wrangle_report

October 8, 2019

0.1 Wrangle Report

The data wrangling project was very interesting and I learned a lot from the process.

Three different sources are referred for the data analysis.

- Enhanced Twitter Archive
- Image Predictions File
- Additional Data via the Twitter API

For the twitter data, using the tweet IDs, I first queried Twitter API for each tweet's JSON data using the Python's Tweepy library I stored each tweet's entire set of JSON data. The data is further used to get the retweet number and like counts.

After analysing the data, I found the following quality issues and tidiness issues

Quality issues

- There are a lot of null values in columns like in_reply_to_status_id and in_reply_to_user_id.
- One suggestion is to change the timestamp to datetime stamp.
- Tweet_id should be string rather than int.
- Rating_numerator and rating_deniminator better to use double
- Need to remove the retweeted twitter
- Convert `id` column from a number to a string
- Remove not relevant columns such as `in_reply_to_status_id` and `in_reply_to_user_id`
- Need to clean the name column
- Transform timestamp to `yyyy-MM-dd HH:mm:ss`
- Add another columns which is the rating ratio rather than two columns only
- Keeping rows with p1_dog, p2_dog, or p3_dog = True. Only keeping entries that we can predict as dog and its species.
- Consulidating image prediction into one column

Tidiness issues

- Concatenate datasets to make one clean dataset using merge
- Tidy the 4 stages of dog column to create variable/value

And I have addressed each of the issue one by one utilising pandas.

```
In [ ]:
```