Data Wrangling and Analysis

The raw data comes with two parts (1)Enhanced Twitter Archive and (2)Image Prediction, the first one recorded all the basic information like timestamp, tweet id, and the full URL etc..,
and the second one used the neural network(CNN) to predict the breed of dogs on the tweets.

The additional data for the like and retweet number was collected through web scraping with twitter's API, it's stored in a file called tweet-json.txt.

Gather

I read in all the three files: (1)Enhanced Twitter Archive (2)Image Prediction (3)tweet-json

Assess

I checked the data manually and programmatically, and briefly documented some issues inside these files as below. The retweets were dropped down first, as it can save some time for the coming problem fixing process.

Quality

main_data Table

- 181 retweets that are not necessary for analysis
- modify denominator and numerator datatype (object -> integer)
- some of the rating denominators are greater than 10
- weird dog names ex.a, all
- timestamp is not datetime data type
- in_reply_to_status_id & in_reply_to_user_id are float
- tweet id 892420643555336193 has 0 as its rating denominator

retweet_like Table

- inconsistent id column name

Tidiness

- image_prediction:Only keep one dog breed prediction that has the highest confidence level
- main_data:Dog stage should be organized in one column

Clean

All the problems listed below was fixed one by one with problem define, code, and test.

Delete Retweets

(1)181 retweets that are not necessary for analysis

Define: All the rows that have retweeted_status_id & retweeted_status_user_id & retweeted_status_timestamp columns were deleted as we only want the original posts.

Tidiness

(1)image_prediction:Only keep one dog breed prediction that has the highest confidence level

Define: Using if statement to build a function to select the most reliable dog breed prediction from three tests.

(2)main_data:Dog stage should be organized in one column

Define: Replace all the None in the dog stage columns and concatenate all the strings in one column. Filter out dogs in multiple stages and separate them with a comma.

Quality

(1)modify denominator and numerator datatype (object -> integer)

Define: using astype to change denominator and numerator datatype to integer

(2)main_data:some of the rating denominators are greater than 10 (pick the first five record to fix)

Define: Reduce the fraction or delete the rating record if the number is not relevant

- 832088576586297345: it's a date not rating (delete)
- 820690176645140481: reduce fraction (84/70 -> 12/10)
- 758467244762497024: reduce fraction (165/150 -> 11/10)
- 740373189193256964: rating should be 14/10
- 731156023742988288: reduce fraction (204/170 -> 12/10)

(3)weird dog names ex.a, all

Define: Use islower() to filter out all the lower case weird names

(4)*timestamp is not datetime data type*

Define: using astype to change datatype to datetime

(5)not useful values (in_reply_to_status_id & in_reply_to_user_id *)*

Define: remove in_reply_to_status_id & in_reply_to_user_id columns

(6)tweet id 892420643555336193 has 0 as its rating denominator

Define: delete a post from id 835246439529840640, since it has an invalid rating denominator (0)

(7)*inconsistent id column name*

Define: rename id column in retweet_like table to tweet_id to match all the other tables