Programming Assignment Analysis

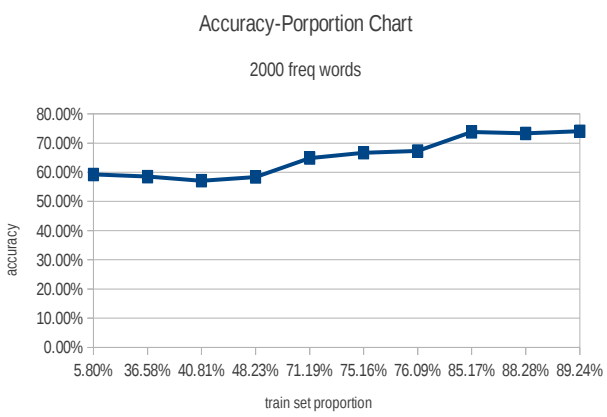
In this Assignment, after setting up the baseline, I mainly focus on how to effectively select the useful feature and then improve the classification accuracy. Generally, I tried two ways to modify the feature and take experiments: Stopword filtering and frequency selection. And my experiments told me that the

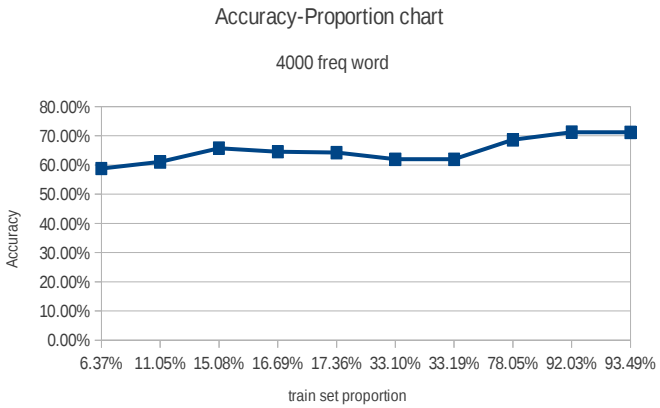
First Attempt

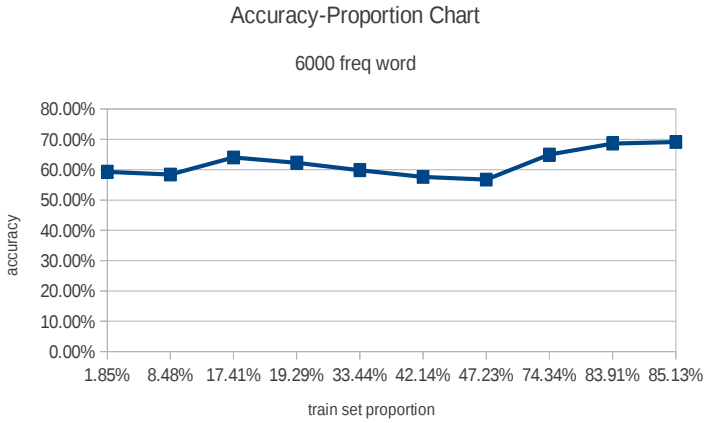
First, I try to build a model trained with all the words served as features, which means I do no feature selection. I tokenized each sentence and recorded almost 30000 entries in the codebook. However, when I ran the program, though it spent a little time on training the model, It took quite a long time to compute the log score and the total testing process is unpractical. So I gave up building the baseline with all-words feature.

Frequency selection

Then I tried to limit the number of features based on certain metric. Quite straightforward, I choose frequency of the word in the whole dataset as my metric to select the feature. By using the nltk.Freqdist(), I am able to get the frequency distribution of all the word in the dataset with occurring times. However, it is a problem that how to choose the number of most frequent feature. In this assignment, I did not take too many tricks to determine the how to choose the right limited number and I just do 3 experiments by separately using 2000, 4000 and 6000 most frequent feature. The result is:



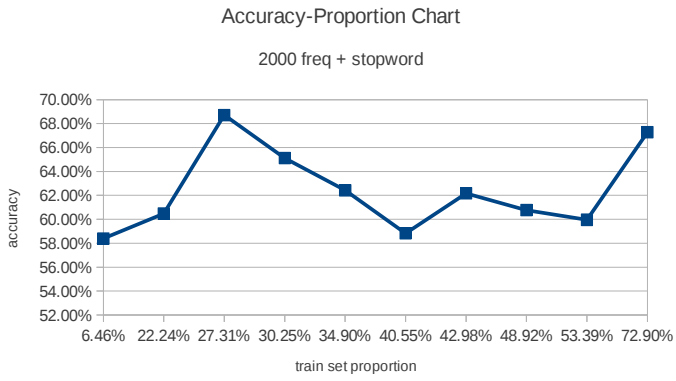


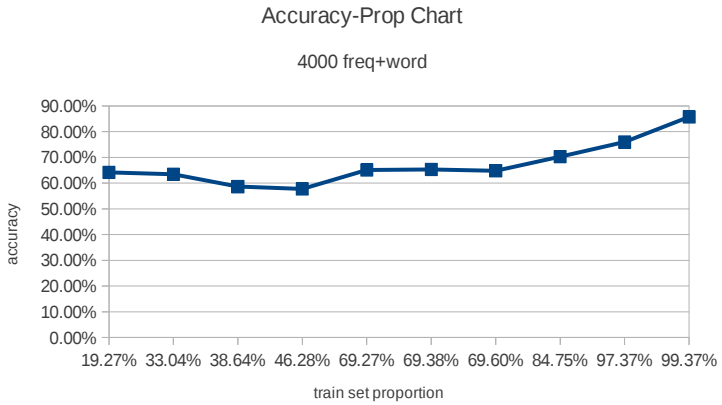


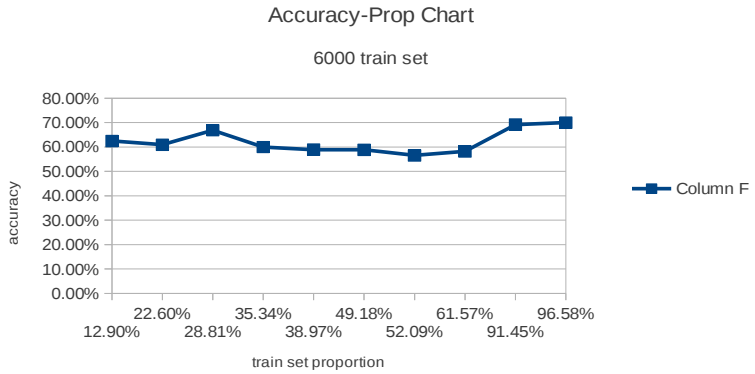
We observe that the classification results is not greatly affected by the number of most frequent word chosen and also the train set proportion have minor effect on the accuracy.

Filtering Stopwords

By using the nltk.corpus.stopwords, we may get the predeifined stopwords set and get rid of the stop words in the feature, since stop words also more frequently occurs in a document. And I also take 3 experiments by adding the filtering on the feature frequency selection process.







By calculating the average accuracy, we can get that by filtering the stop words we can improve the accuracy by 2%.

Conclusion

Though by using the stop words, we get a little improvement, in fact the feature selection method used in this assignment in fact did not have much improvement for the accuracy. And I may seek for other methods and also try another evaluation metric like precision-recall and see what's the result.