EE219 Project 1 Report by

Zixia Weng 305029822 & Shunji Zhan 405030387

January 29

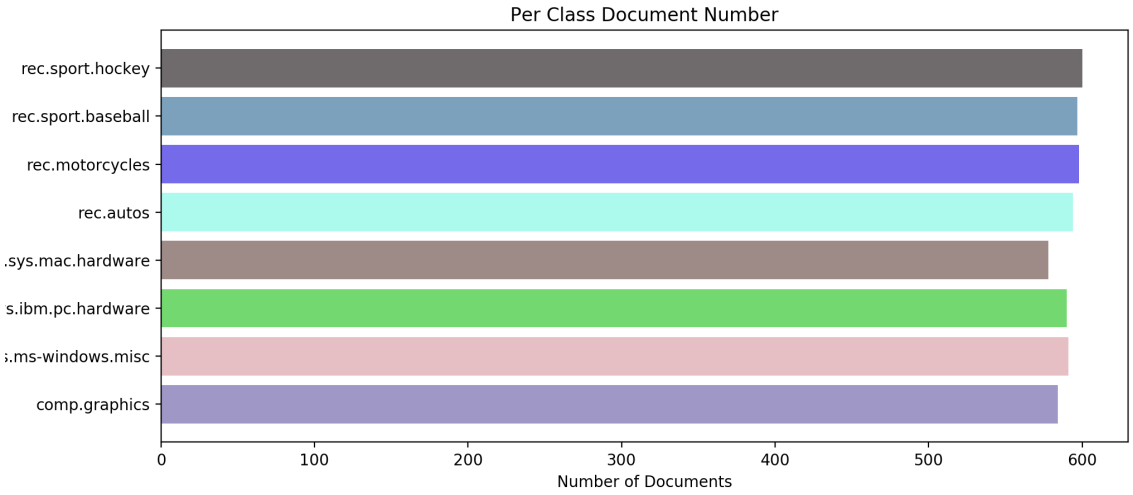
Classification Analysis on Textual Data

Winter 2018

Part 1

Dataset and Problem Statement:

a) For 8 categories, we plot the histogram for each of them with their corresponding documents. And they are evenly distributed.



Part 2

Modeling Text Data and Feature Extraction:

b)

First we create a tokenizer to tokenize each each documents into words:

```
def stemTokenizer(text):
    stemmer = SnowballStemmer("english")
    temp = "".join([i if ord(i) < 128 else ' ' for i in text])                #remove non-ascii
```

```
    temp = re.sub('[,.-:/()?><{}*$#&]','', temp)                    #remove some special
symbols
    tem = "".join(c for c in temp if c not in string.punctuation)    #excluding punctuations
    return [stemmer.stem(item) for item in temp.split()]
```

Excluding Stop Words & using stemmed version of words:

```
from sklearn.feature_extraction import text
stop_words = text.ENGLISH_STOP_WORDS
vectorizer = text.TfidfVectorizer(
    stop_words=stop_words,
    encoding='unicode',
    analyzer='word',
    min_df=5,  #or 2
    tokenizer=stemTokenizer
)
```

When Min_df = 2, our Number of extracted terms is **18723**
When Min_df = 5, our Number of extracted terms is **8032**

c) Finding most significant terms

For the following classes: comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, misc.forsale, soc.religion.christian.

We have 10 most significant terms

'comp.sys.ibm.pc.hardware', 'scsi', 'drive', 'ide', 'se', 'line', 'sbject', 'organ', 'card', 'mb', 'control'
'comp.sys.mac.hardware', 'line', 'mac', 'subject', 'organ', 'use', 'simm', 'appl', 'scsi', 'problem','drive'
'misc.forsale', 'line', 'sbject', 'sale', 'organ', 'nivers', 'new', 'se', 'offer', 'dos', 'nntppostinghost'
'soc.religion.christian', 'god', 'christian', 'jess', 'chrch', 'sbject', 'peopl', 'line', 'say', 'christ', 'believ'

Part 3

Feature Selection:

d)

We applied LSI to the TFxIDF matrix corresponding to the 8 classes by TruncatedSVD from importing sklearn.decomposition:
we got (4732, 50)

And Alternatively, reduce dimensionality through Non-Negative Matrix Factorization (NMF) : we got (4732, 50)

Part 4

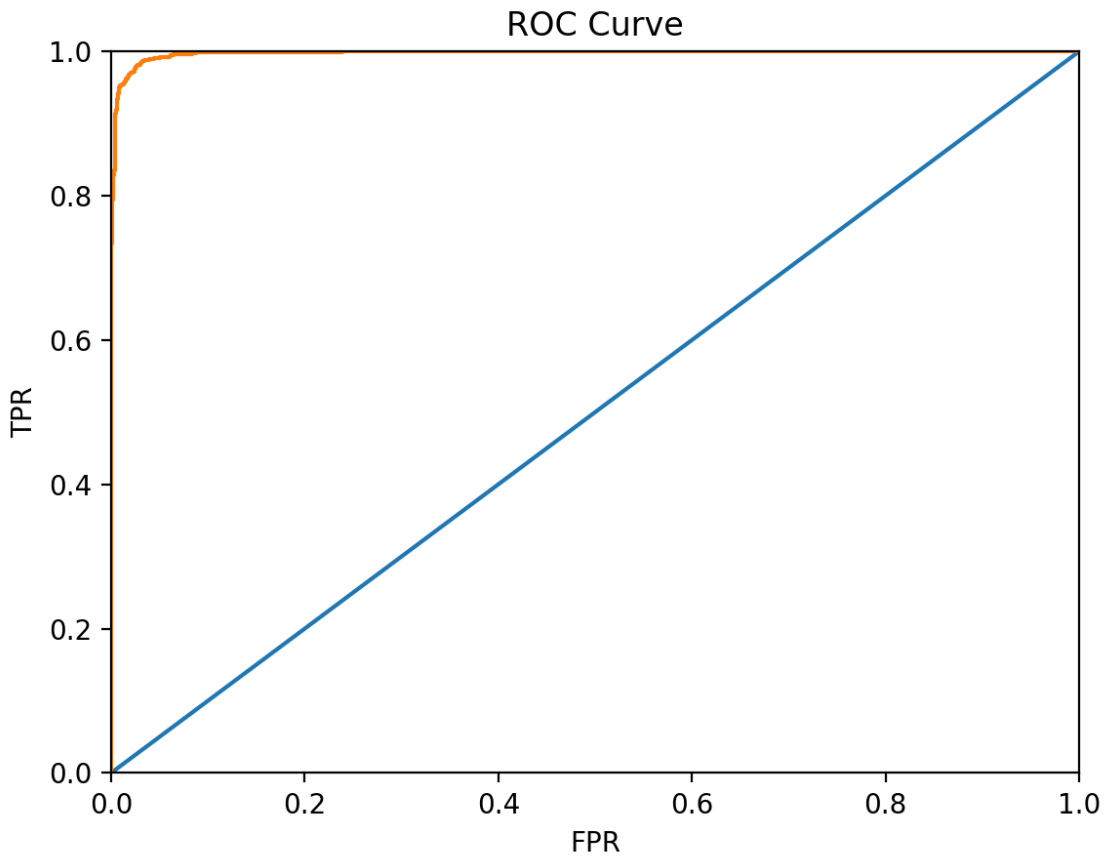
Learning Algorithms:

e)

min_df = 5 with LSI:

hard margin SVM classifier (SVC) $\gamma = 1000.0$

Roc Curve



Confusion Matrix:

