550 assignment1 report

I. Problem set up

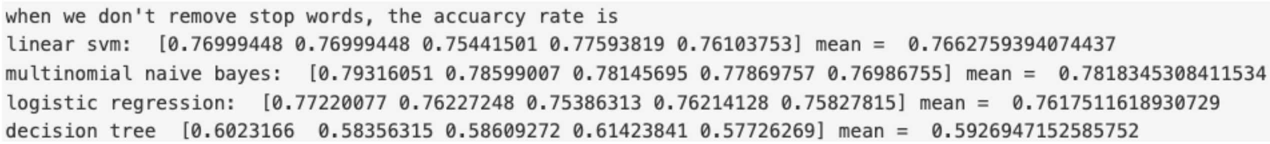
In this experiment, we solve the problem to classify a sentence into either a positive or negative sentiment, with sentences come from a movie review dataset.

II. Experiment procedure

1. Preprocessing and feature extraction: I extracted the feature by TfidfVectorizer, and I tried to use different combination of lemmatization, stemming, remove stop words, and remove rare words with different occurrence, then choose the one give the best results.
2. Cross validation: I did cross validation by for each of the four models by cross_val_score, divide the data set to 5 folds, 1 for test and 4 for training.
3. Selected models: I did the model selection by tuning hyperparameters. I tuned the regularization parameter in linear SVM model, and the smoothing parameter in multinomial NB model. The details are included in "III. Range of parameter setting"
4. The forth model I chose is decision tree, which give an accuracy around 60%, since this is bigger than the random baseline 1/5，so it can do the job of text classification.

III. Range of parameter setting

1. I' ve tried to remove stop words in vectorizer, the result is that all of the four model would have **better** performance when the data **is not removed stop words**, which means use TfidfVectorizer directly. The best result achieved by multinomial naïve bayes, which is 0.781. And we would not remove stop words later, since it doesn' t improve performance.



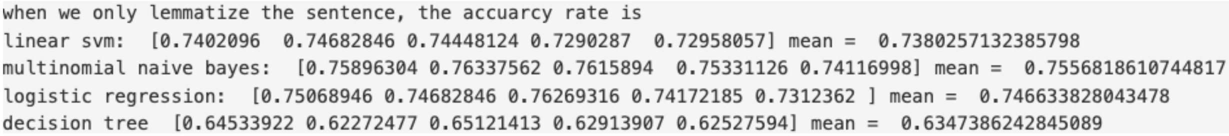
```
when we don't remove stop words, the accuarcy rate is
linear svm:  [0.76999448 0.76999448 0.75441501 0.77593819 0.76103753] mean =  0.7662759394074437
multinomial naive bayes:  [0.79316051 0.78599007 0.78145695 0.77869757 0.76986755] mean =  0.7818345308411534
logistic regression:  [0.77220077 0.76227248 0.75386313 0.76214128 0.75827815] mean =  0.7617511618930729
decision tree  [0.6023166  0.58356315 0.58609272 0.61423841 0.57726269] mean =  0.5926947152585752
```

2. I' ve tried to change the min_df of tfiddvectorizer, to ignore terms that have a document frequency strictly lower than different threshold.

   When we removed words appear **less than 5 times**, the best result is achieved by multinomial NB which is 0.768, with linear svm, logistic regression and decision tree 0.751,0.754, 0.595 respectively.

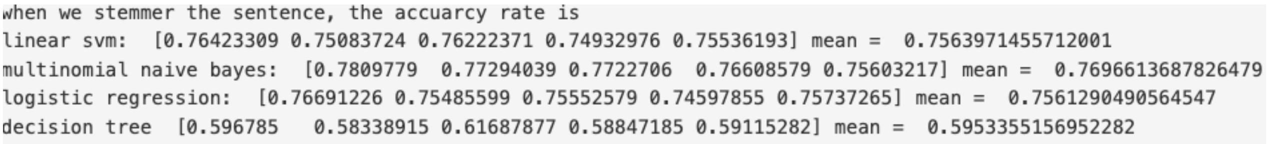
   When we removed words **appear less than 3 times**, the best result is achieved by multinomial NB with 0.775 , with linear svm, logistic regression and decision tree are 0.741,0.725, 0.594 respectively

