CS 434 Report 1

Team number:8

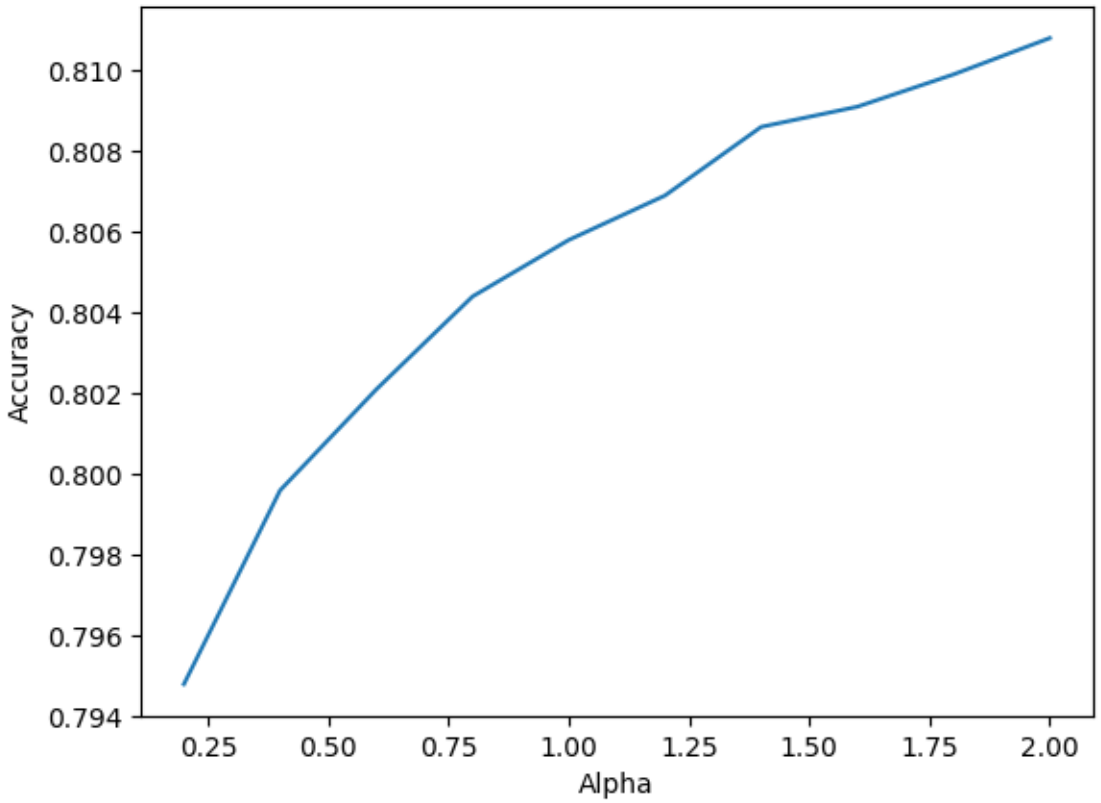
Team member: Adam Stewart, Hao Deng, Yuhang Chen

3. Apply the learned Naive Bayes model to the validation set (the next 10k reviews) and report the validation accuracy of the your model

```
Step3: validation accuracy is 0.8058
```

The validation accuracy of our model is 80.58%

4. Report the results by creating a plot with value of α on the x-axis and the validation accuracy on the y-axis. Comment on how the validation accuracy change as α changes and provide a short explanation for your observation



This graph show that accuracy will increase while alpha increase.

Multinomial:

$$P(x = w_i|y = 1) = \frac{\#\text{ of times word } i \text{ appeared in spam emails} + \alpha}{\text{total \# words in spam emails} + |V|\alpha}$$

Base on this formula, we can tell that alpha will change the probability of each words. And the change of probability will impact the mode and change the prediction eventually.

5. Please describe your strategy for choosing the value ranges and report the best parameters (as measured by the prediction accuracy on the validation set) and the resulting model's validation accuracy

First, I test different values of alpha from 2 to 10, I found the trend that the higher alpha is, the better it is.

```
Step1: Generate BOW
alpha: 2
alpha: 4
alpha: 6
alpha: 8
alpha: 10
[2, 4, 6, 8, 10]
[0.8108, 0.8166, 0.8207, 0.8228, 0.8253]
```

Second, I test different max_features from 3000 to 9000, I found that the higher it is, the better it is.

```
Step5:Tune max_features
max_features_test: 3000
max_features_test: 1000
max_features_test: 5000
max_features_test: 7000
max_features_test: 9000
[1000, 3000, 5000, 7000, 9000]
[0.7786, 0.7974, 0.8136, 0.8201, 0.826]
```

So I decided to test more combos of alpha and max_features. It is a huge computation, but by using parallel programming, I can make it faster than normal.

```
Step1: Generate BOW
Step5: Tune both
max_features_test: 10000 alpha: 10
max_features_test: 10000 alpha: 20
max_features_test: 10000 alpha: 30
max_features_test: 20000 alpha: 10
max_features_test: 20000 alpha: 20
max_features_test: 20000 alpha: 30
max_features_test: 30000 alpha: 10
max_features_test: 30000 alpha: 20
max_features_test: 30000 alpha: 30
[(10000, 10), (10000, 20), (10000, 30), (20000, 10), (20000, 20), (20000, 30),
(30000, 10), (30000, 20), (30000, 30)]
[0.8476, 0.8457, 0.8451, 0.8507, 0.8472, 0.8457, 0.8501, 0.848, 0.8457]
The best combo of (Features,alpha) is:
(20000, 10)
0.8507
```

The best I got is when maxFeatures is 20,000 and alpha is 10. The accuracy is 85 percent.