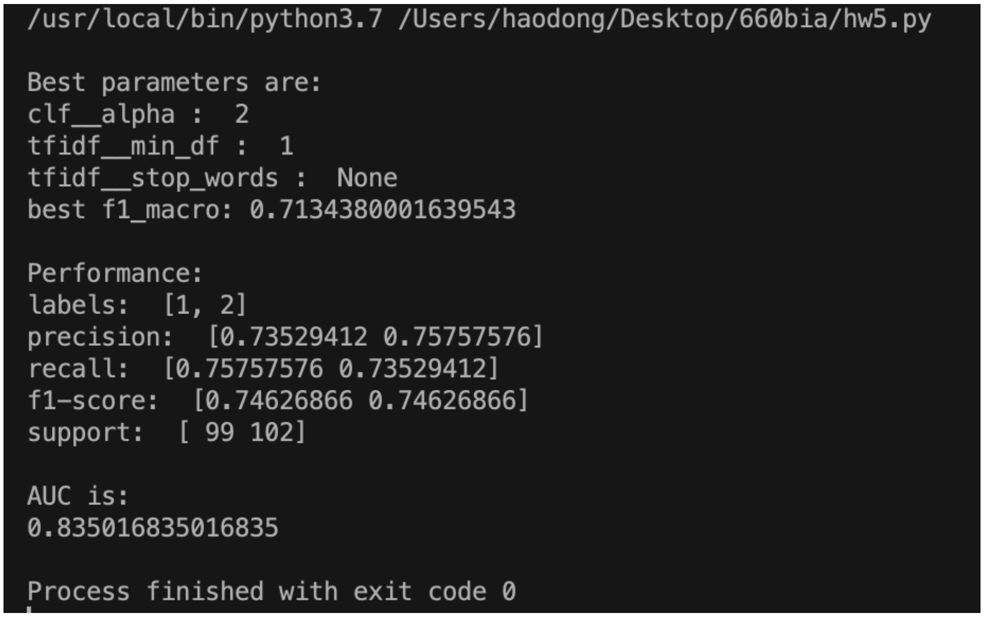
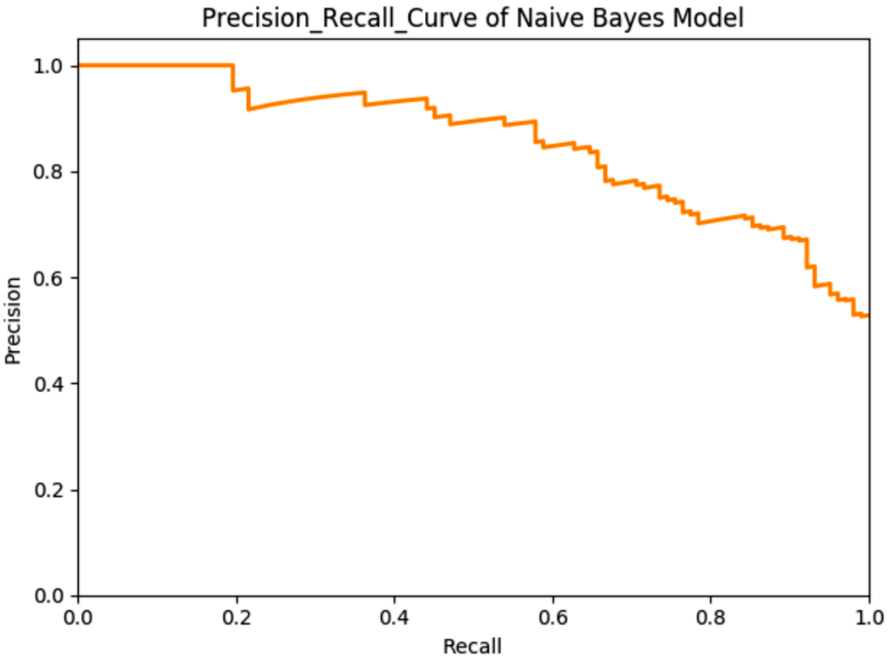
Question1

Following are screenshots of my results:

