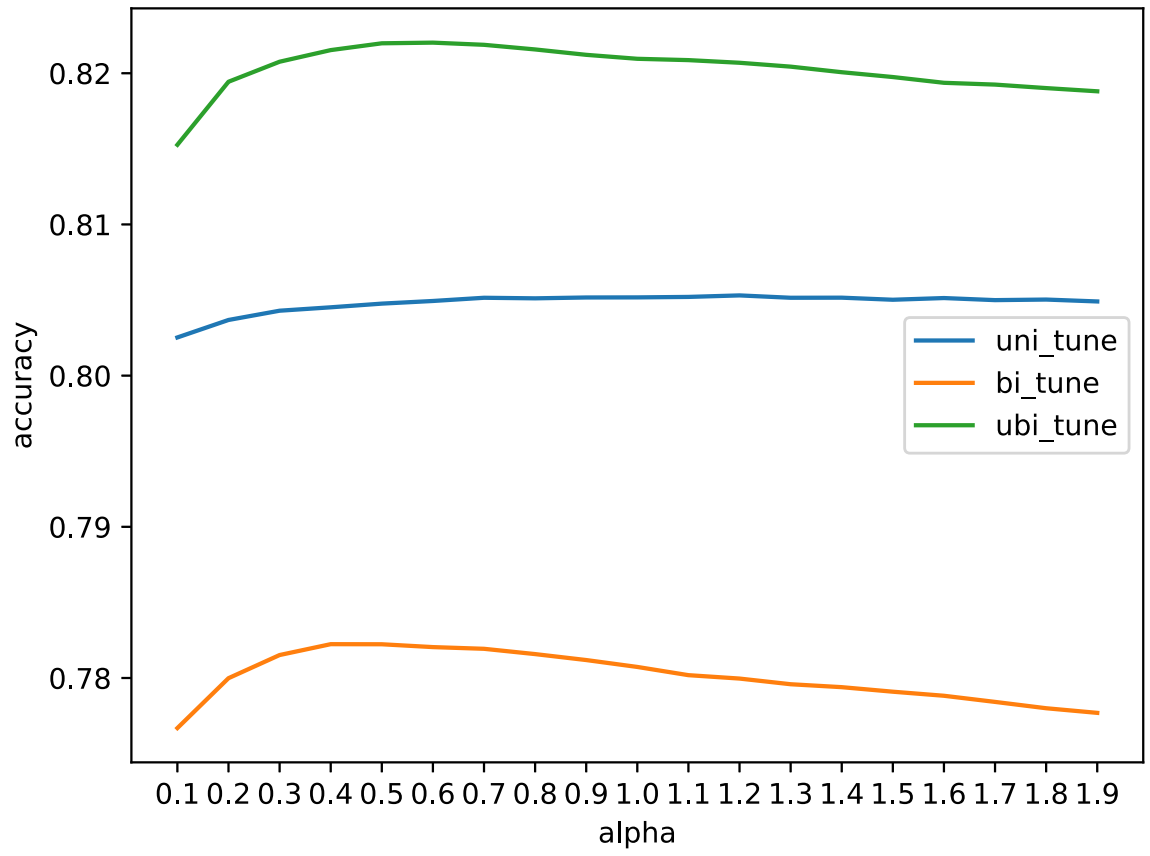
Assignment2

Introduction:

