Clare Schwarzenberg

i) I used the Bag of Words technique to model my text. Bag of words allowed me to extract information about the occurrence of the words in the text file without having to deal with the grammar and order of the words of the sentences. For the features I also extracted the length of each of the sentences. I also attempted to calculate the Euclidean distance but it was difficult to use other features besides the Bag of Words representation when training my model. In the end I just used Bag of Words.

ii) For data preprocessing, I took a list of contractions from the internet and converted the contractions in each sentence to the full phrase. This ensured that contractions would be treated the same as the full phrase because they have the same meaning. Then, I cleaned the data by taking out all of the stop words from the sentences and any punctuation. After that I tokenized the data so that the data was no longer in the form of a sentence but was just all of the words that were in the sentence besides the stop words. I then used stemming and lemmatization to attempt to simplify the words to a common base form. As mentioned above I used Bag of Words and I also attempted TF-IDF but I found that the Bag of Words representation produced a slightly better accuracy.

iii) I tried using both logistic regression and Support Vector Machine. I found that Support Vector Machine produced a higher accuracy on the dev set compared to logistic regression. I also found that linear SVM worked the best. I used the nltk library for preprocessing purposes like stemming and lemmatizing the data. I also used sklearn for logistic regression, SVM, implementing bag of words, and calculating the accuracy of each model.

iv) The biggest lesson I learned from this project is that text data is much more difficult to deal with compared to numerical data. This was the first time I did a big project on text data and it required much more in depth preprocessing compared to numerical data. I also found it difficult to extract useful features from the data since it was textual data. It was frustrating to find how to calculate features like Euclidean distance and Cosine Similarity for text data.