Wrangling report for Weratedog project

Sha Liu

April 2019

This project was part of the data wrangling section of the Udacity Data Analyst Nanodegree program and is primarily focused on wrangling data from the WeRateDogs Twitter account using Python. WeRateDogs provided their Twitter archive of basic tweet data (tweet ID, timestamp, text, etc.) for use with this project. The "enhanced" CSV file provided by Udacity (twitter_archive_enhanced.csv) also contains columns which were extracted programmatically: the rating numerator, rating denominator, dog's name, and dog stages (doggo, floofer, pupper, and puppo). These columns need to be assessed and cleaned as the extraction process wasn't perfect.

However, the provided Twitter archive lacked some useful information: retweet count and favorite count. So first I need to gather the data from Twitter using Twitter API. I found this part is the most challenge part of this project as this is my first time use a Tweenpy to automatically grabbing data. First of all, I need to understand the documentation of Tweenpy. After reading the Tweenpy official documentation and following the instruction, I successfully used the tweet IDs to query the Twitter API 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. Then, I read the text file line by line into a pandas DataFrame only including the desired variables; retweet count and favorite count as required by the project.

Once I had all the data needed, I found cleaning the data is relatively the easy part compared to the process of gathering data. I first used the info and describe function to have an overview of the three data tables and immediately noted spotted some data issue. The number of twitter_id in twitter_archive table is not the same in image_predictions table and tweet_count table. Then based on columns title provided by the info function, I found more data issues. Lots of the columns contain unuseful information, such as “expended url”, “ text” and so on. By looking at the data type of the three tables, I found out that some of the datatypes are incorrect for the context stored in the columns, for example, the datatype for column “timestamp” is an object. To ensure the timestamp information to be useful in my future analysis, the data type of this column needs to be converted to DateTime.

Lastly, I looked at each data table to see if I can observe some data issues. I found that some of the data recorded in “name” columns are not real dog names. I think this data should be replaced by “none” instead.