Wrangling Data

By: Mohammad Tawfik

Gathering data:

This analysis consists of 3 pieces of gathered data. First, ‘twitter-archive-enhanced.csv’ which is provided by the ‘we rate dogs’ twitter account and it was downloaded manually and opened in pandas dataframe ‘df_archive’. Secondly, ‘’image-predictions.tsv’ which is downloaded programmatically and opened in another pandas dataframe ‘df_predict’. Lastly, ‘tweet_json.txt’ which is a text file contains JSON data for each tweet in the account gathered by twitter API using ‘tweepy’ library. And while gathering these data a list called ‘deleted_ids’ were created and saved as text file for later usage in cleaning. From this data, only ‘tweet_id’, ‘favorite_count’ and ‘retweet_count’ were opened in a dataframe ‘df_tweets’

Assessing data:

Assessing data was done both visually and programmatically. For visual assessing, these problems were ‘detected in twitter-archive-enhanced’:

1) Some tweets IDs were retweets or replies.

2) Some tweets IDs were for deleted tweets.

3) When dog name is missing, ‘None’ was used.

4) Dog stage variables were columns.

5) When dog stage is missing, ‘None’ was used.

And for programmatical assessing, these problems were found:

6) ‘timestamp’ was of type ‘object’.

7) ‘rating_numerator’ was of type ‘integer’.

8) ‘tweet_id’ in 3 dataframes was of type ‘integer’.

9) Some ‘names’, ‘rating_numerator’ and ‘rating_denominator’ didn’t make sense.

10) Missing values at column ‘expanded_url’, which then indicated lack of image.

11) ‘Source’ column was not needed.

Cleaning data:

Copies of the dataframes were made to do cleaning on.

1) Columns like ‘in_reply_to_status_id’ and ‘retweeted_status_id’ were first used to drop entries that were either replies or retweets, and then these columns were dropped.

2) The ‘deleted_ids’ file created from API code was used to drop entries from archive df and predictions df that were removed later by the account user.

3) Using numpy library and nan method. ‘None’ values in ‘name’ column were replaced by NaN’s.

4) To fix the variable column name issue, first ‘None’ values were replaced by an empty string. Then, the 4 columns were concatenated in one new column called ‘dog_stage’ and then the 4 columns were dropped. The new column ‘dog_stage’ had values like ‘doggopupper’ due to multiple dogs in the tweet. So this problem was solved manually by replacing it with ‘doggo-pupper’.

5) ‘None’ values in the new column ‘dog_stage’ was replaced by NaNs.

6) Since ‘timestamp’ is an object and contained date and time, date was extracted using slicing in a new column called ‘date’ and converted to datetime type using ‘to_datetime()’ method. Then, Year, month and day were extracted for later use in analysis and the columns were rearranged.

7) ‘rating_numerator’ and ‘tweet_id’ types was converted to ‘float’ and ‘string’, respectively, using ‘astype()’ method.

8) Names, rating numerator and rating denominator that didn’t make sense were reextracted from the text using regular expression formula or replaced with nan.

9) Rows with missing values in ‘expanded_url’ column were removed and then the column was dropped.

10) ‘Source’ column was dropped.

11) ‘favorite_count’ and ‘retweet_count’ column in ‘df_tweets’ dataframe were merged with ‘archive_clean’ dataframe on ‘tweet_id’ column.