Kashyap Patel
Homework 1 Report

The purpose of this homework assignment is to classify movie reviews as positive or negative without using machine learning. I chose to use the NLTK package and their opinion lexicon based on the Hu Liu Opinion Dataset in order to determine the sentiment of the review in question. I used the opinion lexicon and compared lemmatization of different aspects of the input to determine its effectiveness in this example. I either used no lemmatization, lemmatized the opinion lexicon, lemmatized the input sentence, or both the lexicon and sentences. The overall reviews were scored on if they had more positive or more negative words based on the lexicon dataset. The results can be found in the results.md file.

In the case of this data it is interesting to note that lemmatizing the inputs seem to have a negative effect on the evaluation of the sentiment of the movie review. This can be noted in the declining accuracy and f1-score as lemmatization is added to the evaluating script. In addition, adding lemmatization seemed to increase our specificity, the likelihood of correctly identifying a negative review, by a full 11.2%. However, our highest recall, or likelihood to correctly identify a positive review was highest when no lemmatization occurred. Furthermore, lemmatization further increased the time and memory required by the script to complete its evaluation, however, this current script's lemmatization could be further optimized by only performing lemmatization on the lexicon once rather than at each iteration.

Therefore, it would seem that when using this movie review dataset and no machine learning it is better to retain the word form rather than normalizing it. Other interesting methods may be to check for word case (excessive uppercase characters may be weighted higher), and for specific actors/actresses that may be indicators of a better/worse audience rating. Our highest f1-score belonged to the no lemmatization script with a score of 0.61.