Data Wrangling Report

Feb 18, 2018

In this report, We will describe my wrangling process and effort made in this WeRateDogs twitter data analysis project in three detailed part:

- Gathering Data
- Assessing Data
- Cleaning Data

Gathering Data

Gathering data for this project composed from three pieces of data:

1. The WeRateDogs Twitter archive. We manually downloaded this .csv file using pandas: twitter_archive_enhanced.csv
2. The tweet image predictions file. Every image in the WeRateDogs Twitter archive was run through a neural network for classification of breeds of dogs. The results alongside each tweet ID, image URL and how confident of each prediction was stored in this table and we can download it programmatically using python Requests library at the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Additional data via the Twitter API
   We used the tweet IDs in the WeRateDogs Twitter archive to query the Twitter API for each twitter's JSON data using Python tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt with each tweet stored in a line.

Gathering data is always the first step in the data wrangling process. In this project, we applied three different ways to gather the needed data: reading from an existing CSV file using pandas, using requests to download a file from the internet and querying an API to get JSON object.

Assessing Data

After gathering each of the above pieces of data, accessing them programmatically for quality and tidiness issues was our next step. We could detect and document the following quality issues and tidiness issues.

Quality

We assessed the datasets separately in four following standards: Completeness, Validity, Accuracy, and Consistency.
Following are quality issues we found:

The WeRateDogs Twitter archive:

- tweet_id is an integer instead of a string
- Remove records whose denominator is not 10:
    After checking these tweets and pics, we can find these tweets are rated either for the multiple quantity or specific numbers. These special cases would introduce uncertainties to further analysis.
- Rating denominator and numerators are not correctly extracted:
    1) Extract other fraction rather than the rating:
       # This is an Albanian 3 1/2 legged Episcopalian. Loves well-polished hardwood flooring. Penis on the collar. 9/10.
       The rating numerator is 1 and denominator is 2 in the dataset
       # account started on 11/15/15.
       Mistakenly extract 15 as rating denominator
    2) Miss decimal in the rating numerator
       # I've been told there's a slight possibility he's checking his mirror. We'll bump to 9.5/10. Still a menace
       The numerator is 5 in the dataset rather than 9.5
- Missing data in 'expanded_urls' column (can not clean now)
- 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id' are floats insteand of string type.
- 'timestamp' is a string instead of datetime type
- Erroneous names: 'a', 'the', 'an', 'one', 'very', etc, replace these names with None value
- Remove unwanted columns

The Image Predictions file:

- tweet_id is a int instead of a string
- Capitalize the first letter of p1,p2,p3 columns to make it consistent
- Remove underscores '_' between words in p1,p2,p3 columns

Additional Data via API:

- Convert tweet_id column from an int to a string object
- Exclude highly incomplete columns(0 or 1 non-null value): contributors, coordinates, geo, place, favorited
- Remove unnecessary columns that are not needed for later analysis
- Rename id to tweet_id to keep consistency with other two tables

Tidiness

Untidy data is data with structural issues. Following are tidiness issues we observed:
- Original tweets and Retweets are mixed in one table. Remove all retweet records and keep only original records (Not each type of observational unit forms a table)
- Combine all dog stage columns into a single column (Not each variable forms a column)
- Separate timestamp columns into two variables 'Date' and 'Time' (Not each variable forms a column)

Cleaning data

Cleaning data is the final step in data wrangling and it was where we fixed the quality and tidiness issues we found above in the assessing data step. Our cleaning logic was Define, Code and Test one by one and iteratively. After finishing the data cleaning, we revisited the merged dataset and iterated assessing and cleaning before we moved on to our analysis.

Summary

Data wrangling is an important part of any data analysis since real-world data rarely come clean. In this project, we used Python and its libraries to gather, assess and clean WeRateDogs Twitter data and developed some insights and visualization based off the cleaned dataset. From this courses and the practical project, I learned key concepts and characteristics of quality and tidiness of datasets. This project equips me with all the skills I need to query data via API, use pandas to clean dataframes and conduct basic analyses.