Wrangle Report

July 21,2020

Gather

Data were collected from different resources. First resource is archive twitter enhanced, got it from Udacity and put it in the same file location as a CSV file. Second resource was a TSV file that could be downloaded through a URL and read it with a separate **sep=’\t ‘.** Third file which is extracted from Twitter API using Tweepy library and making authentication to save each tweet in dictionary. and each tweet has features defined in a dictionary. Referring that data was saved in text file.

Asses

Managed to convert timestamp to three columns, each one of them represent day or month or year. Numbers of archived data and predicted data did not match. There were duplicated values of tweets_id. Knowing the data types of each column and number of values in it using info() function. All not needed columns should be removed and duplicated ids should be removed but let one lasts.

Clean

First, I made a copy of the three data frames, then try Removing duplicated in tweets id using

df_twitter_clean = df_twitter_clean.drop_duplicates(subset='tweet_id', keep='last')

then I tried to gather four types of dos **in dogs_stage column**. Then I made new rating column and remove the previous two based on the next equation:

$$rating_{total} = \frac{rating_{numerator}}{rating_{denominator}}$$

Also removed the duplicated jpg_url from the prediction clean dataframe. Then Worked on Json clean data frame to choose only original tweets from **WeRateDogs** Through:

df_json_clean = df_json_clean[df_json_clean['retweeted_status'] == 'Original tweet']

While Merging Json text file and the first merged data frame I found that tweet ID need to be changed to object then Merged them successfully. After all, I used the final data frame **df_twitter_merge2**

to maintain answers of many questions on my head, also making good visualizations.