[[1526 34]
[47 1543]]

Accuracy: Accuracy of Hard Margin SVM: 0.97

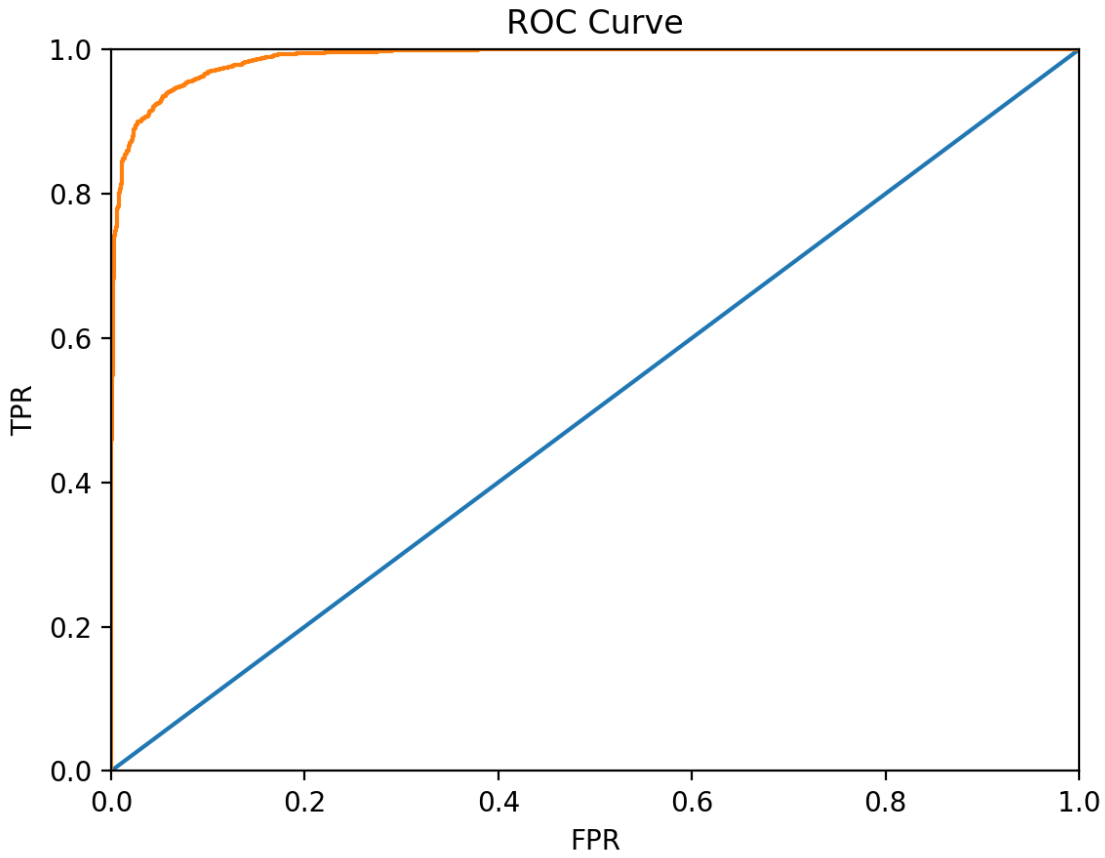
Recall for Computer Technology: 0.98

Recall for Recreational Activity: 0.97
average: 0.97

Precision for Computer Technology: 0.97
Precision for Recreational Activity: 0.98
average:0.97

Soft margin SVM classifier (SVC) γ = 0.001

Roc Curve



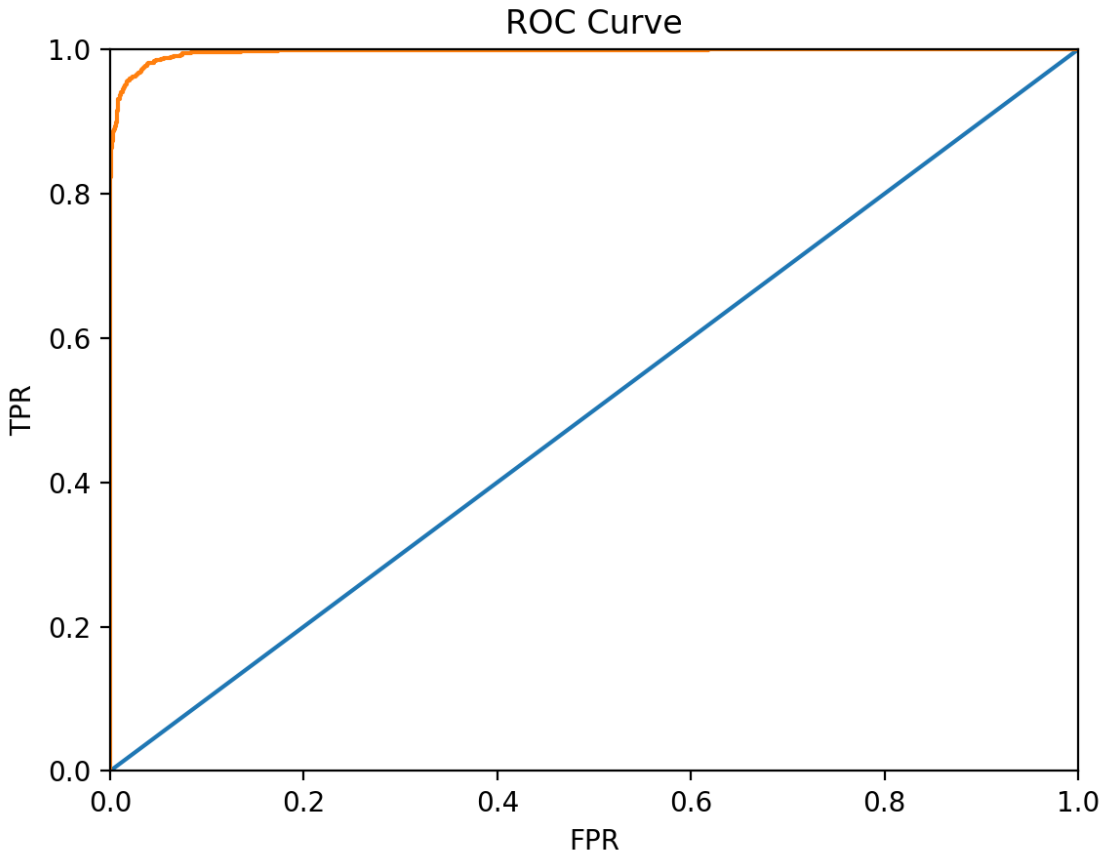
Confusion Matrix:
[[1399 161]
[11 1579]]
Accuracy: Accuracy of Hard Margin SVM: 0.95
Recall for Computer Technology: 0.90
Recall for Recreational Activity: 0.99
average: 0.95

