A program that performs Naive Bayes Classification.

Data: Download the sampleTrain.txt, sampleTrain.vocab.txt, sampleTest.txt form a toy dataset. Another actual dataset comprises the files train.txt, test.txt and train_vocab.txt.

SampleTrain and train are training data, while those files named test are the corresponding test corpus

files. The vocab files give the vocabulary of the respective training file. The classifier is being built using the training data. The classifier finally runs and evaluate on test data. The second column in

the training and test files gives the gold standard true class for each document. The first column of these

files is the document id, the third column gives the words in the document. The columns are separated

by tab spaces.

Classes: There are 2 classes in the toy dataset: 0 and 1.

In the actual dataset, there are 5 classes: 0, 1, 2, 3, 4. Each class stands for a topic. This data was taken

from ``20 newsgroups dataset'' a popular dataset used for topic classification tasks. A newsgroup is

a discussion forum where some topic is discussed by a group of users. The dataset contains 5 of the 20 newsgroups. The topics corresponding to the classes are:

0 comp.sys.mac.hardware

1 comp.windows.x

2 rec.sport.baseball

3 sci.med

4 talk.religion.misc

The Naive Bayes classifier is using the document words as features. It computes a model given some training data and is able to predict classes on a new test set. It uses Laplace smoothing for feature likelihoods. There was no need for UNK token as the dataset has been simplified so that the test corpus only contains words seen during training. Also there was no need to smooth prior probabilities.

The program when run prints two parts, answer to (a) followed by answer to (b).

a) It uses sampleTrain.txt for training the model and the model is being used to predict classes for

documents in sampleTest.txt.

b) It uses train.txt to train the model and the model should be run on test.txt.

The feature likelihoods are only printed for the 4 words (computer, baseball, god, doctor). But the feature likelihoods and prior probabilities are printed for each class.