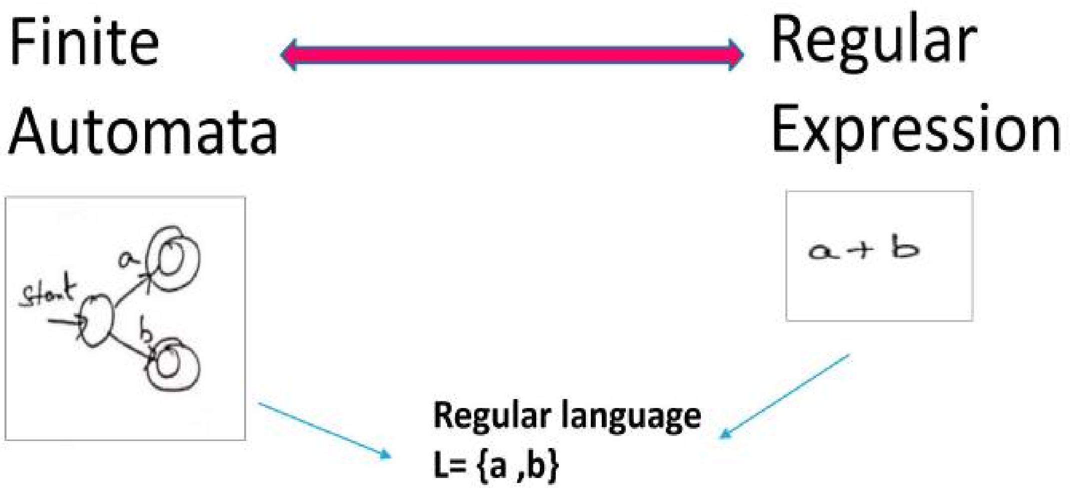
Regular Expression

Regular Language and Expression

- The language accepted by DFA, NFA and ЄNFA is called Regular Language.
- A regular language can be described using regular expression.
- Regular Expression consists of symbols such as alphabet Σ, operators '.', '+', '*'
- Operators used to obtain regular expression:
  - + operator used for union operation (least precedence)
  - . Operator used for concatenation (next least precedence)
  - *(kleene closure), +(positive closure) operator is used for closure operation (highest precedence)

Regular Language, Regular Expression and Finite Automata

- Regular Expressions are useful tools for defining patterns
- The language defined by RE are Regular Language
- Any language defined by RE is accepted by Finite Automata
- Any language accepted by FA can be define by some RE
- RE is generator that generate Regular Language
- FA is acceptor that accepts Regular Language generated by RE



Regular Expression

A regular expression can be formally define as follows

- ϕ is RE denoting empty language
- ϵ (epsilon) is RE denoted the language containing empty string
- a is RE indicating the language containing {a}
- If R is RE denoting the language L_R and S is RE denoting the language L_S then
  - R+S is a RE corresponding to the language L_R U L_S
  - R.S is a RE corresponding to the language L_R . L_S
  - R* is a RE corresponding to the language L_R^*
  - $R^+$is a RE corresponding to the language L_R^+

Kleene Closure L* (zero or more number of concatenation of string in L)

- $L^* = \bigcup_{i=0}^{\infty} L^i$
- $L^* = L^0 \cup L^1 \cup L^2 \cup L^3$ ....
- L^0 = {ϵ}
- L^1 = L
- L^2 = L.L
- $L^3 = L^2$. L = L.L.L

- Ex – a* = $a^0 \cup a^1 \cup a^2 \cup a^3$ … … . .

  = {ϵ,a,aa,aaa,aaaa,...........}

Positive Closure L^+ (one or more number of concatenation of string in L)

- $L^+ = \bigcup_{i=1}^{\infty} L^i$
- $L^+ = L^1 \cup L^2 \cup L^3$ ………. $L^* = \{\epsilon\} \cup L^+$
- L^1 = L
- L^2 = L.L
- $L^3 = L^2$. L = L.L.L