Precision for Computer Technology: 0.99
Precision for Recreational Activity: 0.91
average: 0.95

min_df = 5 with NMF:

hard margin SVM classifier (SVC) $\gamma = 1000.0$

Roc Curve

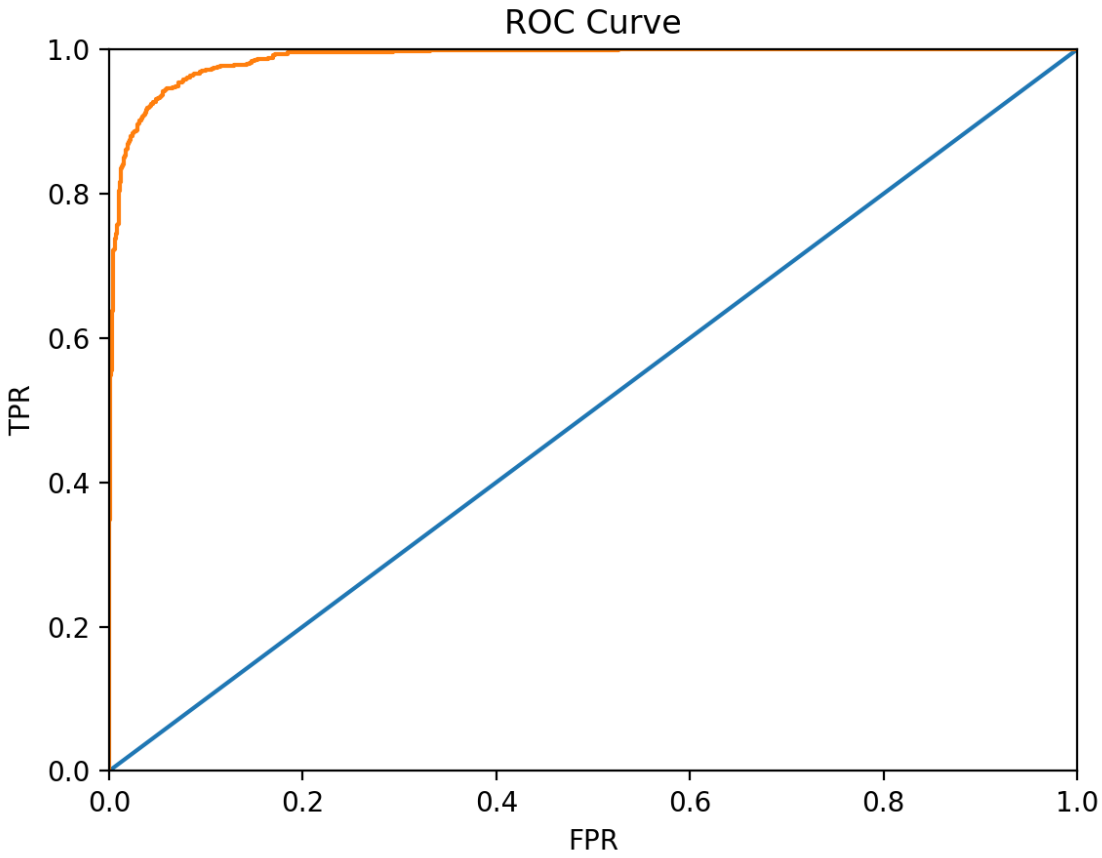


Confusion Matrix:
[[1489 71]
[26 1564]]
Accuracy: Accuracy of Hard Margin SVM: 0.97
Recall for Computer Technology: 0.95
Recall for Recreational Activity: 0.98
average: 0.97

Precision for Computer Technology: 0.98
Precision for Recreational Activity: 0.96
average:0.97

Soft margin SVM classifier (SVC) γ = 0.001

Roc Curve



Confusion Matrix:
[[93 1467]
[0 1590]]
Accuracy: Accuracy of Hard Margin SVM: 0.53
Recall for Computer Technology: 0.06
Recall for Recreational Activity: 1.00
average: 0.53

Precision for Computer Technology: 1.00
Precision for Recreational Activity: 0.52

average: 0.76

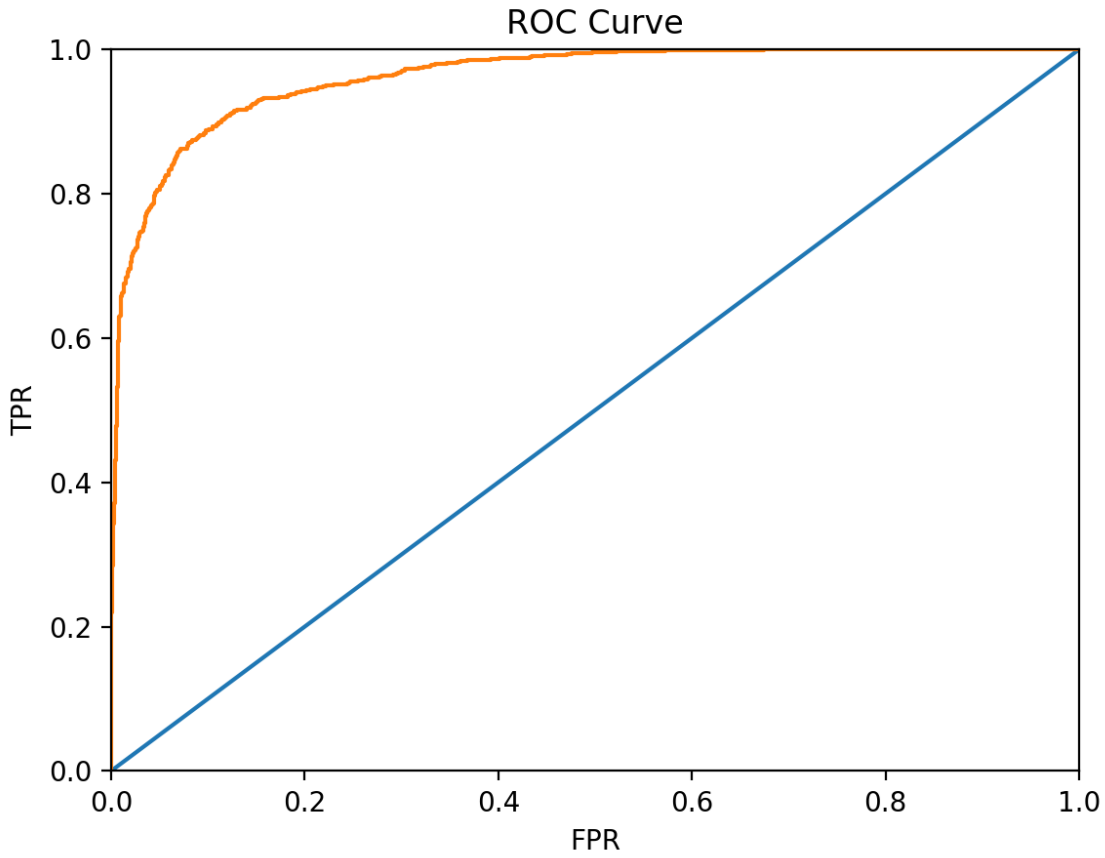
---

f)

Accuracy: 0.94358 | gamma: 0.001
Accuracy: 0.96556 | gamma: 0.01
Accuracy: 0.96767 | gamma: 0.1
Accuracy: 0.97464 | gamma: 1
Accuracy: 0.97655 | gamma: 10
Accuracy: 0.97676 | gamma: 100
Accuracy: 0.97549 | gamma: 1000
Best Accuracy: 0.97676 | gamma: 100

SVM classifier (SVC) $\gamma = 100.0$ with LSI

Roc Curve



Confusion Matrix:

[[1514 46]
[26 1564]]

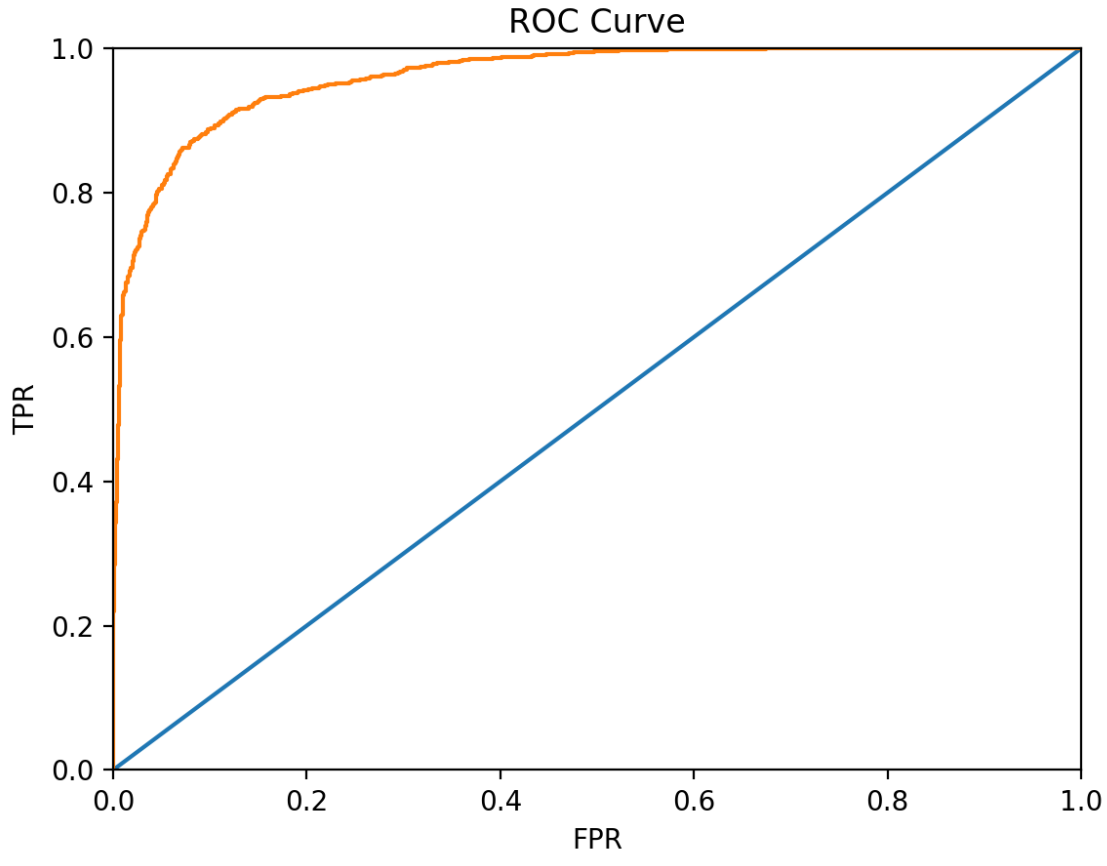
Accuracy: Accuracy of Hard Margin SVM: 0.98
Recall for Computer Technology: 0.97
Recall for Recreational Activity: 0.98
average: 0.98

Precision for Computer Technology: 0.98
Precision for Recreational Activity: 0.97
average:0.98

Accuracy: 0.51839 | gamma: 0.001
Accuracy: 0.92012 | gamma: 0.01
Accuracy: 0.95879 | gamma: 0.1
Accuracy: 0.96640 | gamma: 1
Accuracy: 0.97359 | gamma: 10
Accuracy: 0.97634 | gamma: 100
Accuracy: 0.96576 | gamma: 1000
Best Accuracy: 0.97634 | gamma: 100

SVM classifier (SVC) $\gamma = 100.0$ with NMF

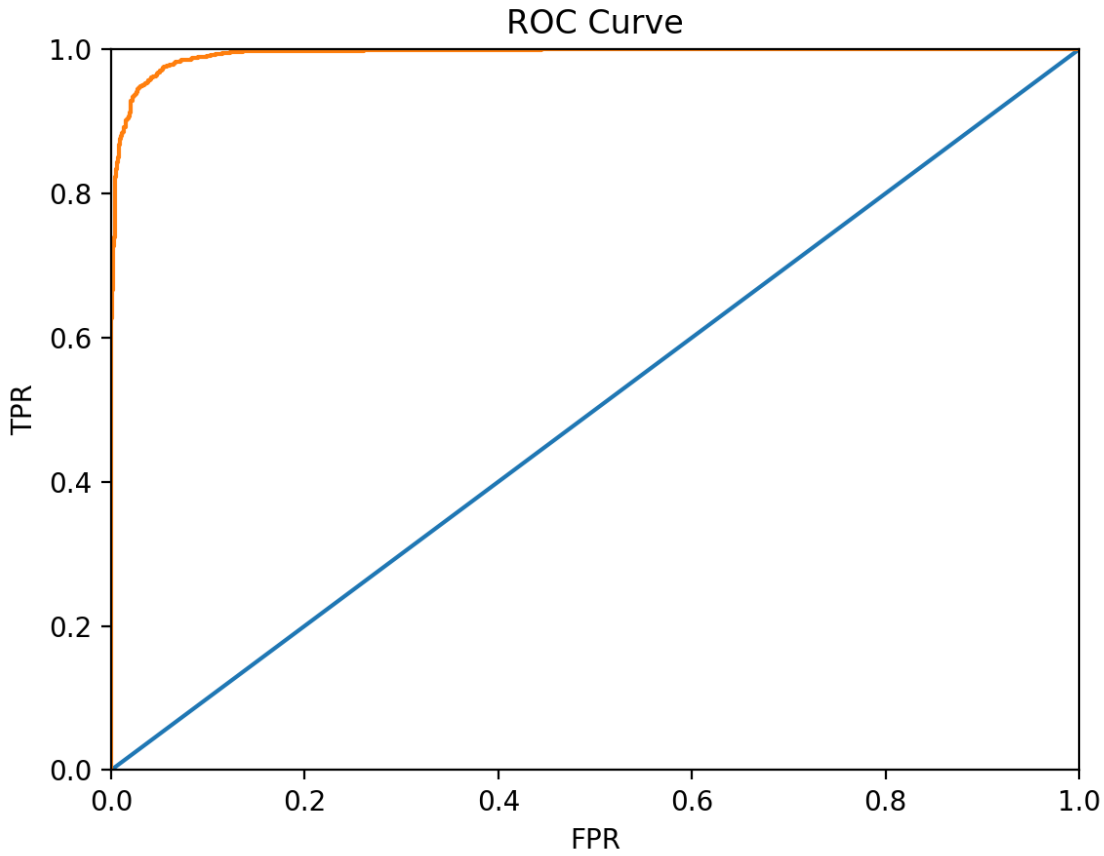
Roc Curve



Confusion Matrix:
[[1494 66]
[35 1555]]
Accuracy: Accuracy of Hard Margin SVM: 0.97
Recall for Computer Technology: 0.96
Recall for Recreational Activity: 0.98
average: 0.97

Precision for Computer Technology: 0.98
Precision for Recreational Activity: 0.96
average:0.97

g)
Naive Beyes Classifier
Using NMF with min_df = 5
Roc Curve:

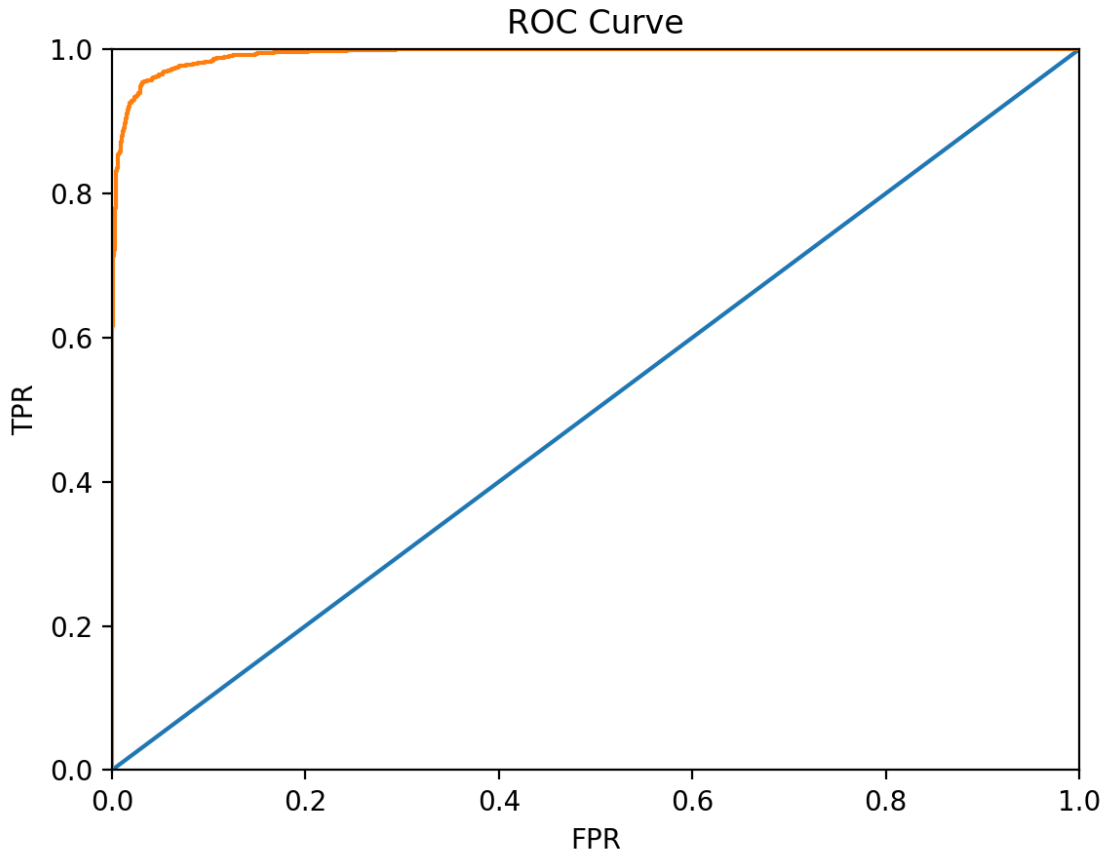


Con
fusion Matrix:

[[1428 132]
[20 1570]]
Accuracy: Accuracy of Hard Margin SVM: 0.95
Recall for Computer Technology: 0.92
Recall for Recreational Activity: 0.99
average: 0.95

Precision for Computer Technology: 0.99
Precision for Recreational Activity: 0.92
average:0.95

Naive Bayes Classifier
Using LSI with min_df = 5
Roc Curve:

