wrangle_report

October 18, 2022

0.1 Wrangle_report

In this project, we wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning was then carried out in order to carry out analysis and visualisations. Various tentative conclusions/inferences/observations will be derived from said analysis. Some of the analysis carried out in this project is aimed at observing the accuracy and extent of the neural network used to predict dog breeds from the tweets. Also the level of user interaction with WeRateDogs tweets. More information like users favorite dog breed were analyzed.

The data wrangling carried out in this project included various oprations;

1. Data Gathering
2. Assessing Data (noting down, quality and tidiness issues)
3. Cleaning Data
4. Storing Data

Data Gathering For the project, data gathering was carried out in various manners i.e. data used was gathered from various sources using different methods.

The first set of data was gathered by programatically downloading and reading a csv file which was a twitter archive dataset containing data on a set of tweets by WeRateDogs page. This dataset was saved as 'twitter-archive-enhanced.csv'.

The second set of data was gathered by using the requets library to programmatically download and read a tsv file gotten from a given url containing data on the tweet images of a set of WeRateDogs tweets. This dataset was saved as 'tweet_image_predictions.tsv'.

The third set of data was gathered by using the tweepy library to query data from the twitter API. This dataset contained favorite count, retweet counts and other statistics on the same set of tweets as the first set of data. The dataset was saved as a JSON file 'tweet_json.txt'.

Asessing Data For this project, the data was assesed both visually and programmatically, and detected data quality issues and data tidiness issues were noted down to be worked on before analysis.

The following quality and tidiness issues were detected;

Quality issues -

1. drop non-original tweets rows i.e. tweets that are retweets and/or are tweets in reply to dog-rating tweets

2. drop/correct rows with incorrect dog ratings numerator (index numbers- 2335,45,)
3. rows with incorrect dog rating denominators
4. correct timestamp data type (from object to timedate...)
5. drop rows without tweet ids... (NaN as value ...)
6. convert url(html file) to a df table
7. note rows in df_2 where p1_dog, p2_dog or p3_dog is false i.e when the prediction is not an actual dog breed
8. image number column????
9. drop 'in_reply_to_status_id ' and 'in_reply_to_user_id' columns (they are not useful for our analysis)
10. drop 'retweeted_status_id' and 'retweeted_status_user_id' and 'retweeted_status_timestamp' column
11. drop 'source' column.... (not needed for our analysis)
12. convert tweet id column values in twitterapi data table from index form to plain values
13. non essential rows in twitterapi data df, leaving only tweet id, favorite count and retweet count columns
14. new tweet_id column should be created in the df_twitterapi_new table, to form and replace the former column in the table...

Tidiness issues -

1. dog type should be a column instaed of 4 different columns (doggo, floofer, puppo, pupper as variables...)
2. predicted dog breed (predicted_breed) should be a column instead of having columns p1, p1_dog, p1_conf, p2, p2_dog, p2_conf, p3, p3_dog and p3_conf. The predicted dog breed column is formed by indicating the prediction with the highest confidence (p1, p2, p3).
3. merge tables to form master dataset...

Cleaning Data For this project, the detected quality and tidiness issues were programatically cleaned using various code blocks as seen in the warngle_act notebook.

Storing Data In this section of the project, the cleaned datasets were saved and then merged to form a master dataset. The resulting clean data sets for each of the initial three datasets, df1_cleaned, df2_cleaned and dftwitterapi_cleaned were then merged programmatically to form 'twitter_archive_master'.

```
In [ ]:
```