- Ex – $a^+ = a^1 \cup a^2 \cup a^3 \; \dots \dots ..$

  = {a,aa,aaa,aaaa,……….}

R.E	Meaning
a*	String consisting of any number of a's (including ϵ) (zero or more)
a^+	String consisting of any number of a's (one or more)
(a+b)	String consisting of either a or b
a.b	String consisting of ab
(a+b)*	String consisting of any combination of a & b including null
(a+b)*abb	String of a and b ending with abb
ab(a+b)*	String of a and b starting ab
(a+b)*aa(a+b)*	String of a and b having substring aa
a*b*c*	Any number of a followed by any number of b followed any number of c
$a^+b^+c^+$	At least one a followed by at least one b followed by at least one c
aa*bb*cc*	At least one a followed by at least one b followed by at least one c a.a* = a. {ϵ,a,aa,aaa,aaaa...} = {a,aa,aaa,aaaa,aaaaa}

R.E	Meaning
(aa)*(bb)*b	Even number of a followed by odd number of b
(11)*	Even number of 1
01*+1	Either 1 or starting with 0 followed by any number of 1
(01)*+1	Either 1 or any number of 01
0 (1* + 1)	Zero followed by any number of 1
(a+b).c	String length of 2, first symbol is either a or b followed by c
(a+b)(a+b)	String of a and b whose length is 2
(1+00)*	Any combination of 1 and 00
a (a+b)* b	Start with a and end with b
(a+b)*(a+bb)	String of a and b ending with either a or bb

- Obtain RE representing string of a and b having length 2

  String – {aa,ab,ba,bb}

  R.E – (a+b)(a+b)

- Obtain RE representing string of a and b having length ≤ 2

  String – {ϵ,a,b,ab,ba,aa,bb}

  RE – (ϵ+a+b) (ϵ+a+b)

  L = $\{(a + b)^n \ n \leq 2\}$

- Obtain RE representing string of a and b having even length

  String = {ϵ,aa,bb,ab,ba,aaaa,bbbb,abab,baba........}

  RE =((a+b)(a+b))*

- Obtain RE representing string of a and b having odd length

  String = {a,b,aaa,aba,bbb,bab,abb,baa..........}

  RE – ((a+b)(a+b))* (a+b)   or   (a+b) ((a+b)(a+b))*

- Obtain a RE representing a language consisting of strings of a and b with alternate a and b or no two consecutive same letter

  String = {ϵ,a,b,ab,ba,aba,bab,abab,baba,........}

  RE = $(\epsilon+b)(ab)^*(\epsilon+a)$

  or

  $(\epsilon+a)(ba)^*(\epsilon+b)$

- Obtain RE representing a language consisting of strings of 0 and 1 with at most one pair of consecutive 0.
- CASE 1: at most zero 0

  RE1 - 1*
- CASE 2: at most one 0

  RE2 – (1+01)* (ϵ+0)
- CASE 3: at most two 0

  RE3 - (1+01)*00 (1+10)*

- CASE 1 + CASE 2+ CASE 3 (RE1+RE2+RE3)

= 1* + (1+01)* (ϵ+0) + (1+01)*00 (1+10)*

- Obtain RE representing a language containing at least one a and at least one b where $\Sigma = \{a,b,c\}$

$$RE - (a+b+c)^* \, a \, (a+b+c)^* \, b \, (a+b+c)^*$$

$$+$$

$$(a+b+c)^* \, b \, (a+b+c)^* \, a \, (a+b+c)^*$$

- Obtain RE for the language $L = \{w \mid w \in \{0,1\}^*$ with at least three consecutive $0\}$

$$RE - (0+1)^* \, 000 \, (0+1)^*$$

- Obtain RE representing string of a and b ending with b and has no substring aa

  String = {b,ab,bab,abb,abab……}

  RE – (b + ab)*(b+ab) = $(b + ab)^{+}$

- Obtain RE representing string of 0 and 1 having no two consecutive 0

  RE – (1+01)* (ϵ+0)

- Obtain RE representing string of a and b containing not more than three a

  RE – $b^*(\epsilon + a)\, b^*(\epsilon + a) b^*\, (\epsilon + a) b^*$

- Obtain RE for the set of all string that do not end with 01 over $\{0,1\}^*$

  RE – $\epsilon + (0+1) + (0+1)^*(00+11+10)$

  OR

  $\epsilon + (0+1) + (0+1)^*(0+11)$

- Obtain RE representing string of a and b whose second symbol from right is a

  RE – $(a+b)^*a(a+b)$

- Obtain RE representing string of a and b whose tenth symbol from right is a

  RE – $(a+b)^*a(a+b)^9$

- Obtain RE representing string of a and b whose second symbol from left is b and second symbol from right is a

