wrangle_report

March 9, 2019

1 Data Wrangle Report

2 Gather

2.0.1 There are three tables in the system.

- (1) `twitter-archive-enhanced.csv` is a given CSV file that contains tweeter ID, URL, rating, description, reweetered status and dog stage. I loaded this file into dataframe df, the main dataframe in this project. There are 2356 rows in this dataframe. It has null values.

- (2) `Image_prediction` is a TSV file that needs to be downloaded from the Internet. I used Requests library and loaded it into a dataframe called df_image_prediction. Each df_image_prediction row contains a URL to the dog image and the probabilities of three possible breeds for this dog. This dataframe has 2075 rows. It does not contain null value.

- (3) Last but not the least, I am required to get the retweet count and favorite count for dogs in `df`. These two values are only available via the Tweepy library. I followed the instruction and created a Twitter developer account. I created a new project and obtained a set of customer key/secrete and token from Twitter. I was successfully connect to the auth and api. However, I found all Tweeter IDs in df except one are no longer available on the Twitter website. This may mean the Twitter IDs in df are removed because the they are the Twitter ID from 2017.

Because of most of the tweeter IDs in df are removed from the website, I decided to use the provided `tweet-json.txt`, a json file contains retweeted count and favorite count for the tweeter IDs in `twitter-archive-enhanced.csv`.

I loaded `tweet-json.txt` and obtained a dataframe. There are so many fields in this dataframe, many are very interesting. Due to the time I have is limited, I extracted three columns, tweeter ID, retweeted count, and favorite count. I called the extracted dataframe `df_tweet_of_interest`. There are 2354 rows in `df_tweet_of_interest`. There is no null value in this dataframe. All tweeter IDs are unique.

3 Assess and Clean

df

I opened df. I never looked at data relating to dogs so closely. According to the introduction, it is okay that the rating numerator is greater than the denominator. Most of the ratings follow this pattern. However, I noticed a few ratings are out of the norm, i.e. 5/10, 960/0 etc. Some ratings are extracted incorrectly. I picked two incorrect ratings and corrected them.

There are several dogs missing names or having wrong name according to the text description. I picked two incorrect names and corrected them.

The 'expanded URLs' field is a concatenation of several URLs. It has a lot of repetitive URLs delineated by commas. I splited the first and second URLs into two columns, 'expanded_URL1' and 'expanded_URL2'. The third and fourth URLs are repeating ones so I did not store them. If the original field has only one URL, 'expanded_URL2' is null.

I noticed several columns have incorrect data types. For example, 'retweeted_id' should be integy. 'rating_numerator' and 'rating_denominator' should be float. I have corrected the three columns mentioned above. df_image_prediction #### img_num is the ranking of image. The range is from 1 to 4. All data falls into this range. There is no issue in img_num. #### The sum of p1_conf, p2_conf and p3_conf should not be greater than 1. There is one row the sum is over 1. I have corrected it by keeping p2 and p3 and modifying p1 so their sum is 1. #### Except tweet ID, there are duplicated rows in this dataframe. I have removed the duplicated rows.

4 Tidiness

The four stages in df are splitted into four columns. I created 'stage' column to store the stage and 'stage II' column if a row has two stages.

Last, df, df_image_prediction and df_tweet_of_interest can be combined. The final master dataframe is a merge of the three dataframes above.