wrangle_report

November 20, 2020

I gathered three datasets related to the rating of dogs on a twitter account: @WeRateDogs.

The first one is the archive tweets provided by @WeRateDogs: "twitter_archive_enhanced.csv". I imported this dataset by pandas function "read_csv()". It contains general information of their tweets, including tweet_id, tweet text, rating information. However, it doesn't contain the number of retweets and number of favorites.

To acquire these additional data, I used *tweepy* to query each tweet_id in "twitter_archive_enhanced.csv" in twitter API. This is the second dataset. Specifically, I generated the authorization handler of twitter API by *tweepy*, then I iterate the tweet_id in the archive dataset to query the current status of tweets. The status data are stored in json format, which could be written into a .txt file. In order to read status data properly in the next step, I started a new line after writing each status data. Because some tweets are deleted, I used try and except block to address the error message. In the "tweet_json.txt" file, each tweet status data is stored in different line. A list of tweet status data generated by function readlines() was iterated. The information of retweet and favorite counts were extracted and appended to a list of dictionaries. After iteration, this list was converted to a dataframe.

The third dataset is the breed prediction of dogs based on the images through a neural network. This dataset was acquired by requesting Udacity's server. Then I save it as a tsv file.