  RE – $(a+b)b(a+b)^*a(a+b)$

- Obtain RE representing the words with two or more letters but beginning and ending with same letter where $\Sigma = \{a,b\}$

    RE – a(a+b)*a + b(a+b)*b

- Obtain RE for $L = \{vuv \mid u,v \in \{a,b\}^* \text{ and } |v|=2\}$

    RE - aa(a+b)*aa + ab(a+b)*ab + bb(a+b)*bb + ba(a+b)*ba

- Obtain RE representing string of a and b whose length is either even or multiple of 3 or both

Case 1: even length - ((a+b)(a+b))*

Case 2: multiple of 3 - ((a+b)(a+b)(a+b))*

((a+b)(a+b))* + ((a+b)(a+b)(a+b))*

- Obtain RE representing string of a and b such that every block of four consecutive symbols contains at least two a

  aa(a+b)(a+b)

  a(a+b)a(a+b)

  a(a+b)(a+b)a

  (a+b)aa(a+b)

  (a+b)a(a+b)a

  (a+b)(a+b)aa

RE = (aa(a+b)(a+b) + a(a+b)a(a+b) + a(a+b)(a+b)a + (a+b)aa(a+b) + (a+b)a(a+b)a + (a+b)(a+b)aa)+

- Obtain RE for the language L = $\{a^n b^m \mid n + m \text{ is even}\}$

  RE = (aa)*(bb)* + a(aa)*b(bb)*

- Obtain RE for the language L = $\{a^n b^m \mid n \geq 1, m \geq 1, nm \geq 3\}$

  n=1, m>=3 a bbb b*

  n>=3, m=1 aaa a* b

  n>=2, m>=2 aa a* bb b*

  RE = a bbb b* + aaa a* b + aa a* bb b*

- Obtain RE for the language L = $\{a^{2n} b^{2m} \mid n \geq 0\ m \geq 0\}$

  RE = (aa)*(bb)*

- Obtain RE for the language L = {w | |w| mod 3 =0 and w ϵ {0,1}* }

  RE – ((0+1) (0+1) (0+1))*

- Obtain RE for the language L = {w | n_a(w)mod 3 =0 and w ϵ {a,b}* }

  RE – (b*ab*ab*ab*)*

- Are (ab)*a and a(ba)* equal?

  (ab)*a = {a,aba,ababa,abababa,ababababa......}

  a(ba)*= {a,aba,ababa, abababa, ababababa,.....}

Prove that (1+00*1)+(1+00*1)(0+10*1)*(0+10*1) = 0*1(0+10*1)*

LHS = (1+00*1)+(1+00*1)(0+10*1)*(0+10*1)

= (1+00*1) [ϵ + (0+10*1)*(0+10*1)]

= (1+00*1) (0+10*1)*

= (ϵ +00*)1 (0+10*1)*

= 0* 1 (0+10*1)*

= RHS

Regular Expression (RE)

→ Any terminal symbol (i.e. an element of Σ), ϵ, ϕ are regular expressions.

→ The union of two REs, R_1 and R_2, written as R_1+R_2 is also a RE.

→ The concatenation of two REs, R_1 and R_2, written as R_1R_2 is also a RE.

→ The iteration (closure) of a regular expression, R, written as R^* is also a RE.

→ If R is a RE over Σ, (R) is also a RE.

→ The RE over Σ are precisely those obtained recursively by the application of above five rules.

Definition

Any set represented by a RE is called as a regular set.

Examples : Consider $\Sigma = \{a, b\}$

(i) a denotes the set $\{a\}$

(ii) $a+b$ denotes the set $\{a, b\}$

(iii) ab represents the set $\{ab\}$

(iv) a^* denotes the set $\{\epsilon, a, aa, aaa, \ldots\}$

(v) $(a+b)^*$ denotes the set $\{\epsilon, a, b, ab, ba, aab, \ldots\}$

↳ All the strings over $\Sigma = \{a, b\}$.

