Project: Wrangle and Analyze Twitter Data

By: Maithili Desai

Data Sources

1. **Name:** WeRateDogs™ Twitter Archive (twitter-archive-enhanced.csv)
    - **Version:** Latest (Download 04.05.2020)
    - **Method of gathering:** Manual download
2. **Name:** Tweet image predictions (image_predictions.tsv)
    - **Version:** Latest (Download 04.05.2020)
    - **Method of gathering:** Programmatic download via Requests
3. **Name:** Additional Twitter data (tweet_json.txt)
    - **Source:** WeRateDogs™
    - **Version:** Latest (Download 04.05.2020)
    - **Method of gathering:** Twitter's API viaTweepy

Data Gathering

1. WeRateDogs™ Twitter Archive (twitter-archive-enhanced.csv)

I commenced the data gathering process by manual download of the twitter archive file which contains basic tweet data for all 5000+ of their tweets. Now the file can be loaded directly into a dataframe using Pandas. It includes attributes related to timestamp, source, tweet text, ratings, dog classifications and retweets information.

2. Tweet image predictions (image_predictions.tsv)

To gather this data file I first defined the source url where this file resides. Then I used the requests method to get the file from the source and then stored the contents of the response in a tsv file name `image_predictions.tsv`. It includes attributes related to image url, three classifications, confidence levels for each and a boolean value to indicate whether the image is of a dog.

3. Additional Twitter data using Tweepy API (tweet_json.txt)

To gather the data from the Twitter API I created a Twitter developer account and gathered the data via tweepy API. This results in a new file called *tweet_json.txt*. This includes additional tweet data related to favorite counts, retweet counts and tweet text length.

Data Assessing

Assessing quality and tidiness issues of the data.

Assessing Summary

Data Quality Issues (Content)

df_twitter table

- Datatype of tweet_id is integer and should be string
- Datatype of timestamp is object and should be datetime
- Some of the dogs are not classified in one of the stages: doggo, floofer, pupper or puppo and contain the value "None" under these columns
- Some of the dog names are incorrect (None, an, by, a, ...)
- Presence of retweets
- Some of the ratings are not correctly extracted (mostly if there are >1 occurence of the pattern "(\d+(.\d+)?/\d+(.\d+)?)"
- Mistakes due to transformation of ratings to integer (there are also floats)
- Source contains html code

df_predict table

- Datatype of tweet_id is integer and should be string
- Presence of retweets (duplicated rows in column jpg_url)
- Presence of pictures that are not dogs
- Predictions are sometimes uppercase, sometimes lowercase
- "_" instead of a whitespace in the predictions

df_api table

- Datatype of tweet_id is integer and should be string

Data Tidiness Issues (Structural)

df_twitter table

- doggo, floofer,pupper and puppo are not easy to analyze and should be in one column under dog_stage

df_predict table

- Prediction and confidence columns should be reduced to two columns - one for the prediction with the highest confidence (dog)

df_api table

- display_text_range contains 2 variables

All tables (Master Dataset)

- All three tables share the column tweet_id and should be merged together

Data Cleaning

Cleaning steps:

1. Merge the tables together
2. Drop the replies, retweets and corresponding columns. Also drop the tweets without an image or with images which don't display dogs
3. Clean the datatypes of the columns
4. Clean the wrong ratings numerators
    - Replace float values
    - Drop multiple occurences of patterns
5. Extract the source from html code
6. Split the display_text_range into two separate columns
7. Transform the doggo, floofer, pupper and puppo columns into one column namely dog_stage
8. Remove the incorrect names in the name column
9. Reduce the prediction columns into two: Mostly likely breed and corresponding confidence
10. Clean the new breed column by replacing the "_" with a whitespace and convert to lowercase

Summary and Conclusions

- Explored the data wrangling process and further analysed the dataset to give insights about 4 questions.
- Executed all the data wrangling steps involving gathering, assessing and cleaning the dataset to augment the data analysis step.
- Gathered the data from 3 sources namely, manual download of twitter archive data, use of requests library for programmatic download of image predictions and use of Tweepy API to get additional data for the tweets
- Assessed the data both visually and programmatically to identify data quality and data tidiness issues
- Cleaned the issues based on the priority of their dimensions
- Made a clean master dataset of WeRatesDogs twitter handle