wrangle_report

November 24, 2020

1 Introduction

This document include a brief report on the efforts that were made in the wrangle and analyze data project.

Data Wrangle have 3 main steps: * Gathering Data * Assessing Data * Cleaning Data

1.0.1 Gathering data

Project data was gathered in three steps:

- First: Twitter archive data was downloded directly from the instructure notes.
- Second: Image predicitions data was gathered using the request librabry and the provided url in the instructure notes.

  url='https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'

```
response = requests.get(url)
response
with open(os.path.join('image_predictions.tsv'), mode = 'wb') as file:
file.write(response.content)
```

- Third: Additional desired data on the tweets in twitter archive data was gathered using twitter API using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt. (The used code for gathering is provided with file name tweeter-api.ipynb)

1.0.2 Assessing

In this step, data was inspected for two things: *data quality issues* (i.e. content issues) and *lack of tidiness* (i.e. structural issues).

Assessing data can be done programmaticaly using pandas, and in this project data was assessed programmaticaly, sample of the used methods: .head() .info() .describe() .value_counts() .nunique()

And while assessing our project data, the issues were found are:

Quality issues:

- Missing data in in_reply_to_status_id.
- Missing data in_reply_to_user_id.
- Missing data retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.
- Missing data expanded_urls.
- 745 dogs with None as their name.
- tweet_id is int not str.
- timestamp is object not datetime.
- rating_denominator sometimes doesn't equal 10.
- rating_numerator is extracted wrong (sometomes less than 11).
-

1.1 0 nun_null values for profile color column.

- 1 varibale (dog_phase) in 4 columns (doggo, floofer, pupper, puppo).
- tweet_df should be part of twitter_archive_df.
- two variables in one column (day and date) in tweet_df.

1.1.1 Cleaning

Here the quality and tidiness issues of the data were fixed, same as assessing data, cleaning it can be done programmatically. First a copy of each data frame was created, then each issue was cleaned step by step; define, code, and test. (. e.g:

Define

```
Twitte-archive: fill missing data `expanded_urls` by filling the `tweet_id` to the end of the li
```

Code

```
for i in range(twitter_archive_clean.shape[0]):
    if twitter_archive_clean.expanded_urls.isnull()[i]:
        twitter_archive_clean.expanded_urls[i] = 'https://twitter.com/dog_rates/status/' + str(t
```

Test

```
twitter_archive_clean.info()
```

After the data was cleaned, it was saved as .csv files.
The next step was analyzing and visualizing the data.

In [ ]: