James Robinson

We-Rate-Dogs Data Wrangle Report

There were many things I noticed wrong with the data from the 'twitter-archive-enhanced.csv' file. This is where I did most of the data wrangling. As for the minor changes, I needed to change category of the tweet_id from int to categorical, and timestamp from a string to a datetime object; join the ratings and clean up the dataframe by removing the denominator and the numerator columns; change the source from html to the nested string within, removing the retweeted_status_id and the retweeted_status_user_id columns; making empty data NoneType; and a couple other small minor changes.

As for the more difficult changes: grabbing the hashtags, url, and rating from the text with regular expressions; comparing the names provided by creating a regular expression to find likely names through the text; removing unlikely names programmatically or visually.

For the other two files, I had to change the tweet_id again to categorical so I could merge the data without running into a typeerror.

In the 'image-predictions.tsv' file, I ended up removing all of the data where p1 didn't come up with a breed of dog. I made this decision because I opened a lot of the jpgs where p1_dog was False, and most pictures weren't even of dogs. Just to reassure myself on that decision, I checked p2_conf of the p1_dog False data, and almost all of them were less than 1%, so I didn't want that bad data to potentially get in the way of my visualization accuracy.

Although it was a time consuming and learning process to understand how to correctly use the tweepy api, the object that I eventually created with the authorization codes that I applied for ended up having accurate data. The only change I made to it was changing the type of the tweet_id, as well as making the tweet_id the first column for uniformity.