1. What is the effect of removing stop words in terms of precision, recall, and accuracy? Show a plot or a table of these results.

Stop words are words that are commonly used to construct a sentence like is, are, an. In context of ourr spam or ham activity, these words can be considered noise as they might appear frequent but contributes less to the spam orr rham prediction algorithm. Removing stop_words increased the accuracy, precision, and recall of the model. The table below will show the difference of each evaluation metrics value if the model is trained with and without the stop words.

	Accurarcy	**Precision**	**Recall**
Stop words rermoved	0.9490	0.9927	0.9298
Stop words not rermoved	0.9429	0.9938	0.9197

2. Experiment on the number of words used for training. Filter the dictionary to include only words occurring more than k times (1000 words, then k > 100, and k = 50 times). For example, the word “offer” appears 150 times, that means that it will be included in the dictionary.

	Accuracy	**Precision**	**Recall**
1000 Words	0.9490	0.9927	0.9298
k > 100	0.9423	0.9933	0.9191
k = 50	0.8061	0.9469	0.7496

The table above shows the evaluation value of our NB model in predicting a spam or ham email. As we can see, in using 1000x common words, our model became exceptional in terms of accuracy, precision, and recall. While using k > 100 provided a little bit of difference from the 1000 common words used, k = 50 shows a significant decrease in predictive performance. This is because when we set k = 50, the model’s vocabulary for common words might have increased affecting its ability to distinguish spam or ham email.

3. Discuss the results of the different parameters used for Lambda smoothing. Test it on 5 varying values of the λ (e.g. λ = 2.0, 1.0, 0.5, 0.1, 0.005), Evaluate performance metrics for each.

Lambda is used for laplace smoothing. This is usually applied to avoid zero probabilities. As we can see in the table below, λ = 1.0 gives the best performance while λ = 2.0 performed worst. This shows that bigger lambda does not guarantee improved performance.

Lambda	Accuracy	Precision	Recall
2.0	0.8960	0.9813	0.8594
1.0	0.9490	0.9927	0.9298
0.5	0.8960	0.9813	0.8594
0.1	0.8960	0.9813	0.8594
0.005	0.8960	0.9813	0.8594

Lambda is significant to apply laplace smoothing in the spam orr ham prediction naive bayes model. Based on the table above, the best value for lambda is 1.0 as it acquires the highest precision, recall, and accuracy and maintains the balance in smoothing and the use of frequent words.

Github Link:
https://github.com/Kaloy2202/CMSC197_jupyter_notebook/tree/de54ee31d8b6a5b4ee929ae5e38133753813a371/hw4