Bayesian learning for classifying netnews text articles:

1) **Problem Statement:**
   Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents. We will provide a dataset containing 20,000 newsgroup messages drawn from the 20 newsgroups. The dataset contains 1000 documents from each of the 20 newsgroups.

2) **Implementation:**
   a) Step 1 – Initially after loading the data, we clean the data. Meaning only words are taken into consideration. This is done by removing special characters and new lines.

   b) Step 2 – Creating a bag-of-words. In this step we create a dictionary containing all the words in dataset.

   c) Step 3 – Bayes Algorithm. Once we have a bag-of-words and training and testing data splits, we calculate the probability vector consisting the probability of a testing instance with respect to all the classes. The prediction of the algorithm is given by finding the maximum probability. The probability is calculated by:

   For each class c:
   Pc = 0
   For each word in cleaned_file:
   Pc = Pc + log(P(word| c))

3) **Results:**
   The code gives output as follows:
   ---------- CREATING BAG OF WORDS ----------
   Reading words from alt.atheism
   Reading words from comp.graphics
   Reading words from comp.os.ms-windows.misc
   Reading words from comp.sys.ibm.pc.hardware
   Reading words from comp.sys.mac.hardware
   Reading words from comp.windows.x
   Reading words from misc.forsale
   Reading words from rec.autos

Reading words from rec.motorcycles
Reading words from rec.sport.baseball
Reading words from rec.sport.hockey
Reading words from sci.crypt
Reading words from sci.electronics
Reading words from sci.med
Reading words from sci.space
Reading words from soc.religion.christian
Reading words from talk.politics.guns
Reading words from talk.politics.mideast
Reading words from talk.politics.misc
Reading words from talk.religion.misc

The bag of words has 173001 words

Calculating Accuracy for Naive Bayes on the Newsgroup data

Accuracy = 84.0

Confusion Martix :

```
[[9711    3    0    9   40   13   72    0   35    0    0    9    0    2
    0   19    4    0    1   82]
 [ 506 7486    0  612  277  227  670   11   36    0    0    9  120   28
   18    0    0    0    0    0]
 [ 501 1615 2482 2854  933  542  968    7   12    0    0   18   56    0
   10    0    0    0    2    0]
 [ 521 1150   38 7012  774   31  374   13    0    0    0    0   80    7
    0    0    0    0    0    0]
 [ 516 1189    6  951 7042    0  280    0    0    0    0    0   16    0
    0    0    0    0    0    0]
 [ 500 1989   24  559  311 6254  346   13    0    0    0    0    0    0
    0    0    4    0    0    0]
 [ 549 1109    6  933  783    5 6340  109   12    0   10    0  144    0
    0    0    0    0    0    0]
 [ 917  908    7  652  701    6  548 6046  132    0    0    0   72    7
    0    0    4    0    0    0]
```

```
[1081  823    2  599  713    9  579  492 5700    0    0    0    0    0
    0    0    0    0    2    0]
 [1422  683   11  558  684    9  940  288  195 5183   13    0    0   14
    0    0    0    0    0    0]
 [1213  975   12  669  468   22 1044  305  167  381 4740    0    0    0
    0    0    2    0    2    0]
 [1774 1581    2  567  765  141  328  386   86    8    0 4326   32    0
    0    0    2    0    2    0]
 [ 584 1527   10 1323 1233   38 1064  602   44   11    0    9 3530    7
   18    0    0    0    0    0]
 [2213 1192    1  277  824   22  818  634  429  145    3   58  178 3191
    6    0    4    3    2    0]
 [1271 2296    1  379  672   82  821  753  201   83    2  114  328  137
 2850    0    6    0    4    0]
 [5664  447    0   81   97    6  153  584  147  127   28    6   36   95
    9 2477    0    0    2    1]
 [3639  168   18  385  138    0  293 1909  510   60    5  750   42   58
   56    5 1915    0   39   10]
 [5688  304    0  149  306    1  370  439  181  126   22  172   54  153
   87  162  347 1382   56    1]
 [4459  205    0  243  245    0  364 1176  369  183   35  613   53  272
  190   21  660   83  764   65]
 [7538   71    0   98   32    2  247  506  126   64   18  131   30   82
18  236  337    1  157  306]]
```

4) NOTE:

In case the code fails to run due to path error, please set the variable “path” to the current working directory.