Task

[This is based on the original assignment I received for CE314 Natural Language Programming, at the University of Essex in 2016. I have edited it down to just the Question relevant to the program at hand.]

Write a computer program that performs Naive Bayes Classification.

Data: 3 files: `sampleTrain.txt, sampleTrain.vocab.txt` and `sampleTest.txt.`

`sampleTrain.txt` is the training data for building a classifier. The vocabulary of the training data is in `sampleTrain vocab.txt`. The classifier should be finally run and evaluated on test data in `sampleTest.txt`. The second column in the `sampleTrain.txt` and `sampleTest.txt` 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.

There are 2 classes in the data 0 and 1.

Task: Build a Naive Bayes classifier using the document words as features. It should compute a model given some training data and be able to predict classes on a new test set. For this assignment, use `sampleTrain.txt` for training a model and the model should be used to predict classes for documents in `sampleTest.txt`. Use Laplace smoothing for feature likelihoods. There is **no** need for UNK token. The dataset has been simplified so that the test corpus only contains words seen during training (so no need for UNK). There is also no need to smooth the prior probabilities.

The program when run should print the following:

```
Prior probabilities
class 0 =
class 1 =


Feature likelihoods
      great sad boring ...
class 0
class 1

Predictions on test data
d5 =
d6 =
d7 =
d8 =
d9 =
d10 =

Accuracy on test data =
```

The features in the feature likelihood table (great, sad, boring,...) can be printed in any order.