COMS W3261: Theoretical Computer Science.
Instructor: Tal Malkin

Problem Set 4

Due: Tue, 10/02/07.
Reading: Chapter 1.3.

1. [1] Download the file "http://www.cs.columbia.edu/~tal/3261/list.txt" to your CUNIX account. This file contains a list of English words up to length 12 acceptable for use in Scrabble.

   We are going to use the UNIX program "egrep" in this problem, which is available on CUNIX. For an egrep mini-tutorial, see:

   `http://www.cs.columbia.edu/~tal/3261/egrep_mini-tutorial.htm`

   You may also have to use UNIX piping to string together egrep commands. A "pipe" sends the output of one command to the input of another. For example, consider the command:

   ```
   egrep 'AA' list.txt | egrep 'C'
   ```

   This means "find all lines in the file list.txt that match the regular expression 'AA'; send those lines as input to the egrep command, which will then find all lines that match the regular expression 'C'. The net effect is that you would get all words that contain two A's in a row, AND contain a C.

   Multiple pipes can be strung together. For example,

   ```
   egrep 'AA' list.txt | egrep 'K'  | egrep 'PS'
   ```

   would output all the words containing two A's in a row, a K, and the substring PS.

   Also, you may need to use the ^ and $ to denote the beginning and end of a line. So, if you want to search for all words that begin with an A, you could say:

   ```
   egrep '^A' list.txt
   ```

   If you want all words of length at least two that begin with an A and end with a letter that is NOT an X or a Z, you could say:

   ```
   egrep '^A.*[^XZ]$' list.txt
   ```

   Answer the following questions; for each one, give the egrep command/commands that told you the answer. (It may be simpler to cut-and-paste your commands into a text file, and just print them out.)

   For the purposes of these questions, the letter Y is considered a vowel.

[1] Thanks to Jon Feldman for this great problem!

(a) **(4 pts)** What word contains two consecutive E's, and also contains two Y's ?

(b) **(4 pts)** What word contains no vowels and starts with an N ?

(c) **(4 pts)** Name a word with 6 consonants (non-vowels) in a row. (There's one with seven, too, but its quite unnatural.)

(d) **(4 pts)** What word contains two Z's, with three A's between them? For example, "**ZA**NIF**A**L**A**TINALI**Z**E" would be acceptable (unfortunately, it's not a real word).

(e) **(4 pts)** Name a word that contains all six vowels in order (not necessarily consecutively). For example, "**A**NTIST**E**REOLOG**I**C**OU**SL**Y**" would be acceptable (again, not a word).

(f) **(4 pts)** What word contains two disjoint substrings of length 3, each substring containing an E, an X, and a T ? (A *substring* is a string of letters that appear consecutively in the word. Two substrings are *disjoint* if they do not overlap in the word.)

(g) **(4 pts)** We define the following languages over $\Sigma = \{A, \ldots, Z\}$:

- Let $L_1 = \{w : w$ contains a T, and contains at least three vowels after the final T $\}$
- Let $L_2 = \{w : w$ does not contain two A's $\}$
- Let $L_3 = \{w : w$ contains an H and an R, with exactly two letters between them $\}$ (For example, SHEAR is is L_2)
- Let $L_4 = \{w : w$ contains an E and an Y, with exactly two letters between them $\}$ (For example, EASY is is L_2)

What word is in all four languages listed above?

(h) **(just for fun, not for credit)** Word-whacking: Find the regular expression with minimum length that matches exactly one word.

2. We now switch back to regular expression using the definition from class (and textbook). Recall that with this definition (unlike the egrep one), a string is in the language if the entire string can be generated with the regular expression – not just a substring match. For example, the language {all strings containing at least one 0} over $\{0, 1\}$, is generated by the regular expression $(0 \cup 1)^*0(0 \cup 1)^*$; the language {all strings starting with a 0 and ending with a 1} is generated by $0(0 \cup 1)^*1$.

   Give regular expressions generating each of the following languages.

   (a) **(4 pts)** $L = \{xy | x$ has an even number of 1s and y has an even number of 0s$\}$, over the alphabet $\{0, 1\}$.

   (b) **(4 pts)** $L = \{w | w$ does not contain two consecutive Cs$\}$ over the alphabet $\{A, B, C\}$.

3. **(6 pts)** Problem 1.21(b) in text (converting a DFA to a regular expression).