Wrangling and Analyzing

February 03, 2019

1 Gathering the Data

We gathered data from 3 sources:

● Enhanced Twitter Archive, which contains basic tweet data for all 5000+ of their tweets → twitter_archive_enhanced.csv

● Additional Data via the Twitter API, which contains retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API → tweet_json.txt

● Image Predictions File, which is a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images). → image_predictions.tsv

2 Assessing the Data

Pandas' .info() method on the merged data showed that timestamp column needed to be a datetime object instead of a string.

There were several empty values in *in_reply_to_status*, *in_reply_to_user_id*, *retweeted_status_id*, *retweeted_status_user_id*, *retweeted_status_timestamp*.

The *name* column had several entries which do not look like a name.

Some of the ratings did not look right. The expected value for numerator and denominator was around 10, but there were many values above 100 also.

The number of rows in archive data and images data did not match.

In several columns, the null values are treated as non-null values. Some entries contain "Nan" as string.

The *Unnamed: 0* column was to be removed.

The columns for dog breed predictions could be condensed into a single column.

The dog stages values were named as columns instead of one column containing stages values.

3 Clean

The data have been cleaned thanks to the programmatic method. With this method, we need to define (definition or instruction list) the cleaning task. Then, we code the issue to get it cleaned (drop, extract, islower, loc, etc., methods). At the end, we test the dataset, visually or with code, to assure that the cleaning operations work correctly.

Conclusion:

Through the data wrangling and analysis, we used many libraries such as pandas, NumPy, requests, tweepy, and json, which allow us to gather, assess, and clean the data. Finally, we put the following documents together:

● wrangle_act.ipynb: code for gathering, assessing, cleaning, analyzing, and visualizing data

● wrangle_report.pdf: documentation for data wrangling steps: gather, assess, and clean

● act_report.pdf: documentation of analysis and insights into final data

● twitter_archive_enhanced.csv: file as given

● image_predictions.tsv: file downloaded programmatically

● tweet_json.txt: file constructed via API

● twitter_archive_master.csv: combined and cleaned data

2