Wrangle report for tweet archive of Twitter user @dog_rates data

Author: Khawla Alqarni
Date: Dec 12, 2022
Purpose: document data wrangling process including gathering, assessing, and cleaning data

Contents

Environment and Tools

The data wrangling process is peformed in the Jupyter Notebook in workspace of Udacity. The libraries used in this project are pandas, requests, tweepy, json, and matplotlib.pyplot,. %matplotlib inline is added for direct outputs in the notebook. pd.options.display.max_colwidth = 100 is set for avoiding text collapes.

Data Gathering

The datasets for this project are from the tweet archive of Twitter user @dog_rates (WeRateDogs).

1. Enhanced Twitter Archive: contains tweet data for all 5000+.

File name: twitter-archive-enhanced
Format: csv
Source: directly download from Udacity website.

2. Image Predictions File: the output from neural network

File name: image-predictions
Format: tsv
Source: get the data from url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'

3. Additional Data via the Twitter API

File name: tweet_json
Format: txt
Source: connect Twitter API to download json format text file and use pandas to read into the notebook.

Note: I read this data directly from tweet_json.txt file provided by Udacity because I don't have access to Twitter API.

Data Assessing and Cleaning

Quality

Enhanced Twitter Archive table:

1. Datatype of 'timestamp' is object not datetime
2. Datatype of 'tweet_id' is integer not object (string)
3. Change 'name' into lowercase for standardize the format
4. The data contain retweets and should be removed
5. 'source' column should contains only device name, but it contains HTML code and URL
6. Names of dogs contains wrong names and should be removed
7. Denominator have wrong numbers, should be 10
8. Remove Unnecessary columns like: in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id retweeted_status_timestamp and expanded_urls

Image Predictions:

9. Tweet_id should be string
10. Reduce prediction and confidence columns into two columns.

Additional Data via the Twitter API :

11. Id should be renamed as 'tweet_id' to normalize the format for all three datasets
12. Id should be string

Tidiness

1. dog stages (doggo, floofer, pupper, puppo) should be transposed into one column, dog_stage.
2. all data related to each other, so we should merge them
3. rating_nemerator and rating_denomirator should be reduced to one column (numerator/denominator)