wrangle_report

August 14, 2020

Wrangle Report
by Kueh Seow Teck

0.0.1 Introduction

For this project, I will be wragling and analyzing and visualizing) a dataset from the tweet archive of a popular Twitter account user @dog_rates, also known as WeRateDogs. The account features dogs along with humorous comments. The tweet also include a rating for these dogs which always ahave a denominator of 10. This Twitter account garners about 8.8 mullion followers up to date.

The dataset which I will be using for this project is sent by WeRateDogs to Udacity via email which includes 5000+ tweets with an end date of August 1,2017. Other than that, I will also be querying extra data (retweets and replies numbers) from their account through the Twitter API. Last but not least, Udacity also provided me with image prediction results derived from neutral network through the input of the dog images from these tweets.

0.0.2 Tool used

1. Below are the Python libraries I have used for this project:
    - `pandas`: importing flat files and wrangling the tables
    - `numpy`: running mathmatical operations
    - `requests`: use for making HTTP request to download the image predictions results in the form of .tsv file from a URL.
    - `os`: navigate through the directories
    - `json`: for reading json data collected in .txt file through Twitter API
    - `tweepy`: for accessing Twitter API
    - `matplotlib.pyplot`: for data visualizations
    - `re`: for regular expressions
2. Other than the Python libraries, I have also registered myself a Twitter account and setup a developer account. From there, i created an app in order to generate the Consumer API keys, and Access Token and Access Token Secret which will be use to access the Twitter API.

0.0.3 Tasks for this project:

1. Data wrangling, consisting of :
    - Gathering data
    - Accessing data

- Cleaning data

Note: Data wrangling process is completed in the Jupyter notebook: `wrangle_act.ipynb`

2. Analyzing and visualizing the wrangled data
3. Create two written reports:

   - `wrangle_report.pdf`: Describe my wrangling efforts
   - `act_report.pdf`: Communicate my insigts and displays the visualizations(s) produced from my wrangled data

0.0.4 Gathering data

The data gathering process involving loading data from three different sources. Below are the details of the data gathering process:

I gathered the first dataset by manually uploading the twitter-archive-enhanced.csv file provided by Udacity into my Jupyter workspace in the Udacity Classroom. I then loaded the .csv file into Jupyter notebook using the Python library.

For the second dataset, I was requested to download the file from Udacity's server from the URL https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv by using the Requests library. After that, I load the .tsv file into my Jupyter workspace using Python library.

Lastly, I gathered the third dataset from WeRateDogs Twitter archive using Tweepy and stored them in a JSON file (tweet_json.txt). Then, I read the file using JSON library and load it as a dictionary before converting it to a Pandas dataframe object by using Pandas library.

0.0.5 Accessing data

I did the initial accessment of the data visually by using head() and sample() functions. After identified some of the key issues, I dive deeper through programatic accessment using the info(), value_counts() etc functions.

I identified and classified issues into two groups: data quality issues and data tidiness issues:

Quality Issues

Here is a list of data quality issues I have identified from the three datasets: ##### `twitter_archive_enhanced` table - retweeted_user_id and retweeted_status_id columns: Replies and retweets among the entries - expanded_urls column: 59 Missing expanded_urls - tweet_id column: tweet_id is an integer instead of a string - timestamp column: timestamp is a string instead of datetime - rating_denominator column: rating_denominator less/more than 10 (should be standardize to 10) - name column: Over 745 tweets have no name with 55 named 'a'/'an'and some with really short names; Below are the detailed observations from these rows - Bo, Jo, Ed, JD, Mo are actual dog names - The actual name of the dog with the name Al is actually Al Cabone - The actual name of the dog with the name O is actually O'Malley - Below are some of the corrected dog names extracted manually from the text column of the rows with name a/an: (Example: Index name) - 1955 Kip - 2034 Jacob - 2066 Rufus - 2116 Spork - 2125 Cherokee - 2128 Hemry - 2146 Alphred - 2161 Alfredo - 2191 Leroi - 2218 Chuk - 2235 Alfonso - 2249 Cheryl - 2255 Jessiga - 2264 Klint - 2273 Kohl - 2287 Daryl - 2304 Pepe - 2311 Octaviath - 2314 Johm - 2204 Berta - source column: long string; can be simplified and data type is string instead of categorical - pupper,puppo,floofer and doggo columns:Around 1976 tweets having no dog stage data (all four stages showing 'None') - rating_numerator and rating_denominator columns: these columns are not extracted properly from text

`image_predictions` **table**

- tweet_id column: tweet_id is integer instead of string
- p1,p2,p3 columns: dog breeds have inconsistent lower/uppercase
- img_num: img_num is float instead of integer
- p1, p1_conf columns:I only want to keep columns with highest confidence level

Tidiness Issues ##### `tweet_json` table - Various stages of dog (ie doggo, floofer, pupper,puppo) in the `twitter_archieve_enhanced` table should be one column

`twitter_archive_enhanced` **table**

- `tweet_json` and `image_predictions` should be merged with the `twitter_arvhive_enhanced` table as they are one observation unit.

0.0.6 Storing data

Once data wrangling process is complete. The final cleaned data is stored in the file `twitter_archive_master.csv`