Regular Expression

Course Name: Compiler Design
Course Code: CSE331
Level:3, Term:3
Department of Computer Science and Engineering
Daffodil International University

Regular Expression

A Regular expression (sometimes called a rational expression) is a sequence of characters that define a search pattern, mainly for use in pattern matching with strings, or string matching, i.e. “find and replace”-like operations.

Regular expression is a notation for defining the set of tokens that normally occur in programming languages.

How are patterns specified?

- Define following operators over sets of strings:
  - 1. Union: L ∪ U
    - S = L ∪ U = {s | (s ∈ L) ∨ (s ∈ U)}
  - 2. Concatenation: LU or L.U
    - S = L . U = {s t | (s ∈ L) ∧ (t ∈ U)}
  - 3. Kleene closure: L∗, set of all strings of letters, including ∈,
    - S = L∗ denotes “zero or more concatenations of” L
  - 4. Positive closure: L+.
    - S = L+ denotes “one or more concatenation of” L

OPERATION	DEFINITION AND NOTATION
Union of L and M	$L \cup M = \{s \mid s \text{ is in } L \text{ or } s \text{ is in } M\}$
Concatenation of L and M	$LM = \{st \mid s \text{ is in } L \text{ and } t \text{ is in } M\}$
Kleene closure of L	$L^* = \cup_{i=0}^{\infty} L^i$
Positive closure of L	$L^+ = \cup_{i=1}^{\infty} L^i$

Operations of Language

- Letters or alphabets and digits are the most important elements of language.
- Let L be the set of alphabets {A, B,…Z, a, b,….z} and D be the set of digits {0, 1, ….,9}
- L could be in form of upper case and lower case.
- Examples:
  - L U D is the set of letters and digits.
  - LD is the set of strings consisting of a letter followed by a digit.
- LLLL = L^4 is the set of all four-letter strings.

Operations of Language

- L^* is the set of all strings of letters, including ε, the empty string
- $L(L \cup D)^*$ is the set of all strings of letters and digits beginning with a letter.
- D^* is the set of all strings of one or more digits.

Examples

- Let L = {a, b}
- Some regular expressions:
- a | b
  - Denotes the set of {a, b} having a or b.
- (a|b)(a|b)
  - Denotes {aa, ab, ba, bb}, the set of all strings of a's and b's of length two.
- a^*
  - Denotes the set of all strings of zero or more a's , i. e., {ε, a, aa, aaa, …..}

Examples

- $(a|b)^*$ or $(a^*|b^*)^*$
- Denotes the set of all strings containing zero or more instances of an a or b, that is, the set of all strings of a's and b's.

- $a \mid a^*b$
- Denotes the set containing the string a and all strings consisting of zero or more a's followed by a b

Language to Regular Expressions

Examples:

- "Set of all strings having at least one ab"
  - $(ab)^+$

- "Set of all strings having even number of aa"
  - $(aa)^*$

- "Set of all strings having odd number of bb"
  - $b(bb)^*$

- "Set of all strings having even number of aa and even number of bb"
  - $(aa)^* (bb)^*$

Language to Regular Expressions

- “Set of all strings having zero or more instances of a or b starting with aa”
  - (aa)(a | b)*
- “Set of all strings having zero or more instances of a or b ending with bb”
  - (a | b)* (bb)
- “Set of all strings having zero or more instances of a or b starting with aa and ending with bb”
  - (aa) (a | b)* (bb)

THANK YOU