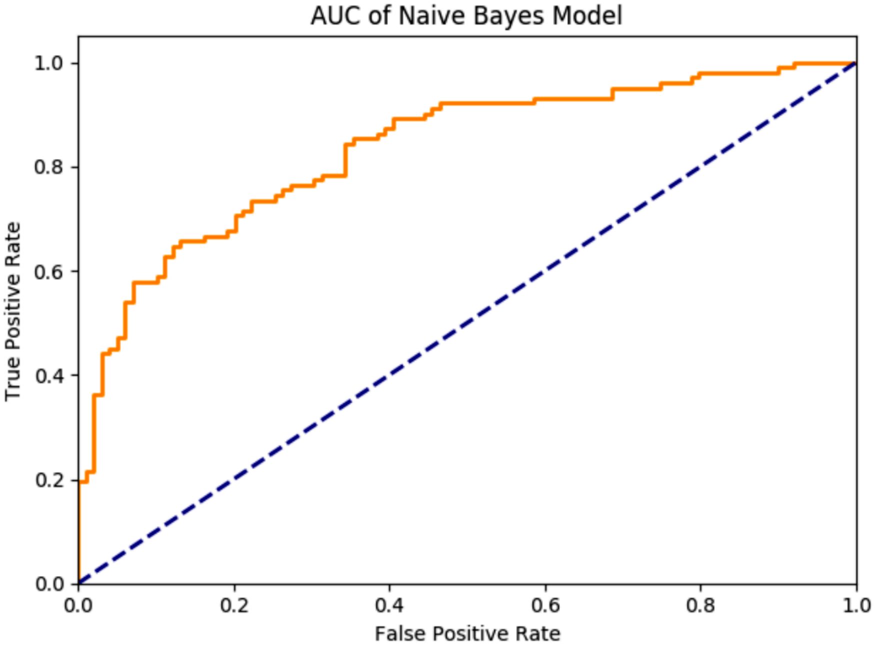


```
/usr/local/bin/python3.7 /Users/haodong/Desktop/660bia/hw5.py

Best parameters are:
clf__alpha :  2
tfidf__min_df :  1
tfidf__stop_words :  None
best f1_macro: 0.7134380001639543

Performance:
labels:  [1, 2]
precision:  [0.73529412 0.75757576]
recall:  [0.75757576 0.73529412]
f1-score:  [0.74626866 0.74626866]
support:  [ 99 102]

AUC is:
0.835016835016835

Process finished with exit code 0
```

Precision-recall curve:

