WeRateDogs Data Wrangling Report

14/03/2021

The dataset that has been wrangled and analyzed belongs to the twitter archive of the user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about and rating about the dog.

Datasets Wrangled and Analyzed

The following datasets were wrangled:

Twitter-archive-enhanced.csv

This is a locally supplied file that contains 2356 records of tweets, all of which have ratings.

Tweet_json.txt

The Twitter API was used to programmatically download the tweets as json data, the data we're interested in is the retweet_count and favorite_count that is not available in the locally supplied file.

Image_predictions.tsv

As above, this file was also programmatically downloaded.

Each of these data sets were loaded into individual data frames (df_archive, df_predictions and df_twitter) and separately cleaned – prior to merging all of them into a single data frame and outputting this as the final, cleaned dataset - twitter-archive-enhanced.csv.

Assessments Made

A total of *2 tidiness* and *15 quality* issues were identified with the data.

Tidiness issues were identified in the locally supplied data – loaded into the *archive* data frame.

1. Remove the individual dog stage columns and merge all into a single dog stage column.
2. Merge the three datasets after cleaning.

Quality issues were identified across all three data sources:

Via Twitter API:

1. id field was renamed to tweet_id so that all three datasets can be merged after cleaning operation.

Predictions data:

2. Remove the underscore '_' from p1, p2, p3
3. Capitalize the first letter of each word in columns: p1, p2, p3

Twitter Archive Enhanced local file data:

4. Using regex, match the ratings given in the tweets and fix the six incorrect numerator values
5. Merge the numerator and denominator fields to create a single rating column
6. Drop useless NaN columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
7. Split the timestamp field to two separate columns: time, date
8. Strip unnecessary HTML code from the source field
9. Delete 'None' from all rows in the 'name' field and replace with NaN
10. Delete the values from the 'name' field that matches with the following nine words: a; actually; all; an; by; quite; not; the; officially; and then replace with NaN
11. Convert the following fields from object to categorical data type: source, dog_stage

12. Replace '&amp', '&gt' and '\n' with &, > and a space, respectively, in the Tweets
13. Convert datatype to 'datetime' for the date column

After the merge, a couple of more cleaning operations were done to tidy up the column data types:

14. Fill blanks with 0 for the columns: retweet_count and favorite_count
15. Convert datatype to 'int64' for the columns: retweet_count and favorite_count

Final Thoughts

This was an interesting project! I liked the process of gathering data from online sources, both a static web address and via API – and merging with a locally available dataset, after cleansing has completed.