Linguistics 384
Homework 3

Language Identification and Spam Filtering

DUE: February 1, 2006

- **Problem Solving** (to prepare for the quiz on February 1, 2006)

  1. (a) 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 5. Remeber that spaces have been converted into underscores (_) and that special characters are found after z in the alphabet. For example, in (i), you will need to look up "en " as "en_", "n k" as "n_k", " kv" as "_kv", etc.

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

     (b) (i) is actually in Finnish, although the trigram frequency distribution method identifies it as Danish. If we were trying to identify all Danish documents and this Finnish one was included in our list (incorrectly, of course), what is the name for this incorrect entry in our list? (Hint: this term is used in spam identification, too. It's the name for a real email message that's accidentally been marked as spam.)

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

     (d) How else could you identify the language of (iii)?

  2. 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 a new message is spam if the word *mortgage* appears in it. Show all your work.

- **Essay** (to hand in on February 1, 2006)

  1. I use spamassassin to filter my email.

     (a) 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 as to trick my spam filter. The information conveyed in the message needs to stay the same, but the way it's presented can be different. 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: If your first message is marked as spam but the second one is not, you will get 10 bonus points. (Warning: this is difficult to accomplish!)

     (b) In the homework you hand in, include 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.

  2. 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.