Byungjin (BJ) Kim
9/2/2019

Wrangling Act Report

Initially the wrangling started by getting the csv and tsv files that I was using. First, I read in the csv file to hold a main data frame of the original data, and then I proceeded to grab the tsv file from the url given to us. After grabbing the tsv file, I put the information inside of the tsv file into another data frame that held the original prediction data. After having these two data frames, then I moved on to grabbing basic json data from the main data frame, which held the data from the twitter_archive_enhanced.csv file. After creating a third data frame to hold the basic json data, I proceeded to add that into a txt file. The main wrangling then began. I ran through the main data frame to see what kind of major cleaning issues that occurred. A couple major ones caught my eye right off the bat. There were numerous numerators for ratings that were extremely high as well as numerous denominators that weren't at 10. I eventually decided to clean this by just removing all of the rows that had denominators that weren't 10 and all the rows that had numerators that were above 20. This is a slightly risky cleaning method, as I could have lost rows of data that were fine for use, but this would have been such a small number that I decided just getting rid of those rows wouldn't impact the investigations much. However, a precaution I took before getting rid of the rows that had numerators over a value of 20 was that I went through the text of the tweet to figure out the exact rating that had been given, in case the rating in the data frame was a typo or a false rating. This prevented against getting rid of multiple rows that otherwise I would not have been able to use. A couple other quality issues that arose were that names sometimes didn't make sense or were not present at all, so I removed those rows as well. For the names that didn't make sense, I used help from a github of the project in the past, which had a lot of the names that ended up just being words (that weren't names), so changing these names to "None" was done by using the code from that github. This github is sourced as a comment under the block of code used. These were a few examples of the quality issues that I found while assessing the main data frame. As for tidiness issues, the two main tidiness issues I found were that the main data frame had columns for each stage for the dog, like "doggo", "floofer" etc, which I decided to just put as one column of "stage_of_dog". The other tidiness issue I found was that the predictions data frame had the predictions to what breed type of dog the dog in the post was, but this was a separate data frame from the main data frame, so I decided to merge the two data frames. Ultimately, this is a quick summary of the wrangling actions taken during this project.