obtained I assessed the data as following:

• Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv files in Excel.

• Programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc).

Cleaning

Cleaning Started with twitter_df named as twitter_df_clean.

I started with cleaning rating_numerator & rating_denominator of twitter_df.
Then Changing the datatypes of the required attributes of twitter_df.

After Cleaning twitter_df_clean I started with image_pred_df named as image_clean

First the false entries during predictions were removed.
Column p2 and p3 were dropped as my prediction was based on p1 column
Datatype of tweet_id column was changed to str instead of int.

Atlast tweets_json_df named as tweets_api_clean was cleaned.

Duplicated Values were removed**.**

Conclusion

Data wrangling is a core skill that whoever handles data should be familiar with.

I have used Python programming language and some of its packages. There are several advantages of this tool (as compared to e.g. Excel) that is used by many data scientists (including the guys at Facebook).

- For gathering data there are several packages that help scraping data off the web, that help using APIs to collect data (Tweepy for Twitter) or to communicate with SQL databases.
- It is strong in dealing with big data (much better than Excel).
- It can deal with a large variety of data (unstructured data like JSON (Tweets) or also structured data from ERP/SQL databases