Q Describe the following sets by REs.

set		RE
(a) $\{101\}$	$\Sigma = \{0,1\}$	101
(b) $\{abba\}$	$\Sigma = \{a,b\}$	$abba$
(c) $\{01, 10\}$	$\Sigma = \{0,1\}$	$01+10$
(d) $\{\epsilon, ab\}$	$\Sigma = \{a,b\}$	$\epsilon + ab$
(e) $\{abb, a, b, bba\}$	$\Sigma = \{a,b\}$	$abb+a+b+bba$
(f) $\{\epsilon, 0, 00, 000, 0000, \ldots\}$	$\Sigma = \{0,1\}$	0^*
(g) $\{1, 11, 111, 1111, \ldots\}$	$\Sigma = \{0,1\}$	11^*

Q// Describe the following sets by RE

(a) L_1 = Set of all binary strings ending in 00. ($\Sigma = \{0,1\}$)

$L_1 = \{00, 000, 100, 0000, 0100, 1000, 1100, \ldots\}$

$$\boxed{(0+1)^* 00}$$

(b) L_2 = Set of all the binary strings beginning with 1 and ending with 0.

$L_2 = \{10, 100, 110, 1000, 1010, 1100, 1110, \ldots\}$

$$\boxed{1(0+1)^* 0}$$

(c) $L_3 = \{\epsilon, 11, 1111, 111111, \ldots\}$

$$\boxed{(11)^*}$$

$$a(a+b)^* b$$

$$+$$

$$b(a+b)^* a$$

Starting and ending letters are different

$$\boxed{a(a+b)^* b + b(a+b)^* a}$$

Identities for RE

I_1: $\phi + R = R$

I_2: $\phi R = R\phi = \phi$

I_3: $\epsilon R = R\epsilon = R$

I_4: $\epsilon^* = \epsilon$ and $\phi^* = \epsilon$

I_5: $R + R = R$

I_6: $R^* R^* = R^*$

I_7: $RR^* = R^* R$

I_8: $(R^*)^* = R^*$

I_9: $\epsilon + RR^* = R^* = \epsilon + R^* R$

I_{10}: $(PQ)^* P = P(QP)^*$

I_{11}: $(P+Q)^* = (P^* Q^*)^* = (P^* + Q^*)^*$

I_{12}: $(P+Q)R = PR + QR$ and

$R(P+Q) = RP + RQ$

Q1(a) Give an RE for representing the set L of strings in which every 0 is immediately followed by atleast two 1's.

$L = \{\epsilon, 1, 11, 111, 011, 0111, 1111, 1011 \ ... \}$

$$\boxed{(1+011)^*}$$

(b) Prove that the RE, $R = \epsilon + 1^*(011)^*(1^*(011)^*)^*$ also describes the set $(1+011)^*$

Q1 Prove that

$$(1+00^*1)+(1+00^*1)(0+10^*1)^*(0+10^*1) = 0^*1(0+10^*1)^*$$

A

$$\underline{(1+00^*1)} + \underline{(1+00^*1)}\,\underline{(0+10^*1)^*(0+10^*1)}$$

$$= (1+00^*1)\,(\epsilon+\underbrace{(0+10^*1)^*}_{R^*}\underbrace{(0+10^*1)}_{R})$$

$$= (1+00^*1)\,(0+10^*1)^*$$

$$= (\epsilon+00^*)1\,(0+10^*1)^*$$

$$= 0^*1(0+10^*1)^* \qquad \text{RHS} \qquad \text{(Proved)}$$

Q// RE = $(aa)^*(bb)^*b$ $\Sigma = \{a, b\}$

L = Set of all the strings over $\Sigma = \{a, b\}$ with even number of a's followed by odd number of b's.

$$L(RE) = \{ a^{2n} b^{2m+1} \mid n, m \geq 0 \}$$

Q// Give a RE for the following language:

$$L(R) = \{ w \in \Sigma^* \mid w \text{ has atleast one pair of consecutive zeros} \}$$

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

$$\boxed{(0+1)^* 00 (0+1)^*}$$

Q) Give a RE for the following language.

$L = \{ w \in \{0,1\}^* :$ w has no pair of consecutive zero's $\}$

$L = \{ \epsilon, 0, 1, 10, 11, 01, 010, 011, 101, 110, \dots \}$

$(1+01)^* (0+\epsilon)$ $\quad\quad\quad\quad\quad$ $(01)^*$

$\uparrow$

$(1+01)^* + (1+01)^* 0$

2(b) Even length binary string where odd position contains 1.