   When we do not remove any rare words, the best result is achieved by multinomial NB with 0.770, and linear svm, logistic regression and decision tree are 0.759,0.748, 0.582 respectively. From above, we can see that the accuaracy rate would increase as we remove less rare words, and **it finally achieved the best when we do not remove uncommon words**. So we would not remove any rare words later.
3. I' ve tried to lemmatize the input sentence, the **accuracy rate is higher when don' t lemmatized data.**



```
when we only lemmatize the sentence, the accuarcy rate is
linear svm:  [0.7402096  0.74682846 0.74448124 0.7290287  0.72958057] mean =  0.7380257132385798
multinomial naive bayes:  [0.75896304 0.76337562 0.7615894  0.75331126 0.74116998] mean =  0.7556818610744817
logistic regression:  [0.75068946 0.74682846 0.76269316 0.74172185 0.7312362 ] mean =  0.746633828043478
decision tree  [0.64533922 0.62272477 0.65121413 0.62913907 0.62527594] mean =  0.6347386242845089
```

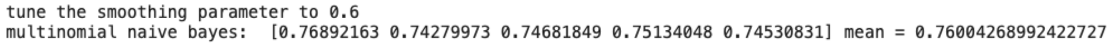
4. I' ve tried to stemmer the input sentence by Porter stemmer, the **accuracy rate is higher when don' t stem data.**



```
when we stemmer the sentence, the accuarcy rate is
linear svm:  [0.76423309 0.75083724 0.76222371 0.74932976 0.75536193] mean =  0.7563971455712001
multinomial naive bayes:  [0.7809779  0.77294039 0.7722706  0.76608579 0.75603217] mean =  0.7696613687826479
logistic regression:  [0.76691226 0.75485599 0.75552579 0.74597855 0.75737265] mean =  0.7561290490564547
decision tree  [0.596785   0.58338915 0.61687877 0.58847185 0.59115282] mean =  0.5953355156952282
```

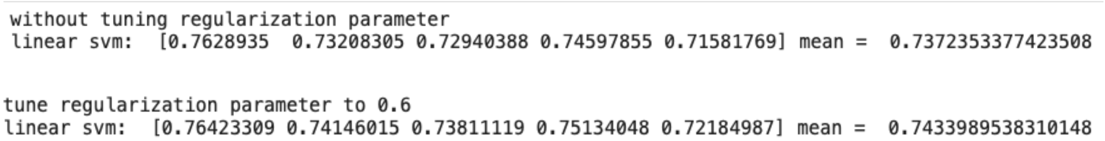
5. When we do smoothing by adding 1 to document frequency, the best result is achieved by multinomial NB with 0.780, and linear svm, logistic regression and decision tree are 0.756,0.754, 0.582 respectively. Since it does not improve the performance, we would not adept it later.
6. **multinomialNB** would produce its **best** result **when smoothing parameter alpha is set to 6**. (the accuracy

rate increases around 0.6% than accuracy rate when alpha is set to default value.)



```
tune the smoothing parameter to 0.6
multinomial naive bayes:  [0.76892163 0.74279973 0.74681849 0.75134048 0.74530831] mean = 0.76004268992422727
```

7. **linearSVC** would produce a better result **when the regularization parameter C is set to 0.6**. (the accuracy rate increases around 0.7% compare to the accuracy rate when C is set to default value.)



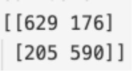
```
without tuning regularization parameter
linear svm:  [0.7628935  0.73208305 0.72940388 0.74597855 0.71581769] mean =  0.7372353377423508


tune regularization parameter to 0.6
linear svm:  [0.76423309 0.74146015 0.73811119 0.75134048 0.72184987] mean =  0.7433989538310148
```

8. I tried to split the dataset into train set and test set with different proportion. **When test size is 15% of dataset, the cross validation score is better than that when test size is 30%**

9. Results and conclusion

Overall, the multinomial naïve bayes gives me the best performance (mean of 0.78 in cross validation sets) when we don' t do any feature extraction to the data. The confusion matrix is



```
[[629 176]
 [205 590]]
```

and the final validation of this model is 0.761.

The second best model is multinomial naïve bayes when we only lemmatize the sentence and remove rare words(min_df=2), this model gives me an accuracy rate of 0.775.

The third best model is naïve bayes when we only stem the sentence, this has accuracy rate of 0.769.

The forth best model is linear svm when we don' t extract any feature from it, this gives accuracy rate if 0.766.

The fifth best model is multinomial naïve bayes when we only remove stop words, this gives accuracy rate of 0.765.