WeRateDogs Twitter Archive

Introduction:

In this project, we have to gather the WeRateDogs Archive data file along with Prediction File and we have to find the status of the tweet which is to be extracted in the JSON format using Tweepy library with API credential.

Process Used:

1) Gathering Data

a. First, we have to download twitter_archive_enhanced.csv WeRateDogs Twitter archive file for which I used data frame twitter_archive.
b. Secondly, we downloaded the image_predictions.tsv tweet image predictions file, having details like breed of dog etc in the data frame twitter_prediction.tsv. We downloaded file programmatically using the Requests library using the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
c. Thirdly, we query the Tweet Id from the Twitter Archive File using the Tweepy library using my API credentials and store each Tweet Id status in the JSON format in the file tweet_json.txt. Each tweet has retweet count and favorite count and some other additional data.
d. Reading this .txt file line by line into a pandas DataFrame and created a csv file df_json.csv having three fields tweet ID, retweet_count and favorite_ count.

2) Assessing Data

a. I created three copy of the original data frames i.e df_json_clean, twitter_prediction_clean and twitter_archive_clean.
b. We did two types of assessment Visual and Programmatic. In the Visual Assessment, I found some Quality issues-
    1. Some of the Tweet_Id's did not work using the twitter API in twitter-archive-enhanced.csv.
    2. Dogs name doesn't match in "Text" and "Names" in twitter-archive-enhanced.csv.For some fields Missing values in 'name' and dog stages showing as 'None' in twitter-archive-enhanced.csv.
    3. There are some duplicate url in expanded_urls field. There are some tweets with has no images. Dataset contains retweets.
    4. One tweet having rating denominator as 0 and image, dog name as none. Some names contain special unicode characters

like FrÃ¶nq. In the Prediction File, names of dog starts with lowercase.

c) In the Programmatic Assessment, I checked that there are four fields for Dog_stage instead of one and I have to join prediction and df_json to archive file. I checked the info(),describe(),tail() and head() methods of the dataframe and found in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id and retweeted_status_user_id are float datatype which should be str, changing datatype of rating_numerator and denominator to float. Replace none with NaN in dog_stage column. Removing rows which have no images.

3) Data Cleaning

I removed all rows having no images.Changed the data types of the fields which are identified in the assessment part. Replaced none to NaN for dog_stage field.Changed the date field data type to timestamp from object. Took only that Tweets which has no retweet and changed the Dog name from lowercase to Title().Finally, joined the prediction and df_json file with the archive file.

4) Master File

Created the twitter_archive_master.csv from the twitter_archive_clean dataframe.