wrangle_report

May 10, 2020

1. Introduction

In this paper I will describe my wrangling effort I made by analyzing tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The goal of the project was to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

The main steps of the project were:

- Data Gathering
- Data Assessing
- Data Cleaning
- Data Analyzing and Visualizing

2. Data Gathering

The data for this project consist of three pieces:

1. The WeRateDogs Twitter archive was downloaded manually as twitter_archive_enhanced.csv and stored as DataFrame in jupyter notebook using Pandas library
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network.
   This file (image_predictions.tsv) was hosted on Udacity's servers and was be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Using the tweet IDs in the WeRateDogs Twitter archive, the Twitter API was queried for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file.

Data gathering Summary: as result of the data gathering process, we have obtained three different Data Frames, which were imported into the programming environment (Jupyter Notebook)

3. Data Assessing

After gathering each of the above pieces of data, I have assessed them visually and programmatically for quality and tidiness issues. We could detect and document the following quality and tidiness issues.

3.1 Quality issues (Completeness, Validity, Accuracy, Consistency)

- `user_id` is numeric

- `retweeted_status_id` contains 181 retweets

-`in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp` contain further information abour retweets

- Missing values in `expanded_urls`

- Column `source` is not properly formatted (HTML tags)

- Column `timestamp` has type object (string)

- Outliers in `rating_numerator` and `rating_denominator` columns

- Column `name` has not only dogs' names but also words like `None`, `all`, `not`, `one` etc.

- Dogs' breeds are written with underscores and can be improved

3.2 Tydiness issues (structural issues)

- `doggo`, `floofer`, `pupper`, `puppo` describe the same measurement unit `dog stage`

- more than one stage is filled for the particular dog (found by assessing the df_image DataFrame further)

- Columns' names don't clear describe the content

- Results of neural net predictions are spread in a lot of columns (can be combined in one column)

4. Data Cleaning and Analyzing

Data cleaning was the third step in data wrangling process. During this step the quality and tidiness issues, defined in 3. were fixed step by step. I used the programmatic type of cleaning. The cleaning process for each issue was divided in three steps: define code and test.

After fixing the issues for each of three Data Frames I have made a copy of a final clean one and merged all three Data Frames in one master Data Frame. This master Data Frame was eventually stored in to a twitter_archive_master.csv file.

As the last step of the project, I have analyzed the data, produces visualizations and insights as well