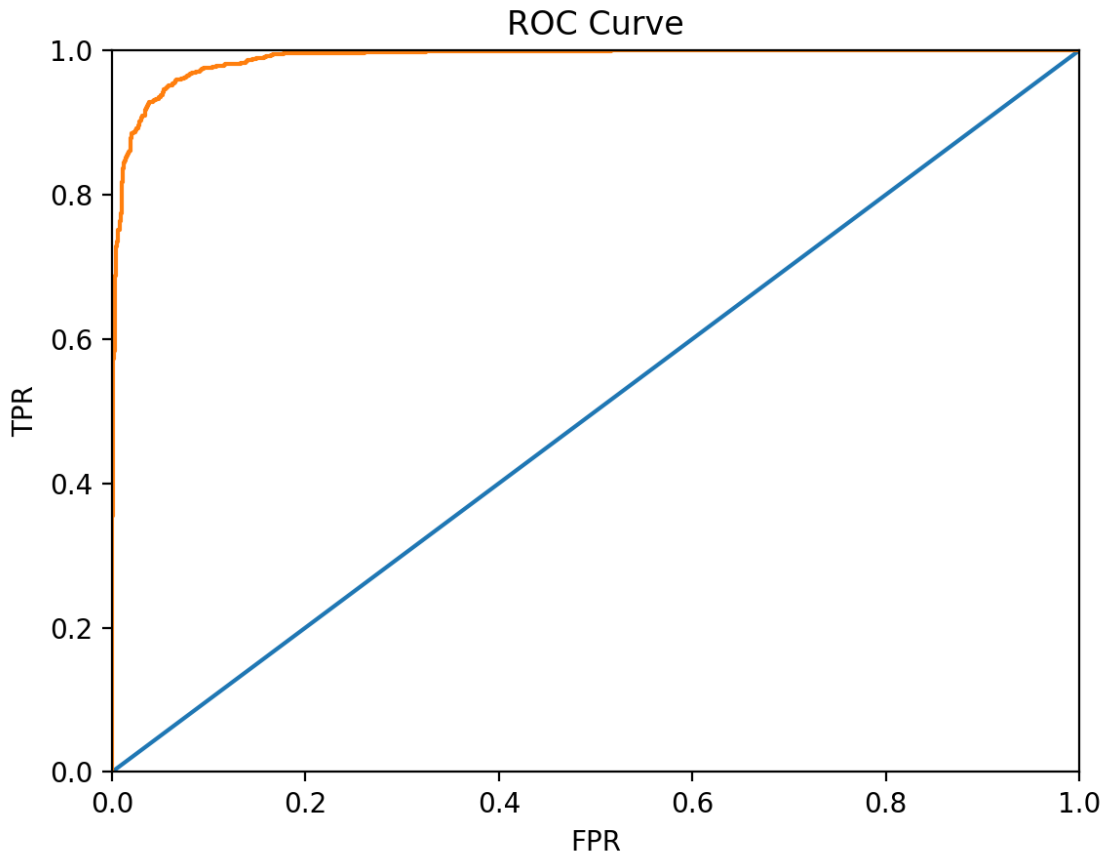


Confusion Matrix:
[[1451 109]
[38 1552]]
Accuracy: Accuracy of Hard Margin SVM: 0.95

Recall for Computer Technology: 0.93
Recall for Recreational Activity: 0.98
average: 0.95

Precision for Computer Technology: 0.97
Precision for Recreational Activity: 0.93
average:0.95

h)
Logistic Regression Classifier
Using NMF with min_df = 5
Roc Curve:

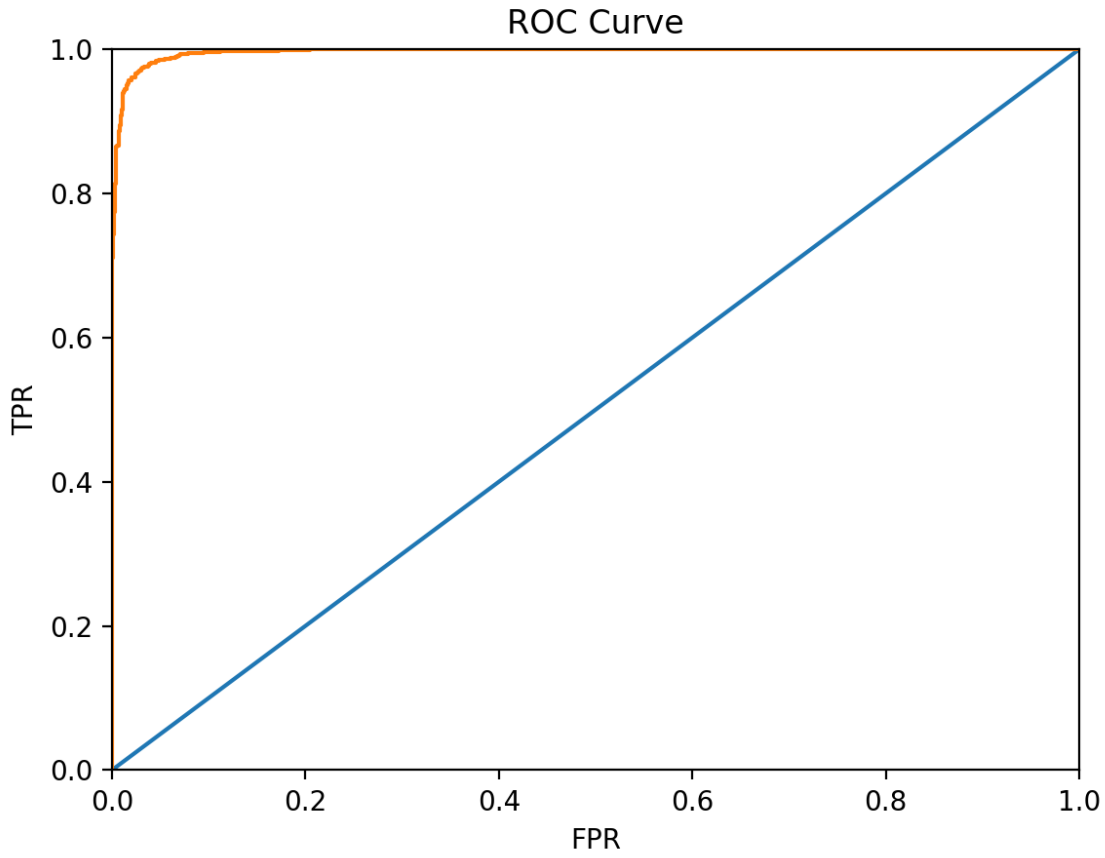


Con
fusion Matrix:
[[1444 116]
[61 1529]]
Accuracy: Accuracy of Hard Margin SVM: 0.94
Recall for Computer Technology: 0.93

Recall for Recreational Activity: 0.96
average: 0.94

Precision for Computer Technology: 0.96
Precision for Recreational Activity: 0.93
average:0.94

Logistic Regression Classifier
Using LSI with min_df = 5
Roc Curve:



Con
fusion Matrix:
[[1549 11]
[187 1403]]
Accuracy: Accuracy of Hard Margin SVM: 0.94
Recall for Computer Technology: 0.99
Recall for Recreational Activity: 0.88

average: 0.94

Precision for Computer Technology: 0.89
Precision for Recreational Activity: 0.99
average:0.94

i)

Now, repeat part (h) by adding a regularization term to the optimization objective. Try both l_1 and l_2 norm regularizations and sweep through different regularization coefficients, ranging from very small ones to large ones.

We tried 0.001, 0.1, 1, 10, 1000 these 5 values to see how does the regularization parameter affect the test error. Wit both NMF and LSI method, they perform worse when c=0.001. After that when c/regularization parameters becomes larger, they performs better than before. And in comparison, LSI tends to be a little bit better than NMF

For l2 and l1, except the situation that c=0.001, in which l2 tends to perform better than l1 and has higher accuracy, in the other situation the hyperplane tend to become higher and higher and become more closer to the (0.0, 1.0)

For more detail, you can check our code and just type make to run.

part 5

Multiclass Classification:

Naive Bayes Classifcation:

Accuracy of multinomial naive Bayes: 0.789137380192

	Precesion	Recall
0	0.66	0.82
1	0.85	0.62
2	0.73	0.74
3	0.95	0.97
Average	0.80	0.79

Confusion Matrix:
[[320 27 43 2]
[90 239 54 2]
[71 15 289 15]

[1 0 10 387]]

Accuracy of multiclass SVM classification(One Vs one): 0.833226837061

	Precesion	Recall
0	0.72	0.83
1	0.83	0.74
2	0.81	0.81
3	0.99	0.94
Average	0.84	0.83

Confusion Matrix:
[[327 37 27 1]
[69 284 31 1]
[52 19 317 2]
[4 2 16 376]]

Accuracy of multiclass SVM classification(One Vs Rest): 0.846006389776

	Precesion	Recall
0	0.75	0.82
1	0.83	0.75
2	0.83	0.83
3	0.98	0.97
Average	0.85	0.85

