Python script is using processed version of 20news-bydate data set, which was downloaded from http://qwone.com/~jason/20Newsgroups/20news-bydate-matlab.tgz.

steps to run the script

1. curl -O http://qwone.com/~jason/20Newsgroups/20news-bydate-matlab.tgz
2. tar xzvf 20news-bydate-matlab.tgz
3. cd 20news-bydate\matlab
4. python ..\..\bayes_classifier.py

results on test data

The script will print below summary for test data as well as a file ‘result.csv’.

The summary print gives gold number labeld in test data set itself, number predicted by model, matched number of gold vs predict, as well as precision and recall for each class. In the end, it gives final average accuracy, precision, recall and f1.

And result.csv gives gold and predicted class for each document.

```
c:\wen\DSE\DSE210\20news-bydate\matlab>python ..\..\bayes_classifier.py
class, gold, predict, matched, precision, recall
   1, 318,    344,   240,     0.70,  0.75
  10, 397,    367,   349,     0.95,  0.88
  11, 399,    401,   380,     0.95,  0.95
  12, 395,    475,   359,     0.76,  0.91
  13, 393,    336,   262,     0.78,  0.67
  14, 393,    363,   324,     0.89,  0.82
  15, 392,    381,   335,     0.88,  0.85
  16, 398,    541,   377,     0.70,  0.95
  17, 364,    477,   324,     0.68,  0.89
  18, 376,    356,   320,     0.90,  0.85
  19, 310,    320,   185,     0.58,  0.60
   2, 389,    450,   299,     0.66,  0.77
  20, 251,    115,    94,     0.82,  0.37
   3, 391,    261,   213,     0.82,  0.54
   4, 392,    503,   304,     0.60,  0.78
   5, 383,    357,   278,     0.78,  0.73
   6, 390,    363,   300,     0.83,  0.77
   7, 382,    265,   238,     0.90,  0.62
   8, 395,    450,   356,     0.79,  0.90
   9, 397,    380,   356,     0.94,  0.90
performance result:
total accuracy     : 0.79
average precision  : 0.79
average recall     : 0.78
f1 score           : 0.39
```