wrangle_report

December 16, 2022

0.1 Reporting: wrangle_report

1 Data Gathering

This project required gathering three data sets. The method used to gather each set of data was different and are as follows:

1. Twitter archive file: This can be downloaded manually or programmatically with the use of the Request library
2. The tweet image predictions: This can only be downloaded programmatically using the Request library because the file image_predictions.tsv is hosted on Udacity's servers and cannot be accessed manually.
3. Tweets: Each tweet's retweet count and favorite ("like") count at minimum and any additional data found to be interesting are scraped. This is done by -
    a. Extracting the tweet's IDs in the WeRateDogs Twitter archive and store in another file (tweet_id.txt)
    b. Querying the Twitter API for each tweet's JSON data using Python's Tweepy library and storing the data in another file (tweet_json.txt)

2 Data Quality Issues

1. A lot of missing data in the features (archive)
2. Change the datatype for some of the columns i.e. timestamp (archive)
3. Change invalid names to "None" (archive)
4. Expanded_url containing more than one url (archive)
5. Lowercase for P1, P2, and P3 occassionally (image)
6. Text column not properly formatted (image)
7. Extract the date from Created_at column (tweet)
8. Rename the Created_at column as Timestamp to bridge uniformity (tweet)

3 Data Tidiness

1. tweet_id is found across all three tables and should be merged into a single column

2. Dog stages should be in a single column rather than having the dog stage values as columns

A new data set named "twitter_archive_master" was produced by merging the three data sets named above on tweet_id. While uniformity of the column names is cruicial for readability, it is also important to enable merging of the data sets. Note that the source link in the archive table was deemed necessary to be extracted from the html tag so as to make is more usable in the browser.

4 Details

1. There is a lot of missing data in the archive table such that the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp columns were deemed to contain little to no meaningful data and were dropped.

2. As for the need to change datatypes, the following columns were addressed:

    a. archive: Timestamp is a datetime not an object
    b. archive: Tweet_id is a object and not an integer
    c. image: P2_dog is a boolean and not an integer
    d. tweet: Created_at is a datetime and not an integer

3. For the column name formatting, the column formally labelled as created_at was changed to timestamp in the tweet table

4. In the archive table, timestamp and source were properly formatted and rewritten to ensure readability and tidiness. These changes include:

    a. Removing the html tags from the source column
    b. Making the timestamp to contain year, month, and day only
    c. Choosing only the expanded url that follows the normal pattern

5. Lastly, two additional columns were engineered as it was deemed that they would be required to answer the research questions.