Confusion Matrix:
[[323 41 25 3]
[61 290 32 2]
[46 18 323 3]
[2 0 8 388]]

In the above, from e to i are all the data of vector when min_df = 5, Here are all the data when min_df = 2, for your favor: (You can always check our code and type make to run)

==========

```
================== Question e ==================
================================================
================== Hard Margin SVM with SVD ==================
Accuracy: 0.98
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.98      0.97      0.98      1560
Recreational activity       0.97      0.98      0.98      1590

          avg / total       0.98      0.98      0.98      3150

Confusion Matrix:
[[1516   44]
 [  32 1558]]
================== Soft Margin SVM with SVD ==================
Accuracy: 0.94
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.99      0.89      0.94      1560
Recreational activity       0.90      0.99      0.95      1590

          avg / total       0.95      0.94      0.94      3150

Confusion Matrix:
[[1389  171]
 [   9 1581]]
================== Hard Margin SVM with NMF ==================
Accuracy: 0.97
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.99      0.95      0.97      1560
Recreational activity       0.96      0.99      0.97      1590

          avg / total       0.97      0.97      0.97      3150

Confusion Matrix:
[[1489   71]
 [  20 1570]]
================== Soft Margin SVM with NMF ==================
Accuracy: 0.54
Classification Report:
                    precision    recall  f1-score   support
```

```
  Computer technology       1.00      0.08      0.15      1560
Recreational activity       0.53      1.00      0.69      1590

        avg / total       0.76      0.54      0.42      3150

Confusion Matrix:
[[ 122 1438]
 [   0 1590]]

==================================================
================== Question f ==================
==================================================
Accuracy: 0.94189 | gamma:  0.001
Accuracy: 0.96619 | gamma:  0.01
Accuracy: 0.96873 | gamma:  0.1
Accuracy: 0.97591 | gamma:  1
Accuracy: 0.97739 | gamma:  10
Accuracy: 0.97654 | gamma:  100
Accuracy: 0.97041 | gamma:  1000
Best Accuracy: 0.97739 | gamma: 10
================== Best SVM with SVD ==================
Accuracy: 0.98
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.99      0.97      0.98      1560
Recreational activity       0.97      0.99      0.98      1590

        avg / total       0.98      0.98      0.98      3150

Confusion Matrix:
[[1510   50]
 [  22 1568]]
Accuracy: 0.52388 | gamma:  0.001
Accuracy: 0.88842 | gamma:  0.01
Accuracy: 0.95816 | gamma:  0.1
Accuracy: 0.96957 | gamma:  1
Accuracy: 0.97443 | gamma:  10
Accuracy: 0.97549 | gamma:  100
Accuracy: 0.96999 | gamma:  1000
Best Accuracy: 0.97549 | gamma: 100
================== Best SVM with NMF ==================
Accuracy: 0.97
Classification Report:
                    precision    recall  f1-score   support
```

```
  Computer technology       0.98      0.96      0.97      1560
Recreational activity       0.96      0.98      0.97      1590

        avg / total       0.97      0.97      0.97      3150

Confusion Matrix:
[[1502   58]
 [  33 1557]]

==================================================
==================  Question g  ==================
==================================================
================== Naive Beyes Classifier with NMF ==================
Accuracy: 0.95
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.99      0.92      0.95      1560
Recreational activity       0.93      0.99      0.96      1590

        avg / total       0.96      0.95      0.95      3150

Confusion Matrix:
[[1433  127]
 [  16 1574]]
================== Naive Beyes Classifier with SVD ==================
Accuracy: 0.72
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.64      1.00      0.78      1560
Recreational activity       1.00      0.44      0.61      1590

        avg / total       0.82      0.72      0.69      3150

Confusion Matrix:
[[1560    0]
 [ 892  698]]

==================================================
==================  Question h  ==================
==================================================
================== Logistic Regression Classifier with NMF ==================
Accuracy: 0.95
Classification Report:
                    precision    recall  f1-score   support
```

```
  Computer technology       0.98      0.92      0.95      1560
Recreational activity       0.93      0.98      0.96      1590

        avg / total       0.96      0.95      0.95      3150

Confusion Matrix:
[[1439  121]
 [  25 1565]]
================== Logistic Regression Classifier with SVD ==================
Accuracy: 0.79
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.70      1.00      0.82      1560
Recreational activity       1.00      0.58      0.74      1590

        avg / total       0.85      0.79      0.78      3150

Confusion Matrix:
[[1560    0]
 [ 663  927]]

==============================================
================== Question i ==================
==============================================
================== Logistic Regression Classifier with c=0.001, penalty=l1 with NMF
==================
/Users/shunji/Library/Python/2.7/lib/python/site-packages/sklearn/metrics/classification.py:1135:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with
no predicted samples.
  'precision', 'predicted', average, warn_for)
Accuracy: 0.50
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.50      1.00      0.66      1560
Recreational activity       0.00      0.00      0.00      1590

        avg / total       0.25      0.50      0.33      3150

Confusion Matrix:
[[1560    0]
 [1590    0]]
================== Logistic Regression Classifier with c=0.001, penalty=l1 with SVD
==================
```

Accuracy: 0.50
Classification Report:

	precision	recall	f1-score	support
Computer technology	0.50	1.00	0.66	1560
Recreational activity	0.00	0.00	0.00	1590
avg / total	0.25	0.50	0.33	3150

Confusion Matrix:
[[1560 0]
[1590 0]]
================== Logistic Regression Classifier with c=0.1, penalty=l1 with NMF ==================
Accuracy: 0.86
Classification Report:

	precision	recall	f1-score	support
Computer technology	0.89	0.83	0.86	1560
Recreational activity	0.84	0.90	0.87	1590
avg / total	0.87	0.86	0.86	3150

Confusion Matrix:
[[1288 272]
[154 1436]]
================== Logistic Regression Classifier with c=0.1, penalty=l1 with SVD ==================
Accuracy: 0.94
Classification Report:

	precision	recall	f1-score	support
Computer technology	0.91	0.98	0.94	1560
Recreational activity	0.98	0.91	0.94	1590
avg / total	0.94	0.94	0.94	3150

Confusion Matrix:
[[1523 37]
[143 1447]]
================== Logistic Regression Classifier with c=1, penalty=l1 with NMF ==================
Accuracy: 0.97
Classification Report:

	precision	recall	f1-score	support

```
  Computer technology       0.97      0.96      0.96      1560
Recreational activity       0.96      0.97      0.97      1590

        avg / total       0.97      0.97      0.97      3150

Confusion Matrix:
[[1502   58]
 [  51 1539]]
================== Logistic Regression Classifier with c=1, penalty=l1 with SVD
==================
Accuracy: 0.92
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.86      1.00      0.92      1560
Recreational activity       1.00      0.84      0.91      1590

        avg / total       0.93      0.92      0.92      3150

Confusion Matrix:
[[1554    6]
 [ 256 1334]]
================== Logistic Regression Classifier with c=10, penalty=l1 with NMF
==================
Accuracy: 0.97
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.98      0.97      0.97      1560
Recreational activity       0.97      0.98      0.97      1590

        avg / total       0.97      0.97      0.97      3150

Confusion Matrix:
[[1506   54]
 [  34 1556]]
================== Logistic Regression Classifier with c=10, penalty=l1 with SVD
==================
Accuracy: 0.94
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.90      1.00      0.94      1560
Recreational activity       1.00      0.89      0.94      1590

        avg / total       0.95      0.94      0.94      3150
```

