wrangle_report

October 25, 2022

0.1 Reporting: wragle_report

1 Gathering Data for this Project

This project involved gathering of data from three different sources as listed below. For each of the data source a different method of data gathering was used:

-Importing data via csv -Using requests to download data off internet -Scrape data from an API

This was challenging and fun at the same time.

My wrangling efforts for the WeRateDogs Twitter project included gathering data from the following sources:

The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided . This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood onAugust 1, 2017.

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 was provided.

Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data I find interesting.

2 Assessing Data:

Once the data was gathered, I began to assess the data on both quality and tidiness issues.

3 Quality issues

3.1 Archive table

1.source column is in HTML-formatted string,not a normal string

2.remove retweets

3.error in dog names: None values

3.2 Image table

4.remove duplicate jpg_url entries

5.remove entries that have p1_dog&p2_dog&p3_dogs values set to false

3.3 Api table

6.remove retweets
7.source column is in HTML-formatted string,not a normal string

3.4 All Table

8.convert data type of tweet_id to object string

4 Tidiness issues

1.Dog stage are spread in three columns in archive table.
2.merge archive,image and api table

5 Cleaning Data

I used my knowledge of python and searching over the internet i.e. google, stackoverflow, stack-abuse github etc for references and possible guideance to resolve the above mentioned issues to the best of my knowledge.

Overall, I learned a lot about how to use python effectively and efficiently to clean data and store it.

Finally, once the data was ready I analyzed it using visualizations as documented in act_report.html

```
In [ ]:
```