COMS 6998-7, Spring 2010
Statistical Methods for NLP

Homework 1
(Solutions)

Due: Feb 16, 2010 (4pm)

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/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]

Solution is available at /home/smaskey/CS6998/solns/hw1

Q2. Perceptron [50]

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

(1) Come up with at least 3 features (or more) that you think are discriminative for two categories of the document 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]

Solution is available at /home/smaskey/CS6998/solns/hw1