Wrangle_Report

Date: December 18, 2017

Data wrangling from the start will always be challenging and I have learned a great deal about gathering, assigning, and cleaning the data in this project. This paper will describe all the steps that were taken for the wrangling process and completion of WeRateDog project.

The first phase of the wrangling process is gathering the data. Our project was composed with three pieces; WeRateDogs Twitter archive, tweet image predictions, and tweepy API json file. The WeRateDogs twitter archive was manually downloaded from this link: twitter_archive_enhanced.csv. This archive contains all @WeRateDogs tweets from August of 2017. The tweet image prediction, identified what breed of dog is present in each tweet according to a neural network. This file was hosted on the Udacity's servers and we downloaded it grammatically using python Requests library on the following link:. The third piece of this process is the Tweedy API json file, which contains each tweet's retweeted count, favorite count and other additional that the team at Udacity found interesting. By using the tweet ID's in WeRateDogs twitter archive, a query was created to retrive twitter API for each tweet's in JSON formatted data using Python's Tweepy library and then stored each tweet's entire set of a JSON data file called Tweet_json file. This part was by far the most challenging, as far as spending a extand amount of time searching and learning how to parse thru JSON file to be queried out to readable format.

The second phase of the wrangling process is assessing the data. Once all the data had been gathered, we assessed them visual and programmatically for quality and tidiness assurance. For quality assurance, our data must guarantee, "accuracy, completeness, consistency, validity" in which in the archived dataset, there were issus with the timestamp columns that needed to be converted from objective to datetime. The name of the dogs in name columns contained invalid name such as: as, a, the, not etc. And the list went on. The images dataset had several duplicates that were dropped. Lastly, the Tweet_JSON dataset also had duplicates and cleared unwanted dataset. For the tidiness assurance, our data must be tidy and completely structured. Therefore, our tables were merged and unwanted columns were dropped, and made sure that each type of observational unit formed a table.

The third phase of the wrangling process is where we fixed the quality and tidiness issue that were identified in the second phase, known as cleaning the data. In this process, we defined, coded, and tested occurrences by coping in files and merging to a single dataset in the end. In this phase, found myself iterating and revisiting the second phase to assure all issues were resolved.

Data wrangling is a core skill set that an individual must obtain in order to wrangle and familiarize themselves with the data. All the world's data are sure not normalized and cleaned, so being able to tackle this issue hands on it's a plus after practicing and struggling thru the process. Believe it or not when you get `dirty` in the wrangling process, your insights and visualization becomes more valuable (ie, dogs gender based on text partitioning) then it was before. **Always, Get Wrangle On It**. (Inspired the song- Earth, Wind & Fire – Boogie Wonderland)