Linguistics 384
Homework 3

Language Identification and Spam Filtering

DUE: Wednesday, October 19, 2005

1. (a) (20 pts) Using the Danish - Finnish Trigram Frequency Distribution table given below, determine the language of the following sentences.

   Danish-Finnish Trigram Frequency Distribution Table:
   http://www.ling.ohio-state.edu/~adriane/384/files/DF-trigrams.txt

   This is the same table we used in ICA 6, but it has been modified slightly since we used it in class. To make the spaces easier to see in the trigrams, all spaces were converted into underscores (_). For example, in (i), you will need to look up "en " as "en_", "n k" as "n_k", " kv" as "_kv", etc.

   Show all your work in order to receive full credit.

   i. Men kvällspressen ger sig inte
   ii. Raision kaavoitus
   iii. På verdensplan dør

   (b) (5 pts) (i) is actually in Finnish, although the trigram frequency distribution identifies it as Danish. What is the name for this case of incorrect identification of language? (Hint: this term is used in SPAM identification, too.)

   (c) Bonus: (5 pts) Give a possible reason why the trigram model fails to identify the correct language in the case of (i).

2. I use spamassassin to filter my email.

   (a) (10 pts) Send two messages to my email account:

   i. A message you think will be marked as spam.
   ii. The same message (content-wise), but now altered in such a way so as to trick my spam filter. Think about the methods spammers use to disguise content.

   Two things to note: 1) Keep your messages clean! 2) Your messages must be less than 100 words.

Bonus: (5 pts) If your first message is marked as spam but the second one is not, you will get 5 bonus points. (Warning: this is difficult to accomplish!)

(b) (15 pts) In the homework you hand in, I want a write-up for both messages. Explain:

- why you thought the first would be marked as spam
- why you thought the second would not be marked as spam

It may help to simply print your messages out and annotate them directly.

3. (20 pts) Go to http://spamassassin.apache.org/tests_3_0_x.html – you'll see a list of rules used by spamassassin, a rule-based filter. Pick three (3) rules that make sense to you. For each rule you pick, do the following:

   (a) Write down the "description of test"

   (b) Write down the part of the e-mail we are looking at (header, body, etc.)

   (c) Explain (in your own words) what this test is looking for in 1-2 sentences

   (d) Do you think this is a useful rule? Why or why not?

4. (15 pts) The word *mortgage* appears in 700 emails, 650 of which are spam, and 50 of which are ham. In total, there are 2000 messages, 1400 of which are spam, and 600 of which are ham. Calculate the probability that the message is spam if the word *mortgage* appears in it. Follow your notes from class carefully and show all your work.

5. (15 pts) Paul Graham says, "Statistical filters yield fewer false positives because they consider evidence of innocence as well as evidence of guilt." Based on the discussion in class of how statistical filters work (or on http://www.paulgraham.com/wfks.html), briefly describe in your own words how statistical filters consider evidence of innocence to avoid false positives.