Maame Yaa Osei

Lab 4 Report

Natural Language Processing

I developed two Naïve Bayes Classifiers and two Logistic Regression classifiers. For each type of classifier, I developed normalized and un-normalized versions.

For both the Logistic Regression classifier and the Naïve Bayes classifier, I used the pandas library to convert the data from the .txt files into a format that could easily be work with (ie a Numpy array), Sci-Kit Learn’s CountVectorizer module to turn the sentences into a vector of counts for the un-normalized classifiers to enable the classifier perform operations on the documents. For the normalized classifiers, I used the TfidfVectorizer because it yielded a vector of normalized data by removing stop words (which were imported from nltk), lowercasing and tokenizing the documents. I also used Python’s pickle library to serialize the different models to prevent the classifiers from training every time they were run. I used the Sci-Kit Learn library because it accepted data easily and returned it for use in formats that were easy to process.

Both the normalized versions of Naïve Bayes and Logistic Regression resulted in accuracy scores of approximately 82 percent. Though this may be credible enough, it could be improved by training on more data. However the normalized versions of both classifiers resulted in accuracy scores of about 79 percent. This may be because normalization eliminates certain words from the document which may be strongly associated with the relaying of a particular sentiment.