MATH 2301 — 21 Sep 2022

* Today: Regular expressions

Fix an alphabet Σ

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

(1) $r = \phi$

(2) $r = \varepsilon$

(3) $r = a$ for some $a \in \Sigma$

(4) $r = r_1 r_2$ where r_1, r_2 are also regexes

(5) $r = r_1 \mid r_2$ where r_1, r_2 are regexes

(6) $r = r_1^*$

[In any of these options, ( ) signify grouping]

[(7) $r = (r_1)$ where r_1 is a regular expression]

Just like in an algebraic expressions.

** We assume that "|", "*", "ϕ", "(", ")" are not in Σ.

** Order of operations:

Brackets first, then *, then concatenation, then "|" (or)

* Examples. Let $\Sigma = \{0, 1\}$

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

$r = \phi^*$, $r = 0^*$, $r = \varepsilon^*$

$r = 0 \mid 1$, $r = 0 \mid 1 \mid 0^*$ # $[(0 \mid 1) \mid 0^*$ # $0 \mid (1 \mid 0^*)]$

(equivalent to $0 \mid 1 \mid 0^*$)

$r = (010 \mid 1)^* 00 \mid 1$

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

$r = r_1^* r_2$

$r_1 = (\underbrace{01}_{r_3} \mid \underbrace{\phi^*}_{r_4} \mid \underbrace{110}_{r_5}) = r_3 \mid r_4^* \mid r_5$

[Continue breaking up the expression mentally until you hit either ϕ, ε, or a letter.]

** Matching

Let r be a regex. Let $w \in \Sigma^*$ be a string.
We say that w matches r if one or more of the following hold:

(1) $r = \varepsilon$ and $w = \varepsilon$

(2) $r = a$ for some $a \in \Sigma$, and $w = a$

(3) $r = r_1 r_2$ for regexes $r_1 r_2$, and $w = xy$, where x, y are strings, and x matches r_1, and y matches r_2.

(4) $r = r_1 | r_2$ and w either matches r_1 or r_2 (or both)

(5) $r = r_1^*$, and either $w = \varepsilon$ or $w = x_1 x_2 \cdots x_k$ where each x_i is a string, and each x_i matches r_1

* No string matches $r = \phi$.

** Examples

. $r = 0$: $w = 0$ only string that matches
$r = 1$: $w = 1$ " " " "
$r = \varepsilon$: $w = \varepsilon$ " " " "

. $r = 010$: $w = 010$ only match

. $r = 1\phi$: nothing matches!

. $r = 0 | 1$: $w = 0$, $w = 1$ only matches

. $r = 1^*$: $w = \varepsilon$, $w = 1$, $w = 11$, $w = 111$, etc

. $r = (01)^*$: ε, 01, 0101, 010101, ...

. $r = (00|11)^*$: ε, 00, 11, 0000, 1111, 0011, 1100, ...

$w = 00|0$ does not match.

~~$r = 01^* | 10$~~

. $r = 01^* | 0^*1$ $w = 001$, 011, 01 ✓
$w = 0101$ not a match
$w = 1$ ✓

** The language of a regex

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

E.g. $r = 0$, $L(r) = \{0\}$
$r = \phi$, $L(r) = \phi$
$r = \varepsilon$, $L(r) = \{\varepsilon\}$
$r = 0(1|0)^*1$, $L(r) =$ strings starting with 0 and ending with 1.