Wrangle Report

Lin Chen, 2021-05-05

Step 1, data gathering

1. Using read_csv of Pandas to extract data from WeRateDogs Twitter archive (enhanced).
2. Save the online tsv page of Twitter image predictions into csv file and then open it.
3. With the help of Twitter API, download web contents from WeRateDogs and save them into a txt file. Then extract retweet and favorite counts for each Twitter from this file.
4. Data are stored in df1, df2 and df3 respectively for the three steps mentioned above.

Step 2, data assessing and cleaning

Quality issues:

1. Retweets shall be deleted since they contain duplicated ratings to original tweets.
2. Some ratings are given by replies, but recipients' id and other info can be deleted, to indicate the rating is given on original post or in reply, use a new column with 0 and 1 to show this info.
3. Wrong datatypes shall be fixed for all dataframes.
4. The column of info source could be simplifies by using some simple words rather than verbose sentences.
5. Duplicated contents shall be dropped for all dataframes.
6. Lots of dog names are mistakenly extracted and should be fixed to enhance accuracy.
   Those wrong names are usually 1) start with lower cases, 2) contain only one single upper case, 3) be categorized into None.
7. Some ratings shall be corrected and some of them shall be standardized.
8. Rows with more than one dog categorizations shall be fixed, since some of them are just result from mistakenly extraction. Some of them has many dogs in the photos, which shall be fixed as a tidiness issue.

Tidiness issues:

1. Cells have multiple URLs of different categories shall be fixed, use two columns instead of one.
2. Four dog category columns shall be unified to one column, since attribute values cannot be used as column names.
3. For tweets that have multiple dog species, additional rows are required for replicate and separate, otherwise one dog specie cell will have multiple values.
4. Combine all three dataframes into one on the common column tweet_id.