CS 6320 Natural Language Processing

Homework 4: Text Categorization

Due: Oct 22nd, 11:59pm

In this assignment, you will implement a simple naïve Bayes classifier to do text categorization.

Data to use: movie review data available from
http://www.cs.cornell.edu/People/pabo/movie-review-data/
Please use polarity dataset v2.0: 1000 positive and 1000 negative processed reviews.

Reference paper:
``A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Proceedings of the ACL, 2004.
There are many papers that have used this data set since then.

What you need to do:

1. Implement a naïve Bayes classifier as we discussed in the class.

There are two parts you need to do:

   a) Training: your program will take two lists of files: one containing all of positive review files, the other containing the negative ones, and output a model file.
   b) Testing and evaluation: your program will take a model file and two lists of files: positive and negative review lists, and output the classification accuracy for the test set.

2. Use your own naïve Bayes classifier to run a 10-fold cross validation classification for the data set and report the overall performance.

Note: what's 10-fold cross validation? You split the data into ten subsets, use one subset as test set, and the other 9 subsets for training; you do this for each of the subsets and report an average performance of the 10 runs.

How to split the data? You will use the same setup as used in the paper:

fold 1: files tagged cv000 through cv099, in numerical order
fold 2: files tagged cv100 through cv199, in numerical order
...
fold 10: files tagged cv900 through cv999, in numerical order

3. Write a report about your experiments and findings.

How to turn in your homework?

Please submit your report and your basic naïve Bayes classifier implementation via eLearning.

Extra credits:

You can also try some existing tools to do text categorization and play with different features.

Tools to consider:

- Weka (http://www.cs.waikato.ac.nz/ml/weka/). This package has implementations of many machine learning algorithms.
- SVM classifiers. (LibSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm/; SVMlight: http://svmlight.joachims.org/)

Features to consider:

- N-gram features (binary or count), possibly with some selection/pruning
- Others you may find from many papers that have used this data set