CS476: Automata Theory and Formal Languages

Homework 1

Due: 19/03/2012 17.00

Questions

1. (20pts) State whether the following statements are true or not. You must give a BRIEF explanation or show a counter example to receive full credit.

   (a) (5pts) A non-regular language can be finite.

   (b) (5pts) If a language is recognized by an 2^n-state DFA, then it must be recognized by some NFA with no more than n states.

   (c) (5pts) $(011+101+110)^*$ is the regular expression for the set of all strings with $2n_0(w) = n_1(w)$.

   (d) (5pts) To show that a language L is not regular by pumping lemma, all possible y strings within the first p characters of w sould be considered for a selected $w \in L$.

2. (20pts) Give a DFA for each of the following languages.

   (a) (10pts) $L = \{w \in \{a,b\}^* : (n_a(w) + 2n_b(w)) \bmod 3 < 2\}$.

   (b) (10pts) $L = \{w \in \{0,1,2\}^* : w$, when read as a ternary (base 3) number, is not a multiple of $4\}$.

3. (20pts) Give an NFA for each of the following languages.

   (a) (10pts) $L = \{0^m 1^n : m, n \geq 1, m+n$ is odd$\}$.

   (b) (10pts) $L = \{w \in \{0,1,2\}^* :$ The rightmost symbol of w is not equal to any other symbol of $w\}$.

4. (20pts) Give a regular expression for each of the following languages.

   (a) (10pts) $L = \{w \in \{0,1\}^* :$ There are at most two pair of consecutive 1's in $w\}$. (Note that $111 \in L$, but $1111 \notin L$.)

   (b) (10pts) $L = \{w \in \{0,1\}^* : w$, with the leading bit 1, when read as a binary number, is not between 10 and 30$\}$.

5. (20pts) Prove or disprove that the following languages are regular.

   (a) (10pts) $L = \{x\$y : x, y \in \{a,b\}^*, n_a(x) = n_b(y)\}$.

   (b) (10pts) $L = \{xy : x, y \in \{a,b\}^*, n_a(x) = n_b(y)\}$.

6. (20pts) The *distance* between two bit strings x and y (notation: $D(x,y)$) is the number of positions at which their symbols differ. For example, $D(011, 110) = 2$. If $|x| \neq |y|$, then $D(x,y) = \infty$. If x is a string and L is a set of strings, the distance between x and L is the distance from x to the closest string in L:

$$D(x, L) = min_{y \in L} D(x, y). \tag{1}$$

   For any set $L \subseteq \{0,1\}^*$ and $k \geq 0$, define

$$N_k(L) = \{x \mid D(x, L) \leq k\}, \tag{2}$$

   the set of strings of distance at most k from L. For example, $N_0(\{000, 001\}) = \{000, 001\}$, $N_1(\{000, 001\}) = \{000, 001, 010, 011, 100, 101\}$, and $N_2(\{000\}) = \{0,1\}^3 - \{111\}$.

   Prove that if $L \subseteq \{0,1\}^*$ is regular, then so is $N_2(L)$. (*Hint:* If L is accepted by a machine with states Q, build a machine for $N_2(L)$ with states $Q \times \{0,1,2\}$.)

7. **Perl:** (30pts) *Perl* is a language with a lot of scripting capabilities. It provides powerful text processing facilities. In this exercise, you will use the regular expression capabilities of *Perl*.

   (a) (10pts) In this part, you will write a script such that given a file the script displays some information about the strings in the file such that

      i. The number of strings that does not contain 011.
      ii. The number of strings that contains at least two 1s and at most three 0s.
      iii. The number of strings that starts with the two symbols that it ends with.
      iv. The number of strings that does not contain more than one occurence of the string 010. (The string 01010 should be viewed as containing two occurences of 010.)

   The alphabet is $\Sigma = \{0, 1\}$ hence the strings are binary strings. The strings can be separated by any kind of whitespaces, i.e., tab, space, newline etc.

   (b) (20pts) Assume you are a TA of a course and each student submits his/her homework by email with the following name format:

$$\{Surname\}\,\{Delimiter\}\,\{Name\}\,\{Delimiter\}\,\{ID\}\,.\,\{Extension\}$$

   Write a *Perl* script to recognize the Surname, Name, and ID of each student and output them in a tab-delimited format

$$\{Surname\}\,\backslash tab\,\{Name\}\,\backslash tab\,\{ID\}$$

   such that

      i. The delimiter between Surname, Name, and ID can be a space, an underscore, or a dot.
      ii. The file extension can be pdf, doc, or docx.
      iii. The student ID starts with 200 or 201 and consists of 8 digits.
      iv. Name may contain an optional middle name. Regardless of the delimiter between the first and middle names of the student, there should be a single space between them in the output format.
      v. Surname may contain an optional married name. To prevent the confusion of the married name with the middle name, a hypen should be used as the delimeter between surnames. There should be a single space between two surnames in the output format.
      vi. Allow Turkish characters in names. You should be able to recognize both lowercase and uppercase letters.
      vii. If you encounter a homework submission with a filename in a wrong format, output

$$\{Filename\}\,\backslash tab\; WRONG\; FORMAT\;!$$

      where Filename is the whole name of the file including the extension.

   SAMPLE INPUT:

   Müftüoğlu.Kuddusi 20081234.pdf
   ABİTOĞLU_Mustafa_Kamil_20101234.doc
   Öztürk-Yılmaz Ayşe 20091234.docx
   homework.pdf

   SAMPLE OUTPUT:

Müftüoğlu	Kuddusi	20081234
ABİTOĞLU	Mustafa Kamil	20101234
Öztürk Yılmaz	Ayşe	20091234
homework.pdf	WRONG FORMAT !	

Answers for questions 1-6 should be returned in hard copy. The answer (perl scripts) for question 7 should be returned in e-mail to `acer@cs.bilkent.edu.tr`, with the subject line cs476hw1, as an attachment zip file named *NameSurname.zip* including q7part1.pl and q7part2.pl. Good luck to all.