Information Retrieval Fall 2016

Programming Assignment 3

Assigned : 28th Oct 2016 **Due: 4th Nov 2016**

Sentiment Analysis (Naive Bays Classifier)

The purpose of this assignment is to build a sentiment classifier using the Naive Bayes classification techniques (Multinomial and Bernoulli) and a bag of words model as discussed in the class.

- You will be using the review polarity dataset (v2.0) which can be found at http://www.cs.cornell.edu/people/pabo/movie-review-data/. It contains 1000 positive and 1000 negative movie reviews. This should give you a balanced corpus.
- The classification should be done in two parts. The first part is for training only. To do that, you should randomly select and separate 66% documents of each class. The rest of the documents should be used for testing.
- You should develop a training program (For example train.exe), which should take two parameters as input: a directory name and a filename. This directory should have one subdirectory for each class. Each subdirectory should have all files of that particular class. For example:

```
txt_sentoken
 |___ pos
 |   |___ file1
 |   |___ file2
 |___ neg
     |___ file1
     |___ file2
```

You should write generic code which can run for more than 2 classes. The second parameter to train.exe would be the name of the file where you will store your Naive Bayes model (which in your case should be the a list of probabilities), For example, a sample call would be

train.exe txt_sentoken / myNBmodel.out

You can decide the format of the model file yourself. For example, it may contain all the class names in the first line, separated by spaces, followed by one word and it's probability for each class per line. You have to use add-one smoothing (Laplace smoothing).

- The second part of the classifications would be the prediction. For this purpose you have to develop a program (say test.exe) which takes two filenames as parameter. First parameter should be the name of the model file (that you have just trained) and the second parameter should be the input test file which needs to be classified. Your program should output the name of the predicted class. For example:

test.exe myNBmodel.out file3
spam

For this part, you can use the leftover data from the initial corpus. This means your task requires 3 programs: 2 for training (Bernoulli + Multinomial) and 1 for testing them.

- You should write a small script to report accuracy of your model along with the confusion matrix in this format.

```
a     b   <-- classified as
265   74 | a = neg
74    267 | b = pos

Correctly Classified Instances  =  532    =    78.2353 %
Incorrectly Classified Instances  =  148    =    21.7647 %
```

Submission Checklist

Output Files:

2 Training file

Source Code

2 code files for training

1 test file

script for calculating accuracy of all test data

Submit these files as one zipped file on Slate (name of file should be your roll number)

Rubric

1 point Implemented Naive Bayes using Bernoulli correctly

1 point Implemented Naive Bayes using Multinomial correctly

1 point test file correctly tests a new document and classification accuracy is calculated correctly