Daniel Stenberg
Udacity Assignment: Data Wrangling
2022-08-02

Report: Data Wrangling of WeRateDogs Twitter Entries

Introduction

To enable analysis of WeRateDogs (@dog_rate) twitter entries, or “tweets”, different data sources have been identified. To facilitate this process, the obtained data must be gathered, assessed, cleaned and stored according to a data wrangling process.

This document reports on these activities which have been largely performed using the Python programming language in a Jupyter notebook environment. Tools and libraries used include:

- pandas – data exploration and manipulation tool
- numpy - Python numeric calculation library
- matplotlib – Plotting tool
- BeatifulSoup – HTML exploration tool
- requests – HTTP(S) request tool
- tweepy – Twitter API explorer tool

The final result is a master pandas Dataframe and corresponding CSV file.

Gathering

The data was gathered from three sources:

- A CSV file of archived tweets manually downloaded and reuploaded in the Jupyter environment
- A line-by-line JSON file of dog image prediction downloaded programmatically
- Additional tweet information obtained from the Twitter web-API

The gathering was completed successfully in all cases and the data loaded into local files and Dataframes.

Assessing

Initial assessment and familiarization of the data was performed in Microsoft Excel. Most assessment was, however, performed using Pandas functions including:

- head()
- tail()
- describe()
- info()
- query() – filtering out different interesting aspects
- value_counts()
- duplicated()
- unique()

The identified issues subject to cleaning are summarized below:

Quality issues

1. Data type issues including IDs as float type and dates not as datetime objects
2. Issues with naming conventions of predicted image objects
3. Source information convoluted with surrounding HTML tags

4. Retweets are not wanted along with retweet information columns
5. Tweets without images are unwanted
6. Strings "None" should be converted to proper Python None
7. The rating values given as numerator and denominator produce instances of highly deviating or invalid values
8. Some values in `rating_numerator` do not match the values in the actual tweet string due to decimal point numerator in the text
9. It would become easier to understand and analyze the rating if the numerator and denominator ratio was in a column of its own
10. In the extended tweet information, a lot of the information is clearly not of interest. Some of these have structural issues, but these should anyway be removed.
11. "In reply to" information is not needed
12. The tweet identifier columns are inconsistently labelled across data sources

Tidiness issues

1. All three source dataframes contain information pertaining to specific tweets. The information would probably be easier to work with if this information was merged into a single master dataframe.
2. The dog types/stages `dogger`, `floofer`, `pupper`, `puppo` should be consolidated to one column whilst handling possible multiple dog types

Cleaning

Cleaning was performed to address the identified issues. Pandas was used extensively with utilization of functions like:

- rename()
- replace()
- drop()
- merge()
- to_datetime()
- astype()
- at()

An example issue was incorrect rating value extracted as an integer from a tweet string. These were quite few and were handled case by case.

Another issue was the spread of dog types/stages over several columns which was consolidated into one column for tidiness with the source columns removed. The rare tweets with multiple dog types were concatenated to strings such as “doggo, floofer”. This tidiness issue was challenging where the pandas melt() function was tried initially, but a more straight-forward approach was suggested as part of the project review.

All cleaning steps were followed by prudent verification steps to assure the cleaning had the desired effect.