Untitled1

September 20, 2018

Wrangling this data was an initially quite difficult task but ultimately rewarding. Gathering the data given in the csv file was quite easy as we've done so many times before in this course. Gathering data from a file programmatically was a bit more difficult. I had to look up some documentation on stackoverflow in order to figure out how to do it correctly. Even after that I still needed to use some trial and error until I could get it to work properly. The final gathering requirement was by far the most difficult. I had an idea of how I wanted to extract the data via the twitter api and place it in a json file, but executing it was far more difficult. It probably took me over 4 different attempts to get this to work properly. Placing an exception in my for loop helped though as I would know immediately if the loop was gathering the tweet Id's correctly or not. I worked with my fellow Udacity student Thomas in order to get this done. Having another person to bounce ideas off of helped a great deal.

The next part required us to asses our data. This was quite easy, I simply used .head(), .info(), .describe(), to get a good look at my data and these functions yielded data messiness. I also tried to see where there was null data in fields in which it was important to have data.

Finally came actually going through the cleaning process for the data. I really enjoyed this part and felt I learned a great deal. The tidiness issues were in some ways a bit easier to fix, I think this is because they were more obvious. I first consolidated my data into one dataframe, I felt it would be better to do this first before cleaning them individually. Next I changed the dog type columns from columns to a single column with the previous columns as the possible data attributes. For quality, I did various things to clean the data set. I enlarged the text from the dataframe to make it easier to read; got rid of underscores and capitalized the possible dog breads; changed dog names that were 'None' to 'NaN'; made the source variable more clear; changed the datatypes for multiple columns; renamed the columns in my json_dataframe so it could be joined with my main dataframe; and removed retweeted tweets; dropped columns with expanded urls that were empty.

Overall, this project was quite challenging, but I learned a great deal. I also did some research and found that these methods are often employed in real life scenarios quite often, so I appreciate the use of non-platonic projects.