WeRateDogs Data Wrangling Report

June 2019

Nasim Khadem

Introduction

This project is about WeRateDogs twitter archive which needs to be cleaned and stored on file.

Gather

The first step of the process was to gather data from three different sources.

- **Manual Download (df)**

First Task was simply downloading the WeRateDogs Twitter archive and open it using Jupyter notebook. This file was read into pandas as df.

- **Programmatic Download (idf)**

The data needed to be downloaded programmatically from Udacity cloud using python requests library. This file was read in to pandas as idf.

- **Twitter Api (jdf)**

The last source was to get the data using twitter api. I used the tweet ids from the image prediction file to get retweet count and favourite ("like") count. I made a data frame from tweet_json and I saved the data frame to file so I can read it easily anytime This file was read in to pandas as jdf.

Assess

Assessing the data was done both through visual assessment and programmatic assessment. There were both quality and tidiness issues in all the data frames.

Quality

Missing values in many of the many of the columns of twitter archive. More exactly in :
”in_reply_to_status_id"
"in_reply_to_user_id"
"retweeted_status_id"
"retweeted_status_timestamp"
“retweeted_status_user_id"
“expanded_urls”

There were also consistency in the missing values. The missing value was sometimes Nan and sometimes None. This issue could be seen in :
“doggo” , “floofer” ,“pupper” ,“puppo” and “name"

There were also wrong values in column name for example “a” , “an”, “the”, ”one”.
There was also timestamp values that were represented by strings.
Also values retweets had to be filtered out.
The values at p1 and p2 and p3 of jdf data frame were sometimes uppercase.

Tidiness

Different stages needed to be one columns and not four.
Also three different data frames were unnecessary . They could all be on table.

Clean
At first copies of each of the data frames was made. This ensures that at any point the original data will stay the same.

I removed the duplicates tweet_ids which I had in Jdf. I made this to ensure that while joining I will not have duplicates value in the final table.

I removed the retweets from the df data frame. Since the retweeted values were removed the columns that were related to retweets became all Nan. These were also removedThen removed the corresponding tweet_ids that were in the idf data frame. (I understand this step was unnecessary cause with the inner join of two data frames this was taken care of.)

I then made a new feature (column) named “stage” which melted the four stage columns “doggo”, “floofer" , “pupper” and “puppo”. Sometimes however there were multiple combinations , like “doggopupper” . This happens when we had pictures of more than one dog to be rated.

Also the if the columns have not been removed the issues at assess step were addressed.

Each step in the clean process has three stages Define (Problem and fix), Code and Test.

Save
The clean data frame df_clean was saved to disk under twitter_archive_master.csv

Future work
The name feature can be extracted with better values. Also different data could be downloaded through Twitter api to enhance our findings. Another thing that can be added is the gender of the dogs. At last a model can be built to predict the dog breeds.

References
https://twitter.com/dog_rates
https://www.slickremix.com/docs/how-to-get-api-keys-and-tokens-for-twitter/
https://stackoverflow.com/questions/22676/how-do-i-download-a-file-over-http-using-python
https://github.com/kdow/WeRateDogs
https://github.com/wanderly0501/Data-Wrangling-of-WeRateDogs-Tweet-Archive
https://github.com/nishchaychawla/Data-Wrangling-of-WeRateDogs