Homework 2: Searching

Scott Martin (Linguistics 384)

Due before class on Wednesday, October 17

(Submit homeworks as PDF, HTML, or plain text to "Homework 2" dropbox in Carmen.)

1. (25 points) Go to http://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 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.

      i. Which worked better?

      ii. In a sentence or two, say why you think this is the case for your example.

      iii. 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) 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

3. (10 points) What pattern matches any lowercase alphabetic string (ex: *a, this, pickles, supercalifragilisticexpialidocious*)?

4. (10 points) What pattern would match would match *theater* and *theatre*?

5. (10 points extra) What pattern would match would match a string of characters that start with an integer and which end a letter? (Ex: *1db, 23*mn*, etc.)

6. (10 points) What pattern would match the following forms of the verb *to be*: *be, being, been*?

7. *Bonus question* (25 points extra) 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).

8. (20 points) Below is a few of the many, many ways people have spelled “Britney Spears” in Google searching. (Source: http://www.google.com/jobs/britney.html) Note: The numbers are the counts of the number of times someone spelled Britney’s name that particular way. IGNORE them for your expression.

488941	britney spears
40134	brittany spears
36315	brittney spears
24342	britany spears
7331	britny spears
6633	briteny spears
2696	britteny spears
1635	brittny spears
1338	britiny spears
1096	britiney spears

Does the following pattern match all of the Britney spellings above? If not, how would you alter the pattern to fix this?

```
brit(a|e|i)?ny
```

9. (20 points extra) Go to http://www.lexmasterclass.com/exercises/regex/index.html and do exercise 2, which is repeated below. Note that you can try out the regular expressions you type in by clicking on the Submit button. The regular expression should match all of the items in the first column (i.e., all characters in the first column are completely red after clicking submit) and none of those in the second column (i.e., none of the items in the second column are red). Write down that expression.

Positive	**Negative**
	aleht
rap them	happy them
tapeth	tarpth
apth	Apt
wrap/try	peth
sap tray	tarreth
87ap9th	ddapdg
apothecary	apples
	shape the

10. (20 points extra) Go to http://www.lexmasterclass.com/exercises/regex/index.html and do exercise 3, which is repeated below. After figuring it out interactively with the website, write down the expression that matches all the words in the left column and none of those in the second column for your homework.

Positive	**Negative**
affgfking	fgok
rafgkahe	a fgk
bafghk	affgm
baffgkit	afffhk
affgfking	fgok
rafgkahe	afg.K
bafghk	aff gm
baffg kit	afffhgk

11. (20 points extra) Go to http://www.lexmasterclass.com/exercises/regex/index.html and do exercise 4, which is repeated below. After figuring it out interactively with the website, write down the expression that matches all the words in the left column and none of those in the second column for your homework.

Positive	**Negative**
assumes word senses. Within	in the U.S.A., people often
does the clustering. In the	John?”, he often thought, but
but when? It was hard to tell	weighed 17.5 grams
he arrive.” After she had	well ... they’d better not
mess! He did not let it	A.I. has long been a very
it wasn’t hers!’ She replied	like that”, he thought
always thought so.) Then	but W. G. Grace never had much

12. (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.