MSCI 641:Text Analytics
Assignment 2

Vyas Anirudh Akundy
Student ID: 20765080

1 Accuracy Table:

Stopwords removed	Text Features	Accuracy(Test Set)
yes	unigrams	80.48%
yes	bigrams	80.4%
yes	unigrams + bigrams	82.24%
no	unigrams	80.8%
no	bigrams	83.16%
no	unigrams + bigrams	**83.46**%

```
Accuracy values:
Unigrams with Stop Words : 80.80388427269996,Alpha : 0.6
Unigrams without Stop Words : 80.48353188507357,Alpha : 0.5
Bigrams with Stop Words : 83.16022624887376,Alpha : 0.4
Bigrams without Stop Words : 80.39843828211032,Alpha : 0.6
Unigrams + Bigrams with Stop Words : 83.46055661227349,Alpha : 0.2
Unigrams + Bigrams without Stop Words : 82.2392131344479,Alpha : 0.5
```

Figure 1: All Acuracies after tuning alpha

2 Answers to questions

2a: ***Question:***Which condition performed better: with or without stopwords? Write a brief paragraph (5-6 sentences) discussing why you think there is a difference in performance.
Answer:

- From the given accuracy table above, we can see that the condition **With Stop Words** is performing better. After tuning the classifier for the best alpha(0.2), the text with the Stop Words has achieved an accuracy of 83.46%
- Removing stop words may not always improve the performance as it depends on the particular application.
- In an application as sentimental analysis, some stop words maybe necessary to distinguish the sentence from being positive or negative which is

very important.
Example sentence : "The phone wasn't in a good condition when it was delivered to me."
Here, "wasn't" is a stop word. If this is removed, then because of the word "good", the classifier may classify it as a positive sentence, when it is clearly expressing a negative sentiment.

- This is one of the reasons I think that the condition with stop words is performing better.

2b: ***Question:***
Which condition performed better: unigrams, bigrams or unigrams+bigrams? Briefly (in 5-6 sentences) discuss why you think there is a difference?
Answer:

- Unigrams + Bigrams performed the best out of all the cases.
- A model will be able to perform better when it is able to capture the context of the sentences and better able to predict the next words.
- As we have seen in class in one of the lectures, bigrams have a lower perpexlity than unigrams which shows that bigrams perform slightly better.
- Also it depends on the size of the dataset. If the dataset is very big then higher N-grams models usually perform better on unseen data.
- Unigrams and bigrams combined seem to capture the actual meaning of the words than any one individually because it is more easier to predict what comes next in case of phrases like "How are " as opposed to when we are trying to predict what comes next when we only have a single word like for example, "What ". The number of possible words to predict are very high for the latter as compared to the former.