We Rate Dogs

Module 4: Project

Vootkoor Samhitha

July 2020

Documentation for data wrangling steps: gather, assess, and clean

---

Gathering Data for this Project

The data is gathered in three steps as mentioned below.

1. The WeRateDogs Twitter archive. It is provided by the Udacity and is downloaded manually.
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and is downloaded programmatically using the **Requests** library and the following URL: **https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv**
3. Finally, with the help of twitter api and using tweepy library the required favorite_count and retweet_count are extracted for the required twitter id's. Twiter API is authorised so this step also involved getting auth keys.

Assesing Data:

Each data set is thoroughly assessed and the following quality and tidiness issues are observed and noted.

Quality Issues

Twitter Archive dataset

- certain dog names makes no sense for example 'a','an'

- timestamp datatype is object but it should be datetime datatype also remove the last +00000
- tweet_id should be object not int
- Source column has anchor tag but we need only the source of the tweets
- certain columns like in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp has lot of missing data and it would be preferably to drop these columns to make the dataframe look complete
- The maximum and minimum denominator values are not 10

Image Predictions dataset

- tweet_id should be object not int
- The dog breed names which consists mutiple words are somtimes separated with - and sometimes with _
- Some dog predictions in p1 makes no sense so compile the actual dog breed from p1,p2 and p3 into one single column to derive the actual breed of the dog
- The names are also sometimes lowercase and sometimes upper case
- Convert the datatypes of Source and Dog Type as categories

Tweets dataset

- The tweet_id has to be object

Tidiness Issues

- combine the different types of dog stages present in different columns as a single one in twitter_archive dataset
- The entire data should be in a single dataframe as they propagate the same purpose
- Drop the unnecessary columns

Cleaning Data:

After documenting the issues the data is cleaned systematically by following three steps

1.Define: What has to be done

2.Code: Programmatically cleaning the code

3. Test: Checking whether the issue is resolved

- All the issues are resolved programmatically.
- Wherever necessary customized functions are written for example getting the dog breeds and changing the improper dog names to 'None'.
- The functions such as melt and merge are used wherever necessary.
- Regular expression is used to extract the source from the url which came with an anchor tag.
- Data types are converted as required
- All the data quality issues and tidiness issues are resolved

Conclusion:

The entire data wrangling process is concluded by creating clean data set twitter_archive_master.csv which can be used in future for further analysis.