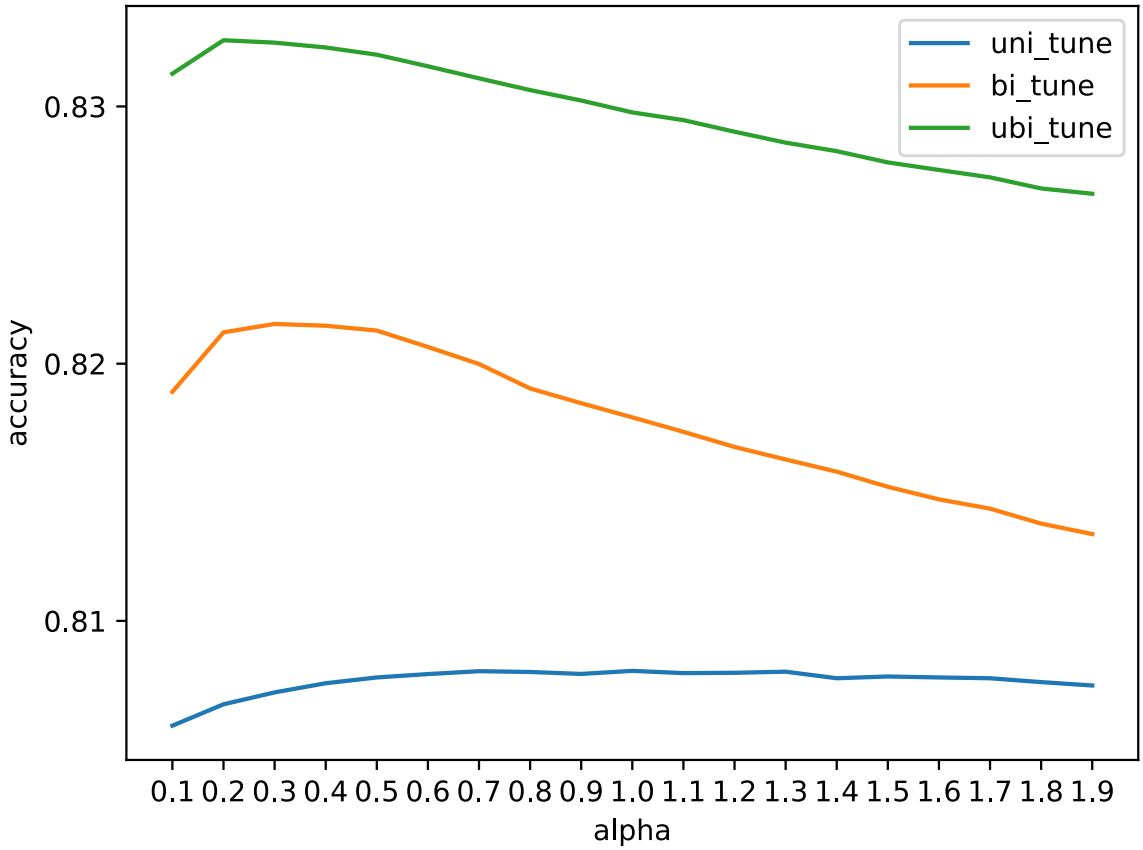
In this assignment, i use python to do a data mining on amazon comment database, and train a classifier to predict the accuracy of the comment, i consider the situation in bigram and unigram. In this process i also tune the parameter alpha in MNB to make sure it can get a best accuracy.

Tune paramater:

Dataset without stop words:



Dataset with stopwords:



So, i find that from two tables, for dataset with stop_words, the best alpha is 0.3, for dataset without stopwords, the best alpha is 0.4.

Consequence:

Stopwords removed	Text features	Accuracy(test)
Yes	Unigrams	0.80515625
Yes	Bigrams	0.7803125
Yes	Unigrams+Bigrams	0.82016875
No	Unigrams	0.807525
No	Bigrams	0.820825
No	Unigrams+Bigrams	0.82898125

After i choose the best alpha in stopwords dataset and no_stopwords dataset, i train the classifier and get the consequence on test dataset.

Q1:

From the consequence table, i can find that the accuracy on stopwords_dataset is apparent higher than dataset on no_stopwords dataset, especailly on bigrams situation. In my opinion, because stopwords has some words like 'not', 'few', 'nor' etc. which directly imply the negative effect of a sentence, if we remove that, we will find that in training dataset, some sentence which remove the stopwords look like a positive sentence but labeled with a negative sentence, as a result, in test dataset, some positive sentence will be predicted as a negative sentence, especailly for bigram, like 'not happy', after removing not, it definitly a positive word.

Q2:

From the consequence table, i can find that Unigrams plus bigrams is the best situation, because, when use unigrams plus bigrams, it can solve the problem of Q1 and reduce the influence of removing negetive stopwords. Instead of this, by combining unigram and bigram, they will increase the units of frequency vector have a larger training dataset, so it can increase the accuracy. The vector will contain the possibility of words one by one and also contain the words two by two. so it apparently increase the accuracy.