Praneel Trivedy
Professor Wenpeng Yin
CIS 4526
20 October 2022

System Description

Prior to performing on the training data, there needed to be preprocessing steps for this specific ML problem. This included putting all words to lowercase, removing punctuations, and removing meaningless articles. These steps helped the algorithm clear out some of the noise which would have been observed in the data otherwise. I could not use simple preprocessing such as splitting by "." because it would splice the data in undesirable ways and ruin sentence meaning.

After this step, sentences needed to be converted to tokens so the algorithm can perform on the vectors. Tokens were used to split up sentences into their component parts, words. I split the data into two feature vectors of containing each sentence. After this, the vectors were normalized.

Libraries used included Sklearn and NLTK. I implemented logistic regression as a binary classifier to determine if the sentences were similar. In my research this was described as a great starter algorithm to classify sentence paraphrases. This was also easier to implement than other algorithms.

From this project I learned some valuable skills in time management and thought experimentation. The project was assigned a month ago so I found it difficult to continuously work on the assignment while managing other coursework and a job. From a machine learning perspective, I found it helpful to understand the issue before breaking into code. There are key ideas for specific problems in NLP that must be addressed based on the task and available data. Furthermore, this project was a test to my coding skills which certainly need improvement. I hope to move toward the final project with a stronger command over these two ideas.