Francisco Pineda

Wrangle Report

Gathering:

For this project we gathered data from three sources. Specifically our sources are CSV file called “twitter-archive-enhanced.csv”, a tsv file “image_predications.tsv”, and the twitter API.

1. I read in the CSV file (‘twitter-archive-enhanced.csv’) using panda’s read_csv() and stored it as df_archive.

2. I used the URL provided to download the image_predications.tsv. https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. I read in the TSV file using read_csv() using “\t” as the delimiter.

4. Using the Twitter API I downloaded the favorites and retweet counts for all the tweets listed in the ‘twitter-archive-enhanced.csv’ using tweepy library. I stored the data in a list on the jupyter notebook. I then stored that list as a text file (tweet_json.csv) so that the data can be used without reusing the twitter API to download the information.

Assessing:

Using .info() and .head() methods we can find possible issues with data.

Specifically I found the following issues regarding the data’s completeness, accuracy, validity and consistency.

- Names should be in lower case
- Timestamps should be converted to datetime
- Tweet_id columns for all dataframes should be a string
- Convert 'none' to NaN for names column
- Remove retweets
- Dog type names should all be lower case
- Remove in reply tweets
- Remove columns for retweet and in reply tweets

Untidiness:

- All dataframes should be combined using the tweet_id columns
- Create one rating value for each tweet (rating_numerator/rating_denominator)
- Remove unused columns.

Cleaning Steps:

1. Change all retweet_id columns in all three dataframes to datatype string instead of integer. I did this using the .astype() and using string as the input to that function for all three dataframes for the tweet_id column.

2. Change the names to all lower case. This was done with str.lower() function.

3. Replace incorrect names (‘none’, ‘a’, ‘an’, ‘the’) in the name column with NaN. I did this using the replace() function.

4. Changing the timestamps column to datetime. I did this using the pd.to_datetime() function.

5. I removed the retweeted tweets using the isnull() function

6. I removed the in reply tweets using the isnull() function

7. I dropped the columns “in_reply_to_status_id”, ”in_reply_to_user_id”, ”retweeted_status_id”, ’retweeted_status_user_id’, ’retweeted_status_timestamp” from the df_archive dataframe.

8. I used str.lower() to change all dog breed names to lowercase.

Tidiness steps:

1. I combined the numerator_rating and denomintor_rating into a single rating value for each tweet.

2. I removed the the numerator_rating and denominator_rating columns as they are no longer useful.

3. I then combined all three dataframes into one dataframe. I did this using the merge() function.

Conclusion:
Now that we have cleaned and tidied up the data we can analyze and visualize the data. To generate insights I will explore the ratings, the amount of retweets, amount of favorites and the names of the dogs within this dataset.