CSCI 6350/7350 Homework 4 (Due : beginning of class, Monday February 27th)

Experiment with text classification using naïve Bayes classifier.

Part A

Download the mini-newsgroup data set, mini-newsgroup.tar.gz, from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups).

Once downloaded, do :

```
gunzip mini-newsgroup.tar.gz
gar –xvf mini-newsgroup.tar
```

This creates the 20 directories corresponding to the 20 news groups. Each group contains a varying number of newsgroup entries. Download the naïve Bayes program, naïve.py, from the course page.

Perform the following tasks:

1. Learn the most characteristic words (at least 10) of each newsgroup. This involves learning the naïve probabilistic model based on the entire set of data, and display the words that have the highest conditional probabilities for each category. Discuss whether you think these words are good set of words for differentiating among the categories. (Program modification required)
2. Estimate the classification accuracy of naïve Bayes classifier for this data. Hold out 10 entries from each category and form a test data (use the 10 entries with the lowest file numbers in each categories). Use the remaining entries in the 20 categories as training data. Report the classification accuracy.
3. Learning curve study – for this task, we will study the learning performance of the naïve Bayes system. The learning curve is derived by gradually decreasing the size of the training data, starting from 100% training data, down to 5%. You need to perform at least the following set of experiments: 100%, 80%, 60% 40%, 20%, 10% and 5%. Record the classification accuracy at each step.
    - Create a table listing classification accuracies obtained for different training data sizes.
    - Then draw a plot with y-axis representing the classification accuracy and the x-axis representing the percentage training data.
    - Does the learning curve generated from your experiments confirm with your expectation? Discuss the results obtained.

Part B

Create your own text classification data and perform the classification on your data. Here are some ideas to get you started:

- Collect 20 or more junk mails and 20 or more regular mails from your mailbox and see if the classifier can learn to correctly classify new mails into junk vs. non-junk mails.
- Perform a search on the web for certain query. Collect all the retrieved articles and divide them into “interested” and “not interested” categories and learn to predict whether you will be interested in other pages returned from the same query.
- Collect movie synopsis online, and create categories of movies you like and you don’t like.
- … something related to your own interest.

You are not limited to learn between 2 categories, e.g., you can divide your mails to “extremely important”, “weekend reading”, “junk” categories. Try to include an equal number of articles in each category. For your learning task, make sure to hold out at least 10 articles from the categories to test for classification accuracy. Discuss your data (what is it about, how many instances in each of the categories, number of test data, etc.), the classification accuracy obtained for this task, and discuss your ideas to improve the data collection process, data preparation process, or classification method that may improve the classification accuracy.

Turn in a report addressing the questions listed in Part A and B. Also, electronically submit a tar file containing your modified program for Part A and all the data used in Part B. Submit the tar file through peerspace.cs.mtsu.edu, Tools/Assignment.