Confusion Matrix:
[[1553 7]
[182 1408]]
================== Logistic Regression Classifier with c=1000, penalty=l1 with NMF ==================
Accuracy: 0.97
Classification Report:

	precision	recall	f1-score	support
Computer technology	0.98	0.97	0.97	1560
Recreational activity	0.97	0.98	0.97	1590
avg / total	0.97	0.97	0.97	3150

Confusion Matrix:
[[1506 54]
[30 1560]]
================== Logistic Regression Classifier with c=1000, penalty=l1 with SVD ==================
Accuracy: 0.55
Classification Report:

	precision	recall	f1-score	support
Computer technology	0.52	1.00	0.69	1560
Recreational activity	1.00	0.11	0.19	1590
avg / total	0.76	0.55	0.44	3150

Confusion Matrix:
[[1560 0]
[1420 170]]
================== Logistic Regression Classifier with c=0.001, penalty=l2 with NMF ==================
Accuracy: 0.50
Classification Report:

	precision	recall	f1-score	support
Computer technology	0.00	0.00	0.00	1560
Recreational activity	0.50	1.00	0.67	1590
avg / total	0.25	0.50	0.34	3150

Confusion Matrix:
[[0 1560]
[0 1590]]

========== Logistic Regression Classifier with c=0.001, penalty=l2 with SVD ==========
Accuracy: 0.89
Classification Report:

	precision	recall	f1-score	support
Computer technology	1.00	0.77	0.87	1560
Recreational activity	0.82	1.00	0.90	1590
avg / total	0.91	0.89	0.89	3150

Confusion Matrix:
[[1207 353]
 [1 1589]]
========== Logistic Regression Classifier with c=0.1, penalty=l2 with NMF ==========
Accuracy: 0.91
Classification Report:

	precision	recall	f1-score	support
Computer technology	0.99	0.82	0.90	1560
Recreational activity	0.85	0.99	0.92	1590
avg / total	0.92	0.91	0.91	3150

Confusion Matrix:
[[1284 276]
 [10 1580]]
========== Logistic Regression Classifier with c=0.1, penalty=l2 with SVD ==========
Accuracy: 0.82
Classification Report:

	precision	recall	f1-score	support
Computer technology	0.74	1.00	0.85	1560
Recreational activity	1.00	0.65	0.79	1590
avg / total	0.87	0.82	0.82	3150

Confusion Matrix:
[[1558   2]
 [552 1038]]
========== Logistic Regression Classifier with c=1, penalty=l2 with NMF ==========
Accuracy: 0.95
Classification Report:

```
                precision    recall  f1-score   support

  Computer technology       0.98      0.92      0.95      1560
Recreational activity       0.93      0.98      0.96      1590

        avg / total       0.96      0.95      0.95      3150

Confusion Matrix:
[[1439  121]
 [  25 1565]]
================== Logistic Regression Classifier with c=1, penalty=l2 with SVD
==================
Accuracy: 0.79
Classification Report:
                precision    recall  f1-score   support

  Computer technology       0.70      1.00      0.82      1560
Recreational activity       1.00      0.58      0.74      1590

        avg / total       0.85      0.79      0.78      3150

Confusion Matrix:
[[1560    0]
 [ 663  927]]
================== Logistic Regression Classifier with c=10, penalty=l2 with NMF
==================
Accuracy: 0.97
Classification Report:
                precision    recall  f1-score   support

  Computer technology       0.98      0.95      0.97      1560
Recreational activity       0.95      0.98      0.97      1590

        avg / total       0.97      0.97      0.97      3150

Confusion Matrix:
[[1486   74]
 [  28 1562]]
================== Logistic Regression Classifier with c=10, penalty=l2 with SVD
==================
Accuracy: 0.82
Classification Report:
                precision    recall  f1-score   support

  Computer technology       0.74      1.00      0.85      1560
Recreational activity       1.00      0.65      0.79      1590
```

```
        avg / total       0.87      0.82      0.82      3150

Confusion Matrix:
[[1560    0]
 [ 561 1029]]
================== Logistic Regression Classifier with c=1000, penalty=l2 with NMF
==================
Accuracy: 0.97
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.98      0.96      0.97      1560
Recreational activity       0.96      0.98      0.97      1590

        avg / total       0.97      0.97      0.97      3150

Confusion Matrix:
[[1502   58]
 [  33 1557]]
================== Logistic Regression Classifier with c=1000, penalty=l2 with SVD
==================
Accuracy: 0.75
Classification Report:
                    precision    recall  f1-score   support

  Computer technology       0.67      1.00      0.80      1560
Recreational activity       1.00      0.51      0.67      1590

        avg / total       0.83      0.75      0.74      3150

Confusion Matrix:
[[1560    0]
 [ 782  808]]

================================================
==================  Question j  ==================
================================================
================== naive bayes ==================
Accuracy: 0.83
Classification Report:
                      precision    recall  f1-score   support

comp.sys.ibm.pc.hardware       0.73      0.86      0.79       392
   comp.sys.mac.hardware       0.87      0.66      0.75       385
           misc.forsale       0.80      0.82      0.81       390
```

```
  soc.religion.christian       0.96      0.99      0.97       398

          avg / total       0.84      0.83      0.83      1565

Confusion Matrix:
[[338  24  28   2]
 [ 75 254  49   7]
 [ 48  13 320   9]
 [  2   0   3 393]]
================== one vs one ==================
Accuracy: 0.85
Classification Report:
                     precision    recall  f1-score   support

comp.sys.ibm.pc.hardware       0.73      0.84      0.78       392
   comp.sys.mac.hardware       0.81      0.78      0.79       385
          misc.forsale       0.89      0.83      0.86       390
  soc.religion.christian       1.00      0.96      0.98       398

          avg / total       0.86      0.85      0.85      1565

Confusion Matrix:
[[329  51  12   0]
 [ 63 300  22   0]
 [ 52  16 322   0]
 [  7   5   4 382]]
================== one vs rest ==================
Accuracy: 0.86
Classification Report:
                     precision    recall  f1-score   support

comp.sys.ibm.pc.hardware       0.76      0.84      0.80       392
   comp.sys.mac.hardware       0.83      0.76      0.79       385
          misc.forsale       0.88      0.84      0.86       390
  soc.religion.christian       0.97      0.99      0.98       398

          avg / total       0.86      0.86      0.86      1565

Confusion Matrix:
[[330  45  15   2]
 [ 58 291  30   6]
 [ 43  13 329   5]
 [  4   1   0 393]]
```

Thank you so much for looking our report into the end.

Best,

Zixia Weng & Shunji Zhan

Jan 30 2018