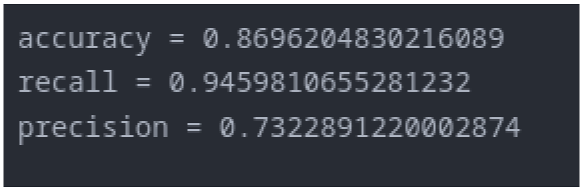
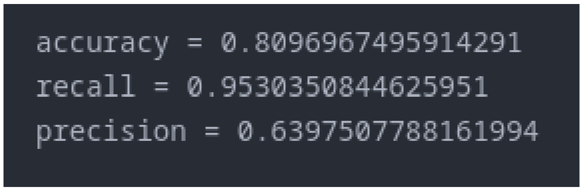
John Gabriel Cabatu-an
CMSC 197 - 1
Hw4

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.
    - Using stop words:



```
accuracy = 0.8696204830216089
recall = 0.9459810655281232
precision = 0.7322891220002874
```

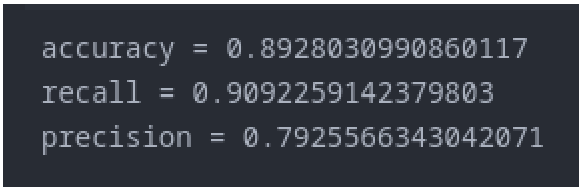
    - Not using stop words:



```
accuracy = 0.8096967495914291
recall = 0.9530350844625951
precision = 0.6397507788161994
```

    - Not using stop words improves recall for a little bit but decreases accuracy and precision.

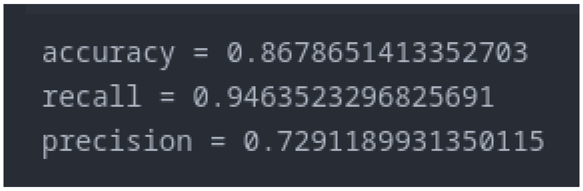
2. Experiment on the number of words used for training. Filter the dictionary to include only words occurring more than k times (1000 words, then k > 100, and k = 50 times). For example, the word “offer” appears 150 times, that means that it will be included in the dictionary.



```
accuracy = 0.8928030990860117
recall = 0.9092259142379803
precision = 0.7925566343042071
```

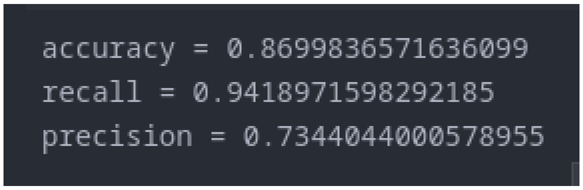
    - Doing so improves accuracy and precision but reduces recall.

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 = 2:



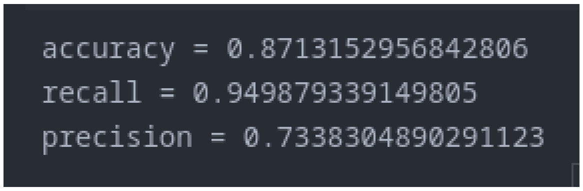
```
accuracy = 0.8678651413352703
recall = 0.9463523296825691
precision = 0.7291189931350115
```

- Lambda = 5



```
accuracy = 0.8699836571636099
recall = 0.9418971598292185
precision = 0.7344044000578955
```

- Lambda = 0.05



```
accuracy = 0.8713152956842806
recall = 0.949879339149805
precision = 0.7338304890291123
```

- As observed with the performance metrics the lower the lambda is the better it performs.

4. What are your recommendations to further improve the model?
    - I did not see any use for the feature matrices although they took the most runtime when running. With 10000 words the data frame is prone to fragmentation which slows down performance. I would not make the feature matrices altogether and filter more words that contain numbers.