COMS 6998-4, Fall 2012
Statistical NLP for the Web

Due: October 4th, 2012, Thursday (11:59pm)

Total Points: 100

Text Categorization

You are given Newsgroup data set that has been previously used by researchers to compare text categorization algorithms. The data set consists of newsgroup postings in 20 different domains. You are required to use the subset of the given corpus for all of your experiments. You can implement you classifier in any programming language you wish; but it has to compile and run in one of the clic machines.

Q1. Naïve Bayes Classifier [50]

Data directory: /home/smaskey/CS6998-0412/hw1/q1_q2

(1) Implement a Naïve Bayes Classifier and train it using the documents in the train directory. The classifier should classify any new document into one of the two given classes of hockey or baseball. [35]
(2) Test your classifier using the test files in test directory. Report the precision, recall and f-measure and accuracy of your classifier. [10]
(3) Update your classifier with Laplace smoothing. Does the performance improve? [5]

Q2. Perceptron [50]

Data directory: /home/smaskey/CS6998-0412/hw1/q1_q2

(1) Come up with at least 3 features (or more) that you think are discriminative for two categories of the document from Q1 and implement them to represent the documents [10]
(2) Implement perceptron classification algorithm and estimate the weights using the feature vectors and corresponding class labels [35]
(3) Test your classifier using the test files in test directory. Report the precision, recall and f-measure and accuracy of your classifier. [5]

Extra Credit [5]

Can you think of a better smoothing technique than Laplace smoothing? Implement your smoothing technique. Does the performance improve further in Q1?

Instruction for Submission:

- Create a **submit** directory.
- For each submission and Final submissions, copy the source code files into the **submit** directory and include all the other files that are necessary for your program to run.
- **cd** into the **submit** directory.
- Mail your files to mulinksi@cs.columbia.edu using the command:

```
$ tar cvf - . | compress | uuencode temp_file | Mail -s "submit cs6998 hw1"  mulinski@cs.columbia.edu
```

After a short time you will get an automatic acknowledgement of your submission. Please note:

- If you do not get an answer after a few minutes, then your program did not go through. Please resubmit
- If you do not get back a listing of ALL your files (please check the file sizes to ensure that everything arrived without any problems) then resubmit!

If you submit once, and then decide to submit again, your second submission will overwrite the first. All the files from your first submission will automatically be wiped out.