$L = \{ \epsilon, 10, 11, 1010, 1011, 1110, 1111,$

$$\left(1(0+1)\right)^*$$

$$(10+11)^*$$

~~Q// Find a RE over $\Sigma = \{a, b\}$ for~~

Q// Find a RE for all the strings over $\Sigma = \{a, b\}$ that do not end with aa.

$L = \{\epsilon, a, b, ab, ba, bb, \ldots\}$

$$(a+b)^* ab + (a+b)^* ba + (a+b)^* bb + a + b + \epsilon$$

$$= \boxed{(a+b)^*(ab+ba+bb) + a + b + \epsilon}$$

Q1) Find a RE for all the strings over $\Sigma = \{a, b\}$ that contains at most one occurrence of aa.

b   b

ab   aa

one time $(b + ab)^* aa (b + ba)^*$

$+$

zero time. $(b+ab)^* \epsilon (b+ba)^*$

$+$

only a $(b+ab)^* a (b+ba)^*$

$$(b+ab)^*(aa + a + \epsilon)(b+ba)^*$$

Q// Find a RE ~~[illegible]~~ that represents all the even length binary strings. (Here $\Sigma = \{a,b\}$).

✓ $((0+1)(0+1))^*$

$((0+1)^2)^*$

✓ $(00+01+10+11)^*$

([illegible] $+011)^*$

Q) Find an RE that will represent all the even length strings over $\Sigma = \{a, b\}$, where number of a's and b's are ~~equal~~. also even.

Ans: $L = \{ \epsilon, ab, ba, aabb, bbaa, abba, baab, abab, baba, \ldots \}$

$$\boxed{(aa+bb)^* \left((aa+bb)^* (ab+ba) (aa+bb)^* (ab+ba) (aa+bb)^* \right)^*}$$

$(aa+bb)^*$ ϵ, aa, bb, aaaa, aabb, bbaa, bbbb, aaaaaa, aabbaa.

abbaabab . abab, baba baab

$(aa+bb)^*$ × (ab+ba) × (ab+ba) × × (ab+ba) × (ab+ba) ×

Q// Find an RE for the following. Here $\Sigma = \{a, b\}$.

(i) $a^m b^n \mid m+n$ is even.

case 1: Both m and n are even

$(aa)^*(bb)^*$

case 2: Both m and n are odd

$a(aa)^* b(bb)^*$

$(aa)^*(bb)^* + a(aa)^* b(bb)^*$

(ii) $a^m b^n \mid m+n$ is odd.

Case I: m is even and n is odd

$(aa)^* b(bb)^*$

case 2: m is odd and n is even

$a(aa)^*(bb)^*$

$(aa)^* b(bb)^* + a(aa)^*(bb)^*$

Q// Find RE for the following over $\Sigma = \{a, b\}$

$L = \{ a^n b^m \mid n \geq 2 \ \& \ m \leq 3 \}$

$$\boxed{aaa^*(\epsilon + b + bb + bbb)}$$

Q/ The set of binary strings whose sixth symbol from right end is 1.

$$\boxed{(0+1)^* \, \underline{1} \; \underline{(0+1)} \; \underline{(0+1)} \; \underline{(0+1)} \; \underline{(0+1)} \; \underline{(0+1)}}$$

$$\downarrow$$

$$\boxed{(0+1)^* 1 \, (0+1)^5}$$

Q/) $\Sigma = \{a, b\}$. Find RE for all the strings that contain exactly two a's.

$$\boxed{b^* a b^* a b^*}$$

Q/) Find RE for at most two a's over $\Sigma = \{a, b\}$.

$$b^* + b^* a b^* + b^* a b^* a b^*$$

Q// Find the shortest string that is not in the language represented by the following RE.

$$a^*(ab)^*b^*$$

Ans $\boxed{ba}$.

Application of Regular Expression

- Automata = finite state machines (or extensions thereof) used in many disciplines
- Efficient string searching
- For recognizing the pattern using regular expressions.
- Pattern matching with regular expressions (example: Unix grep utility)
- Lexical analysis (a.k.a. scanning, tokenizing) in a compiler (the topic of a lecture later in the course)