COMS 6998-7, Spring 2010
Statistical Methods for NLP

Homework 1

Due: Feb 16, 2010 (4pm)

Total Points: 100

Text Categorization

You are given the 20 newsgroups dataset that is frequently used by researchers to compare text categorization algorithms. The corpus consists of newsgroup postings in 20 different domains and you will use a subset of it for all of your experiments. You can implement your classifier in any programming language that you prefer; but it has to compile and run in one of the CLIC machines.

Q1. Naïve Bayes Classifier [50]

Data directory: `/home/smaskey/CS6998/hw1/q1_q2`

(1) Implement a Naïve Bayes Classifier and train it using the documents in the `train` directory. The classifier should be able to classify a new document into one of the two given classes of hockey or baseball. [35]
(2) Test your classifier using the documents in the `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/hw1/q1_q2`

(1) Come up with at least 3 features that you think are discriminative for the two given categories and implement them to represent the documents [10]
(2) Implement the 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 the `test` directory. Report the precision, recall and f-measure and accuracy of your classifier. [5]

Extra Credit [5]

Can you devise a better smoothing technique than Laplace smoothing? Justify your proposed smoothing technique and implement it for the Naïve Bayes classifier in Q1. Does the performance improve further?