Project 02

Dipta Roy

Question 1)

To start off this NLP project, the first goal is to get a sense of our data. To do this I needed to figure out the most common words to effectively get rid of any stop words that might not be helping us in the data. Looking at the most common words in the whole document, “the, and, a, I, is, to, of, was, this, it, in”, were in the top 10. These words we can get rid of as they won’t really add value to our modelling later. Now we have to build a stop words list, there is one in NLTK library which they caution not to use. Further investing why this might be the case, I looked at what is actually in the stopwords “English” library in the NLTK library.