Name: Allen Solomon L. Tam
2019-03019
BS in Computer Science IV

HW – 4

Guide questions:
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.

If we remove the stop words in the datasets, the precision, recall, and accuracy should improve because the model can focus on the words or statements that are typically used in a spam email.

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.

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.

4. What are your recommendations to further improve the model?

I think trying to use different machine learning algorithm altogether, such as decision tree , SVM, or any models that we think can enhance the performance of the naive Bayes model will be helpful. Also using a different type of naive Bayes such as Gaussian naive Bayes or Multinomial naive Bayes will also improve the model. Using different type of feature extraction of feature selection method can also be helpful to improve the quality of the features being used in the prediction.