MATH 2301

* Regular expressions (and pattern-matching)

Let Σ be an alphabet

** Defn: A regular expression (or regex) r is a string in the letters of Σ, together with the symbols "$|$", "$*$", and "ϕ", satisfying one of the following:

1) $r = \phi$
2) $r = \varepsilon$
3) $r = a$ for some $a \in \Sigma$
4) $r = r_1 r_2$ for regexes r_1 and r_2 ⇝ concatenation
5) $r = r_1 | r_2$ for regexes r_1 and r_2 ⇝ "or"
6) $r = r_1^*$ for a regex r_1. ⇝ star

Additionally, we can use parentheses "(" & ")" to signify grouping.
(Just like in an algebraic expression)

** We assume that Σ does not contain $*, |, \phi, \varepsilon, (,)$.

** Order of operations

Brackets are subexpressions, so they come first.

Apply $*$ first, then concatenation, and then or.

Concatenation & $|$ are associative, so we don't need to bracket multiple concatenations or multiple "or"s

** Examples

$\Sigma = \{0, 1\}$

$r = \phi$, $r = \varepsilon$, $r = 0$, $r = 1$

$r = \phi^*$, $r = \varepsilon^*$

$r = 0^*$

$r = 01$, $r = \varepsilon 1$ $r = 1$

different, but they'll have the same meaning

$$r = \underbrace{(01 \mid \phi^* \mid 110)^*}_{r_1} \underbrace{010}_{r_2} \underbrace{(0 \mid 1 \mid \varepsilon)}_{r_3}$$

How to parse?

Mentally, break it up as: $r = r_1 r_2 r_3$

$r_1 = (\dots)^*$

$\underbrace{01}_{} \mid \underbrace{\phi^*}_{} \mid \underbrace{110}_{}$ an or of 3 subexpressions

(keep going)

$r_2 = 010$ = concatenation of $0, 1, 0$

$r_3 = 0 \mid 1 \mid \varepsilon$ = or of $0, 1, \varepsilon$.

** <u>Matching</u>

A word $w \in \Sigma^*$ is said to match a regex r if one or more of the following hold.

1) $r = \varepsilon$ and $w = \varepsilon$
2) $r = a$ and $w = a$, for some $a \in \Sigma$
3) $r = r_1 r_2$ and w can be written as $w = xy$, where x and y are words, and x matches r_1 and y matches r_2.
4) $r = r_1 | r_2$ and w either matches r_1, or matches r_2, or matches both.
5) $r = r_1^*$ and either $w = \varepsilon$, or $w = x_1 x_2 \cdots x_k$, where each x_i is a word, and each x_i matches r_1.

** <u>Examples</u>. Let $\Sigma = \{0, 1\}$

1) $r = 0$: $w = 0$ is the only string that matches
   $r = 1$: $w = 1$ is the only string that matches
   $r = \varepsilon$: $w = \varepsilon$ is the only string that matches.

2) $r = \phi$: no strings match this.

3) $r = 01$: $w = 01$ is the only string that matches

4) $r = \phi 1$: no strings match this

5) $r = 0 | 1$: $w = 0$, $w = 1$ match.

6) $r = 1^*$: $w = \varepsilon$, $w = 1$, $w = 11$, $w = 111111$ etc

match.

7) $r = (00|11)^*$

<u>Match</u>: $w = 00$, $w = 11$, $w = 000000$, $w = 111111$

$w = \varepsilon$, $w = 0011110011$

each of these match $(00|11)$

(and many others...)

8) $r = 0(1|0)^*1$

w starts with a 0 and ends with a 1.

these are exactly all the words that match.

$w = 0\varepsilon 1 = 01$

matches $(0|1)^*$

** The language of a regex

Let r be a regex. The language of r, denoted $L(r)$ is the set of all words that match r.

*** Example

$r = 0(1|0)^* 1$

$L(r) = \{ w \in \Sigma^* \mid w$ begins with a 0 and ends with a 1$\}$

** Question : Given some $L \subseteq \Sigma^*$, is there a regular expression r such that $L(r) = L$?

*** Example

$L \subseteq \Sigma^*$, $L = \{ w \mid w$ begins with a 0 or ends with a 1$\}$.

(Homework... finish next time)