**Term Project Summary**

For my final project, I chose to pull job information from indeed.com. I thought it would be good to start looking into the data science job market prior to my expected graduation in August. I was planning to set-up reusable code that could be used to extract data from indeed on a daily basis, as I thought it was important to keep up to date on the data science job market. I planned to use fields from both the main indeed site, as well as the link to the company page, where the company reviews were located.

I began by analyzing the html of the indeed website on my Firefox browser. I was most comfortable with the xpath format so mainly used an xpath selector object to extract the fields I needed. I took some care to extract as much as possible in the formats I needed so that later clean-up would hopefully be minimal. The results returned from the xpath selector were in the form of lists containing the data for each variable. After iterating through each page of indeed, I took the separate data lists and zipped them together with the header data. I then converted the zipped list into a pandas data frame. Once in the form of a data frame, I performed clean-up on the data to remove line breaks and concatenate the substrings returned for the web links.

I was having some trouble integrating the different Python data structures (i.e. string, list, dictionary) together with the pandas data frames. I was comfortable with each data structure on its own, but I found that using them together was a challenge. For example, after zipping the separate lists together and transforming them into a pandas data frame, I was left with a list of lists inside of a data frame. I had some difficulty recalling how to reference the individual components within the data frame since they were several layers down. I decided to use for loops to extract the fields from each data frame and then transform them into the necessary format.