DataWrangling

March 4, 2018

In this data analysis project there were three different sources of Data. One was the twitter-archive-enhanced.csv which contained 2356 observations. The dog names , ratings were scraped from the text of the tweets. Next was the image prediction file image_predictions.tsv which has the predictions of the dog breed and confidence percentage of the predictions. Finally , the last dataset was created by querying the twitter API to get few attributes like retweet_count , favorite count for the tweets that were present in the twitter-archive-enhanced dataset.

My data wrangling effort started with visually assesing at the data in the twitter-archive-enhanced.csv in excel. I saw that there were many dogs whose names were “a”, “an”, “the”. There was a problem with the ratings of the dogs also. I observed that for few dogs there were multiple fraction values in the Text field and the 1st fraction available was assigned to the ratings. For instance there was a 3 1/2 legged dog which had a rating of 9/10. But 1/2 was the 1st fraction in the data the rating were considered to be 1/2. There were many entries with None in the dog types columns eventhough the dog type was present in the Text field. The retweet related fields like retweetid, retweeted_status_user_id and retweet_timestamp had data related to a tweet that has been retweeted. Since we already have an entry with the tweet that has been tweeted ,having an additional entry with retweet was just a duplicate.

I started my data cleaning process by making copies of the original three dataframes. Then I started with correcting the dog names programatically. Firstly i cleaned all the dog names that were starting with lowercase alphabets to None. Then Once this was taken care of there were few cases when the dog names were present after the word “Named” in the text of the tweets , tried to get the names of the dogs from the text after the keyword named. There were 4 columns in the data with the dog types. Since each dog can be of ony one type I converted this to a single column called dog_type and found the type of the dog from the text field. Then I shifted my focus to the retweet columns as observed these were just creating duplicates, so I deleted all the observations for which the retweet columns were not null and then dropped these columns from my dataframe. Then I shifted my focus to the ratings column which had few problems like incorrect ratings when the rating was in decimals. To correct this issue I converted the rating_numerator and rating_denominator to decimal datatype. There were few observations where the text field had multiple fractional values which were misconstrued as ratings . I corrected ratings for these records and updated proper ratings.

Once the data cleansing was completed. I combined the twitter-archive-enhanced, image prediction and programatically downloaded tweet file from API on the tweet_id. Then I loaded this dataframe to a twitter-enhanced-master.csv file. After creating this file. I have proceeded with my data analysis part on the we rate dogs data which I had assesed , cleansed and stored.