Linguistics 384
Homework 2

Searching

DUE: January 25, 2006

- **Problem Solving** (to prepare for the quiz on January 25, 2006)

  1. Boolean expressions

     Below is a list of 7 book titles:

     1) cats
     2) dogs
     3) birds
     4) how to make cats and dogs get along
     5) dogs hunt cats and birds
     6) dogs and birds
     7) cats who love birds

     For each of the following Boolean expressions, write down which book titles match the expression. e.g. the Boolean expression *dogs* matches 2, 4, 5, 6.

     (a) cats
     (b) cats AND birds
     (c) cats OR dogs OR birds
     (d) cats AND (dogs OR birds)
     (e) (cats AND dogs) OR birds

  2. Regular expressions

     (a) Write down the four (4) matches to the following regular expression: /(a)|(the) pigs?/
     (b) In addition to the four (4) from the previous example, which two (2) other strings does the following regular expression match? /((a)|(the))? pigs?/
     (c) You want to write a regular expression which matches the various spellings of *e-mail* and some of its derivatives. For this problem, you are not allowed to use the period (.) operator (which matches any single character). You'll write the regular expression in three steps:

i. First, write a regular expression which matches just the following two items: *e-mail, email*

ii. Now write a regular expression which includes the *s* ending: *e-mail, email, e-mails, emails*

iii. Of course, there are other possible endings, so let's also include *ing* (which can interact with *s*). Write a regular expression that matches all of the following items: *e-mail, email, e-mails, emails, e-mailing, emailing, e-mailings, emailings*

3. Evaluating search results

   You are searching for titles containing the word *chipmunk* in the library catalog. You write a search query that returns 50 matches. 40 of those matches are what you were looking for (they contain the word *chipmunk*) and 10 don't contain the word *chipmunk* at all (because your search query wasn't quite perfect). The librarian tells you that there are in fact 60 books in the catalog whose titles contain the word *chipmunk*.

   (a) What is the precision of your search?

   (b) What is the recall of your search?

4. Indexing

   You are given a set of documents and you need to create an index so that you can search the documents efficiently.

   The documents:

   1: I have three dogs. They like to play outdoors.
   2: My dog is playing catch with a frisbee outside.
   3: Dogs like catching frisbees.

   (a) Create an inverted index for these three documents. In this index, capitalization and word endings should count, so "dog", "dogs", and "Dogs" should each have their own entry in the index.

   (b) Which words in the index do you think should be considered stop words?

   (c) How would your index from (a) change if you ignored capitalization? Give a concrete example.

   (d) How would your index from (a) change if you used stemming? Give a concrete example.

   (e) How would your index from (a) change if you worried about synonyms? Give a concrete example.

- **Essay** (to hand in on January 25, 2006)

  1. Googlewhacks

     Go to `http://www.googlewhack.com` . This website lists pairs of words which generate exactly one – i.e. one and only one – result on google.com. Some previous examples (January 16, 2006) are *spoonable brickbat* and *birthname requiems*.

     (For each of the following, you may try as many times as you want, but you are only required to write up one response.)

     (a) Think of two unrelated words, and write them down.
        i. About how many hits do you expect to get with these words? (dozens? hundreds? thousands? tens of thousands? etc.) Why?
        ii. How many actual hits do you get at www.google.com? How were your words related?

        If you get zero hits, record that and try again with two less unrelated words.

     (b) Now pick one word. Write it down.
        i. About how many hits do you expect?
        ii. How many actual hits do you get?
        iii. Now carefully select a word which appears in one of the resulting webpage descriptions. What word did you pick? Enter it with your original word. How many actual hits do you get now?

     (c) You have just tried 2 different search strategies for finding a "google-whack". One required you to know exactly what you were looking for; the other required you to search and then narrow your search. Which worked better? In a sentence or two, say why you think this is the case for your example. If you wanted to find a single result using as many keywords as needed, which method is guaranteed to work?