Yoofi Brown-Pobee Lab 4 Report

Choice of Libraries
I used the following libraries in my worl

- Sci-kit Learning
- NLTK

NLTK was used for normalization as it has a library of common words like ‘the’, ‘a’ and ‘an’ that are removed when normalizing.
Sci-kit was the more important library used. It contained the Vectorizer Object (CountVectorizer) which was necessary for converting the reviews into a matrix of word counts. This was necessary because the classifiers (Multinomial Naive Bayes and Logistic Regression) need vector features in order to perform the classification task. It also contained Multinomial and Logistic Regression objects that were used to build models and fit to the training set.

Results of The Classification with amazon test file

Classifier Type	Precision	Recall	F1-score
Naive Bayes with Normalization	81%	78%	78%
Naive Bayes without Normalization	79%	76%	76%
Logistic Regression with Normalization	81%	80%	80%
Logistic Regression without Normalization	79%	78%	78%

From the above table, normalization had higher precision, recall and F1 than no normalization in both classifiers. Logistic Regression with normalization had the highest average of the three measures among the four classifications. I think normalization has a higher score than no normalization because it tries to reduce how much the morphology of sentences impacts the classification by reducing all words to lower case, removing punctuations and common stopwords. The Logistic Regression classifier was better than the Naive Bayes classifier due to what I believe is the Naive Bayes classifier’s assumption that features are conditionally independent even though this is not true. The Logistic Regression overcomes this bias and attempts to make a prediction directly.