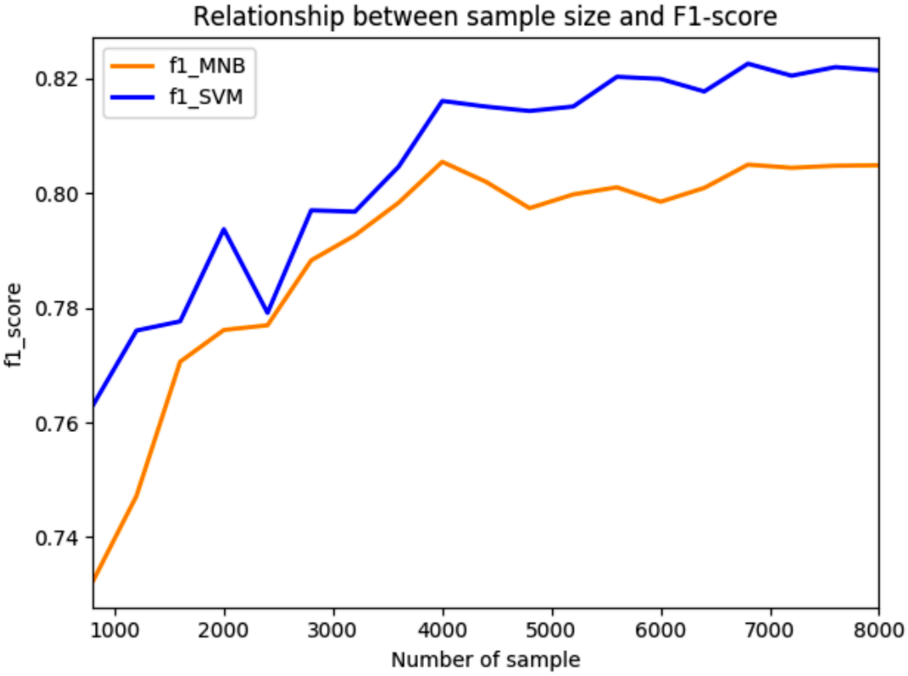


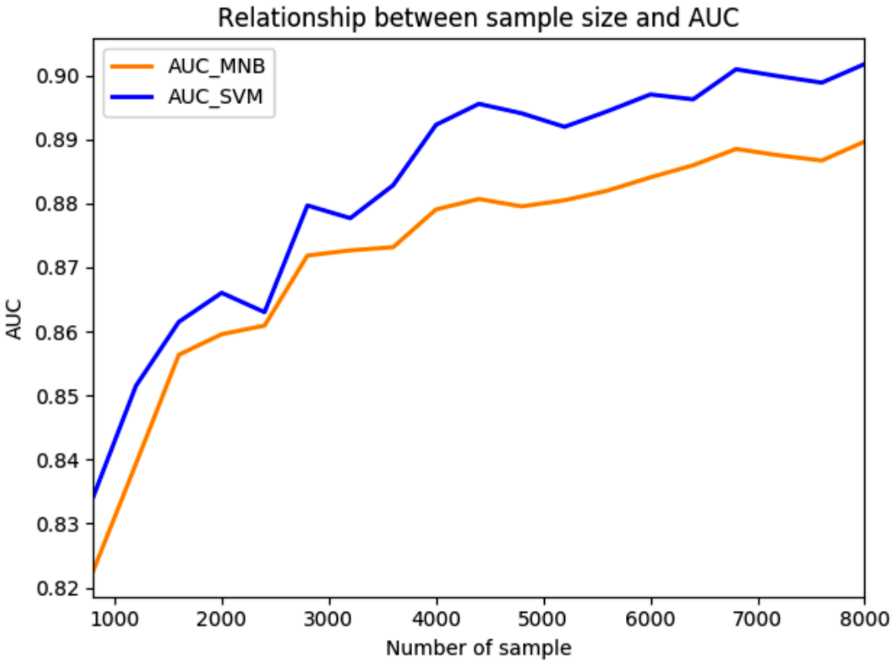
ROC curve:



Question2:

Screenshot of my results:



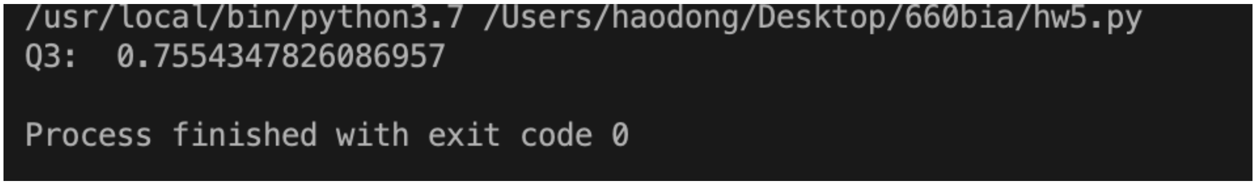


Analysis:

1. When the sample size increase, the AUC and f1-score also increase. Therefore, the bigger the sample size is, the better these two models will perform.
2. From the plots, I think at least 4000 samples are needed for both models, because after 4000 samples, the accuracy will increase slowly.
3. From the plots, we can find that the SVM classifier is always perform better than Naïve Bayes classifier, especially when the sample size over 4000.

Question3:

Following is my result:



```
/usr/local/bin/python3.7 /Users/haodong/Desktop/660bia/hw5.py
Q3:  0.7554347826086957

Process finished with exit code 0
```

I have features ‘unigram’, ‘bigram’, ‘trigram’ and the cosine similarity result. And the model I use is still SVM. The AUC value I got is about 0.7554