Onesmas Ngugi

11/25/2017

Wrangling Report

Data collection for this project involved the following steps:

- Manual download of the `WeRateDogs` twitter archive as a csv file.
- Programmatic download of the `image_predictions` tsv file from Udacity's servers using the requests library and a url.
- More tweet information, from Twitter using Twitter's API and the tweepy library as json objects which were then written onto a text file, `favs_and_retweets`.

After collecting all the data, I assessed it to determine which issues needed to be fixed so as to end up with clean data that could then be used for various analyses.

The first step in the cleaning process was to merge the fav_and_retweets data with the tweets from the `WeRateDogs` twitter archive. The merging was on tweet_id.

The following are some of the issues I dealt with:

- Filtered retweets and replies since only original tweets with images were required in the final DataFrame, then dropped all columns related to retweets and replies from the DataFrame.
- Columns with wrong data type such as timestamp represented as objects.
- Some tweets did not have favorite and retweet counts.
- Incorrect numerator and denominator ratings. One issue here was that some images included more than one dog and thus ratings were not out of 10 but rather multiples of 10 depending on number of dogs in the image. Another issue was that some tweets had more than one number group besides the rating which were then picked up when pattern matching, leading to wrong ratings. Such number groups as 24/7 or 50/50.
- I collapsed the four dog 'state' columns into one column by first concatenating the four dog 'states' for each row into a new `state` column, then using string methods to split the values into a list, removing duplicates in each list and finally dropping the four original dog 'state' columns.

After cleaning the `tweets` DataFrame, I lastly merged it with the `image_predictions` DataFrame using an `inner join`.

One issue that I did not address in this project was incorrect dog names extracted from the tweet text using regex. When assessing the data, I noticed incorrect dog names such as 'a', 'an' 'such' and more. To clean this column required a lot of effort which did not seem worth it as analysis could still be done on the DataFrame even with the incorrect names.

The final cleaned DataFrame was stored as a csv file called `twitter_archive_master.csv`