wrangle_report

February 17, 2019

1 Data Wrangling Report

1.1 Gathering the Data

I started the gathering process by downloading the 'twitter-archive-enhanced.csv' manually, and then programatically downloading the 'image_predictions.tsv' file off of Udacity's server using Python's requests library.

Then, I had to use the Twitter API with Tweepy in order to acquire the rewteet and favorite count data of all the tweets in the twitter archive. To do this, I looped through every tweet ID in the 'twitter-archive-enhanced.csv' file, and used the ID to query Twitter's API, saving each JSON response into a new line of a file called 'tweet_json.txt'.

After the query was completed (which took around 30 minutes, accounting for the timeouts to wait for the request limits to refresh), I accessed the favorite count, tweet ID, and retweet count for every line in the 'tweet_json.txt' file, and saved the data into a dictionary, and appended that dictinoary into an empty list. Finally, I used the list of dictionaries to create a pandas DataFrame and saved the DataFrame into a csv file called 'retweet_favorite.csv'.

1.2 Assessing and Cleaning the Data

After I had all of the required datasets, I read them into separate pandas Dataframes and assessed both their quality and tidiness, through programatic and visual means.

Through visual assessments of samples from each dataset, I noticed that there was some missing or incorrect data, particularly in the columns which containing name and rating information, as they were generated programatically. Through the use of regex, I was largely able to clean this up.

Through checking the .info() method, I also noticed problems with some of the data formats, so I converted them using the .astype() method.

I also noticed that the dog stage was represented in four separate columns, when it should be represented by one (according to the tidiness rules), and thus I deleted those columns, and created a new column by checking for the dog stage using a regular expression.

Lastly, I combined all the DataFrames into a single dataframe, as they all concerned the same observational unit, and deleted the tweets which were deleted from the Twitter account.

1.3 Storing the Data

After the cleaning was complete, I saved the new DataFrame into a csv file called 'twitter_archive_master.csv'.