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

There are multiple quality and tidiness issues observed, especially for the given twitter archive (df1).

Quality issues:

1. Retweets shall be deleted since they are not written by WeRateDogs, and the ratings inside are replicates from original WeRateDogs ratings.
2. Some useless columns can be deleted, such as in_reply_to_status_id and in_reply_to_user_id. Those information are not needed for further analysis, and we could add a column called in_reply to indicate whether a rating is given as original post or reply to others. In this new column, original post can be represented by 1 while 0 is used for replies.
3. Wrong datatypes shall be fixed for all dataframes, such as timestamp shall be datetimes, all id values shall be strings, not numbers.
4. The column source could be simplifies by using some simple words (iPhone, Web Client, Vine and TweetDeck), rather than verbose sentences.
5. Duplicated contents shall be dropped for all dataframes, and there are lots of duplicated URLs, so I build two new columns twitter_url and extra_url, one for Twitter urls and another for other urls such as GoFundMe links.
6. Lots of dog names are mistakenly extracted in the original csv file. Those wrong names are usually 1) start with lower cases, 2) contain only one single upper case, 3) be categorized into None. Those rows are taken out separately, and their names are extracted again by get the word after the word “named” in the original text. “None” will be given to those Twitters without a word of “named”. It’s still pretty raw but at least could enhance accuracy for a little bit.
7. Some ratings shall be corrected and some of them shall be standardized. All denominator shall be 10 or 10x, and after fixing denominators, then extreme numerators shall be fixed. Some outliers shall not be fixed for fun (like the 1776 and 420 one), and for those photos that are not dogs and given low ratings, they shall not be deleted since most photos with ratings under 8 are not dogs.
8. Rows with more than one dog categorizations shall be fixed, since some of them are just result from mistakenly extraction, and some of them has many dogs in the photos. Those wrong ones shall be fixed at this step.

Tidiness issues:

1. The dog category columns shall be unified to one column, since in original csv file, four columns are used to represent one attributes. For those photos that have multiple categories of dogs, a duplicated row can be created, and such circumstances are actually rare.
2. Many repeated IDs shall be removed, especially in df3, the dataframe extracted by Twitter API from WeRateDogs directly. After deleting repeated rows, there are still few left with various data for one single Twitter. The rows with wrong information shall be deleted.
3. All three dataframes shall be combined into one large dataframe by using merge in Pandas on tweet_id, and here we need to drop duplicate rows again and reset index. The final dataframe shall be saved into a csv file Twitter archive (master). The brand-new csv file will be used for further analysis.

Step 3, data analysis

Results can be found in another report (act report), and details can be found at the wrangle_act.html.