Wrangle Report

Parth Thakur

11th September, 2018

The data wrangling project began relatively easy, but as I soon realised, there were many challenges ahead. Getting the twitter API to work, and querying it to get the data was easy. The hardest part was understanding the received data, and determining what was important. The most challenging task in this part was to figure out the structure of the JSON data. Then there was the task of reading it efficiently and storing it in a text file. Once the tweet_info.txt file was created and JSON structure understood, reading it and adding extracting the missing information went smoothly.

After getting the missing data, the real wrangling efforts began. Looking at the data in a spreadsheet program made some issues plenty clear. The dog_stage was split in four columns, and duplicated expanded_url to name some. As I began to clean the obvious problems, some deeper issues came into perspective. These issues would require knowledge of regular expressions. I had to press pause on the project to learn about regex. Once I developed necessary skill in regular expressions, most of the tidiness issues became easy.

Above all, the most difficult challenges to overcome was whenever pandas would give an error saying “Truth value of dataset is ambiguous.” I ran into this error too many times to count, and each time I considered stopping. Though that would mean all my efforts until now would ne for naught. The solutions to this problem were similar but every case required using a new perspective.

Further, the most important lesson I learned through this project, is that no matter how much effort I put into it, real world data will always have issues. Whenever I rectified one issue, another would pop out. I was beginning to think that I would never be able to complete this project. In the end I asked considered the impact solving an issue would have on my analysis. After that point, prioritising issues became easy.