Wrangle_report

April 13, 2021

1 Data Wrangling Report

1.1 Step-1(Data Gathering)

Basically for this project, I gathered data from three different sources as listed below. I used a different method of data gathering for each data source.

- Importing data via CSV
- Using requests to download data from the internet
- Scraping data from an API

Three data sources

- Enhanced Twitter Archive

The WeRateDogs Twitter archive is provided by Udacity. This contains basic tweet data for all 5000+ of their tweets. I downloaded this file manually by clicking the link.

- Image Predictions File

The tweet image predictions, i.e., what breed of dog (or another 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 should be downloaded programmatically using the Requests library and the following URL: image_predictions.tsv

- Data via the Twitter API

Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. I did not include my Twitter API keys, secrets, and tokens in the project.

1.2 Step-2(Assessing data)

```
After gathering the data, I used a different method to check the data. I came up wi
```

```
Tidiness issues
```

1.Combine three different dataframes into one master data set(Information about one type of observational unit (tweets) is spread across three different files/dataframes.)

2.There are 4 columns for dog stages (doggo, floofer, pupper, puppo). The 4 columns for one variable doesn't conform to the rules of "tidy data".

```
Quality isuues
```

3.Remove unwanted columns (retweeted_status_timestamp, retweeted_status_user_id,retweeted_status_id) and clean the duplicate rows and NaNs

4.Clean text column to get dog gender

5.Clean sources columns, which is difficult to as such

6.Refine respective predictions and confidence columns

7.Drop columns with one low values or similar values

8.Fix numerator and denominators

9.Convert NaNs/Nulls to None

10.Fix datatypes of various columns

1.3 Step-3(Cleaning Data)

For data cleaning, most of the content I have learned from udacity classes and the internet i.e. google, StackOverflow, etc. to sort out the above-mentioned issues. I dropped some unused columns to save my time. There were a lot of difficult cases where I had to use regular expressions. For every issue, I followed the steps such as Define the problem, code to resolve it then test if it works.

After cleaning the data I created and saved the master file for further visualizations and analysis.