Program 4
LING 5832
Yuan Chai

Program 4 Report

Classifer: Naïve bayes

What features I tried and why:

1. The existence of most frequent 2000 words: provided by the textbook. **Reason:** The most frequent words might identify the sentiment of the text. The 10-fold validation accuracy is 0.6845.
2. The existence of positive/negative lexicons provided by *The General Inquirer.* The features are in the format of *{contains(good) : True/False}*. The lexicon is stored in "posneg_inquirer.txt". **Reason:** Positive and negative usually characterize the positivity and negativity of texts. Positive reviews are more likely to contain positive lexicons while negative ones would more likely to contain negative lexicons. The classifier including Feature 1 and 2 yields a frequency is 0.7905.
3. The count of positive and negative words in text. The features are in the format of *{freq(pos):<count of positive category>, freq(neg):<count of negative category>}.* **Reason:** Potts (2011) suggests that negative sentiment tends to have more negations. Thus the count of words in positive and negative category may help to characterize the sentiment of the text. Result: However, this feature decreases the accuracy to 0.7885.
4. Remove the objective lexicons (lexicons without a positive/negative tag) suggested by *The General Inquirer* from the most frequent 2000 words. The objective lexicons are stored in *"neutral.txt"*. **Reason:** objective lexicons may not be very contributive to the sentiment of a text. Extract the irrelevant information might be helpful for the text. **Accuracy:** However, the accuracy declined to 0.71, which suggesting that objective lexicons still have a role in sentiment analysis.
5. Extract preposition and articles from the most frequent 2000 word. The preposition and article list is from Wikipedia and stored in *"prep.txt"*. **Reason:** Prepositions and articles are usually the most frequent word in corpus but barely have any sentiment value. And removing them did increase the accuracy.
6. The same as Feature 2, but the positive/negative lexicon is from Bing Liu Opinion Lexicon. The lexicons are stored in *"positive_bing.txt"* and *"negative_bing.txt"*. **Reason:** I was testing which pos/neg lexicon corpus would yield the best accuracy. Accuracy: Bing Liu Opinion Lexicon turns out to fit the movie_review corpus better than The General Inquirer. The accuracy was improved to 0.8145.
7. Remove the count of positive lexicon feature but preserve the negative count. **Reason:** There is study suggesting that negative count could better discriminate sentiment. **Accuracy:** The accuracy was improved to 0.815.
8. Remove the feature of the existence of most frequent 2000 words. **Reason:** it is possible that the most frequent 2000 words are less informative than positive/negative lexicon. The **accuracy** was improved to 0.8245.

Final decision:

I used **Feature 6:** The existence of positive/negative lexicon provided by Bing Liu Opinion Lexicon; **Feature 7:** The count of words in negative category; and **Feature 5:** Remove preposition and articles from Feature 6.

Reason: this combination yields the best accuracy. Positive/negative lexicon provided

by Bing Liu Opinion Lexicon fits movie_review corpus better than The General Inquirer. Prepositions and articles are not very informative for sentiment characterization and removing them from features improve the classifier accuracy.
Final accuracy: 0.8245.