Hey Professor,

When comes to this week’s assignments and discussion board, I have a new found respect and understanding of what goes into not only cleaning unstructured data but the various methodologies and depth one can employ in doing so at the time.

With the first half of the coding exercise, I had some trouble in translating the examples from the book on removing stop words and stemming the words into the large dataset that we working with this week. The trouble for me was utilizing all of the different packages from the NLTK library and paying attention to the datasets to actually see the changes present. At the same time, with the large dataset, I have been used to working with smaller datasets so the program ran a little faster. However, I am now understanding more how time can now play a role in how you clean up the data. With every time I running my script on the dataset, it was getting a point where I simply waited for it to be done.

With this week, I had written about the various options people have when approaching unstructured data and how they can organize and clean said information. It seemed with the unstructured data that it is where most data is being originated from and that we do not have a cache-all to handle all of it. Along with not having a system that can curate all of the data no matter the origin, I do not think you can get your data to be 100% perfect in the end. It is a battle that you must fight, but will never truly completely conquer in the end. Like how it is mentioned in various articles and posts, we want to ensure that it is in the best position for to be used in any model or else it will cause consequence towards the future.

Looking back on the quiz for this week, I was able to understand the mistakes that I had one the first attempt at the quiz. However, the next attempt, I did not understand what the correct answer was for the first question since it seemed that each response was incorrect when I submitted for the next attempt.