Homework #1: Text classification

Assigned: September 27, 2002  
Due: October 11, 2002

This assignment covers problems related to corpus analysis, information retrieval, information theory, text classification, and evaluation of the output of automatic systems that address these tasks. You will build and experiment with a simple system that discriminates between six categories and you will answer questions about evaluating this system and applying similar techniques to other problems.

1. (10 points for parts 1–7 combined) The directory /clair3/corpora/aquaint-split contains 3 GB of news data from the Associated Press, the New York Times, and Xinhua. For this assignment, a small subset of these articles (346 in total) have been extracted and copied to /clair4/class/lni-f02/data/APW19990520.

2. Each of these articles is assigned to a category. You can use the following unix command to see how many articles are in each category:
   
grep -i "<CATEGORY>" * | cut -f2 -d " | sort | uniq -c

3. Copy all these files into your own directory. Clean them up of any HTML markup. Given the simple markup used, you should be able to use a command like this:
   
grep -v "<" APW19990520.0222 > APW19990520.0222.clean

4. Use the first 299 files (0001-0299) for “training”. Use the rest for “testing”.
   
cat APW*.00??.clean APW*.01??.clean APW*.02??.clean >| training
   
cat APW*.03??.clean >| test

5. How many lines of text are there in each of the two new files (training and test)?
   
   wc -l training test

6. Using NSP, find the 50 most frequent n-grams (for n = 1…4) in the training file.
   
   /clair4/tools/nsp-v0.51/count.pl --ngram 1 training.out training

7. Produce a histogram of all n-grams. What is the percentage of n-grams (for n = 1…4) that appear only once?

8. (10 points) Compute the mutual information for each of the 50 most frequent bigrams. You can use a spreadsheet or NSP or write your own program.

9. (10 points) Compute the log-likelihood and the dice coefficient for each of the 50 most frequent bigrams. The NSP toolkit will come in handy again.
10 (10 points) Refer back to the list of categories for each file. Create six files: one that contains all (cleaned) training files in “yes-entertainment”, another for “yes-sports”, etc. Similarly, create six files with the complements of each category (e.g., “not-entertainment”, “not-sports”, etc.). Ignore the “political” category as it is much more infrequent than the others.

11 (10 points) For each category, determine which words have a high power for discriminating among the documents in each “yes” file from the documents in the corresponding “no” file. Use the methods presented in class, or any reasonable alternative of your own.

12 (10 points) Using the frequencies of the unigrams that you identified in the previous step as variables, design a simple discrimination system for each category. Assume the Poisson distribution for the frequencies, and pairwise independence between the variables (both are simplifications). Note that you will have to design six independent classifiers (one for each category).

13 (10 points) You can either implement the discrimination system that you designed or simulate it by hand. In any case, run or simulate your system on the articles in each test set, and construct a contingency table for the results for each of the six classifiers.

14 (10 points) For each of the six contingency tables above, compute recall, precision, and F-measure (with beta 1) as measured on the evaluation (test data). In other words, you will have to evaluate how well your classifiers work on the test data.

15 (10 points) Several other discrimination criteria were discussed in class, including word and sentence length, and n-gram character models. Thus far, you used a method based on words not necessarily because it is best but because it was simple to implement/simulate (it is actually quite good for the problem of topic identification). Which of these criteria would have been most appropriate for selecting discriminatory variables if the problem had been language identification? Author identification? Justify your position.

16 (10 points) Discuss other methods that may be applicable for author identification if the texts that we need to assign come in unedited form (i.e., not from a book but from email or a newsgroup).