Wrangle and Analyze Data

Sourojyoti Paul

Wrangle Report

Synopsis

Along the Data Wrangling process, in the twitter_archive_enhanced.csv file, we found several problems in the dog's name column, probably the regex used to gather/nd it (from the Twitter user @dog_rates also known as WeRateDogs™ (https://twitter.com/dog_rates)) was not well calibrated, and in many cases has gathered articles, nouns, etc. or any other ordinary word. I have xed it assuming these problematic dog's names as None .

We also found problems in rating_numerator and rating_denominator columns, both from image_predictions.tsv file, which has required a new process of "scrapping" these values from the text column. Finally, combined the files twitter_archive_enhanced.csv and image_predictions.tsv into a new data frame called twitter_archive_master.csv, which we have aggregated some new features:

- Retweet_count
- Favorite_count

Both features are gathered from the WeRateDogs™ tweets using the tweepy package.

1. Introduction

This Wrangle Report is a part of a Data Science Course Project offered by Udacity. The project aims to gather data from Twitter and combine it with a third party data frame to create analysis about the tweets and the predicted dog's breed.

2. Data Gathering

We have gathered the files image_predictions.tsv and twitter_archive_enhanced.csv using the requests package. Although the image_predictions.tsv file has almost all the information from the WeRateDogs™ user, there is some missing variable, which has been gathered using the tweepy package.

3. Data Assessing

Quality issues

twitter_archive_enhanced.csv

1. Invalid names or non-standard names.
2. Invalid ratings numerator. Value varies from 1776 to 0. Data Structure must be converted from int to float.
3. Invalid denominator, I expected a fixed base. Data Structure must be converted from int to float.
4. timestamp is a String, It needs to be converted to date.
5. tweet_id must be a string.
6. retweeted_status_id : The same dog could be recorded twice or more in cases of retweets.
7. in_reply_to_status_id : The same dog could be recorded twice or more in cases of reply.
8. source column is having HTML tags, URL, and content in a single column.

image_predictions.tsv

9. Columns p1,p2 & p3 : Dog's breed has no standard. Capital letter or lowercase names.
10. tweet_id needs to be converted as string
11. Column jpg_url has duplicated images and consequently double entry.

Tidiness issues

twitter_archive_enhanced.csv

1. doggo, floofer, pupper, and puppo. These are categorical variables, can be combined into one column.
2. text : There is two information in a single column. need to split the text from the URL.

df_ach : Loaded data frame from twitter_archive_enhanced.csv
df_img : Loaded data frame from image_predictions.tsv
twt_ach_mstr : Loaded data frame from twitter_archive_master.csv

4. Data Cleaning

The dog’s name issue was solved by evaluating if it starts with a capital letter. It was a name, if not it was an ordinary word and I have converted it to “None”. Most of the issues involving non-usual values to rating_numerator and rating_denominator were solved using a new tailored regular expression to gather the ratings from the text column. In respect to the data type problems in timestamp and tweet_id columns, were fixed using the .astype() method and .loc[] . In regard to the duplicated information, I decided to remove all retweets and reply to avoid double entries of the same dog. Finally, I have solved the tidiness issues combining the tables twitter_archive_enhanced.csv and image_predictions.tsv in one called twitter_archive_master.csv .We also have merged 4 columns (doggo, pupper, puppo, and oofer) into one, which have been bundled and named as dogtionary.

5. Conclusions

We have documented 13 issues but this final file version is not totally free of issues, because I faced the Data Wrangle as an iterative process, what has been done so far was the first iteration. For this reason, the twitter_archive_master.csv file is the final file version with a minored number of issues, and ready for a Data Analysis. This file has 1968 observations and 24 features. Caveats.
Bear in mind, there are some tweet_id that do not have retweet_count and favorite_count , which means there are observations with NaN.