Linguistics 384
Homework 2

Searching

DUE: Wednesday, October 12, 2005

1. (20 pts) **Boolean queries**

   Below is a list of 7 book titles:

   1) cats
   2) dogs
   3) birds
   4) how to make cats and dogs play nice
   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. (20 pts) **Searching**

   Your friend tells you the following:

   > Whenever I fall asleep watching TV, my back hurts when I wake up. I want to find sofas and easy chairs that are good for my back.

   Note: Be sure to write down for each step (except f) what you did (very briefly, only what is being asked for; in particular, do NOT enter any queries and report on their results until instructed to do so).

(a) Identify the words to be queried.

(b) Identify synonyms of those words.

(c) Decide which synonyms are best by determining which are least ambiguous; explain in one sentence why you made this decision.

(d) Decide which words need to be kept in the query, but might still be problematic; explain in one sentence why you think so.

(e) Formulate a query.

(f) Enter this query at:
http://www.altavista.com
(NOTE: The query language for altavista is described at: http://www.altavista.com/help/search
Be sure to capitalize AND and OR.)

(g) How many of the first 10 results were what you wanted?

(if none, formulate a different query in (e) until you get at least one intended result)

(h) How could you tell that these results were what you wanted?

(i) If these first 10 results are the only results, what is the precision?

(j) If there are 20 documents total that match your query, what is the recall given the number found in (g)?

3. (20 pts) **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 (October 4, 2005) are *slanting minitowers* and *creaseproof snaggletoothed*.

(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 "googlewhack". 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?

4. (20 pts) **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 likes to play catch outside.

3: Dogs like to sleep. Cats do not like to play catch.

(a) Create a simple inverted index for these three documents. In this index, capitalization and word endings should count, so "dog", "dogs", and "Dogs" should all have different entries.

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

(c) Think about the following queries. Which documents would be returned using the index you just made? For each query, which techniques could be used to improve the results?

i. dogs like

ii. dogs outside

iii. animals play catch

Some of the issues to consider are:

- stemming
- capitalization
- synonym checking
- word ambiguity

5. (20 pts) **Regular expressions**

   (a) Write down the four (4) matches to the following regular expression: /(a)|(the) dogs?/

   (b) In addition to the four (4) from the previous example, which two (2) other strings does the following regular expression match? /((a)|(the))? dogs?/

   (c) We're going to write a regular expression which matches the various spellings of *e-mail* and some of its derviatives, and we'll do this in pieces. For this exercise, you are not allowed to use the period (.) operator (which matches any single character).

      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*