

Homework 2: Searching

Linguistics 384 (Detmar Meurers)

Due at beginning of class on Tuesday, October 19, 2004

1. (25 points) Go to 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 (listed there as of October 13, 2004) are *blueish outstands* and *rastafarian supernatants*.

(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 web page 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 site using as many query words as needed, which method is guaranteed to work?
- (d) *Bonus question* (10 points extra): What other strategies might you use to find a googlewhack? Describe an example you tried.

2. (25 points) Your friend tells you the following:

When I fall asleep watching TV, I always wake up with pain in my lower back. 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.
 - (d) Decide which words need to be kept in the query, but might still be problematic.
 - (e) Formulate a boolean 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/syntax>
 - (g) How many of the first 10 results were what you wanted?
 - (h) How could you tell that these results were what you wanted?
3. (25 points) We're going to write a regular expression which matches the various spellings of *e-mail* and derived words and we'll do this step by step. For this exercise, you are not allowed to use the period (.) operator (which matches any single character).
- (a) First write a regular expression which matches just the following two items:
 - e-mail
 - email
 - (b) Now write a regular expression which includes the *s* ending:
 - e-mail
 - email
 - e-mails
 - emails

- (c) Of course, there are other possible endings, so let's also include *ing* (which can interact with *s*):

e-mail
email
e-mails
emails
e-mailing
emailing
e-mailings
emailings

4. (25 points) Write down the smallest regular expression you can come up with which finds any of the following words:

suncream
full-cream
ice-cream
scream
screams
screamed
screaming
cream
creams
creaming
creamed

Try it out at <http://logos.uio.no/cgi-bin/opus/opuscqp.pl?corpus=EUROPARL;lang=en> and note how many hits it finds (set “show max” to 1000 to do this).

5. *Bonus question* (20 points extra): Go to <http://logos.uio.no/cgi-bin/opus/opuscqp.pl?corpus=EUROPARL;lang=en> and define a regular expressions to query the corpus for words starting with “un” and ending with “ing”. Report the regular expression you used to search for this and the first five words it finds.