Regular Expression: Regular expressions are used for representing certain sets of strings in an algebric notation.

Example:

1) Algebraic (ξ) and $\wedge, \phi$
2) Union of two RE. $(R_1 + R_2)$
3) Concatenation of two RE. $(R_1 \cdot R_2)$
4) Iteration or closure of two RE. $R \to R^*$

$a^* = \wedge, a, aa, aaa \ldots$

5) applying the rules above multiple times.

N.B:

① When empty symbol is present,

$1^* = \{\wedge, 1, 11 \ldots\}$

② when empty symbol is not present

$1^+ = \{1, 11, 111 \ldots\}$ → Kleene plus one or more occurences

Kleene* (cleany) star: "zero or more occurences of the immediately previous character / RE"

sets as RE:

Ex: 1) {0, 1, 2} R = 0 + 1 + 2

2) {ε, ab} R = ε ab

3) {abb, a, b, bba} R = abb + a + b + bba

* Characteristics of RE:

① Case sensitive: /s/ ≠ /S/

② Delimited by slashes (/)

/woodchucks/ ; /a/ , /!/

③ [ ] means disjunction of characters.

/ [pP]romi - promi or Promi

/ [abc] - 'a', 'b' or 'c'

④ dash (-) means range of a set of chars.

/ [A-Z] - an upper case letter

⑤ Canet (^)

a) 1st symbol after [— negation

Ex: /[^A-Z]/ - not an upper-case

/[^Ss]/ - neither S nor s.

b) otherwise — simple chanacten

/[a^b]/ — 'a' on '^' on 'b'

/a^b/ — a^b

c) stant of a line - /^The/ — The dog ✓ — in The ✗

⑥ question mank /?/

previous chanacten — 0 on 1 time

ex: /colou?r/ — colon on coloun

/woodchucks?/ — woodchuck on woodchucks

N.B: /[0-9][0-9]*]

— at least one digit

can also be written as: /[0-9]+/

N.B: caret (^), dollar ($), \b and \B are known as anchors.

⑦ The period (/./)

- matches any single character

/beg.n/ — begin on beg'n on begun....

⑧ .* — any string of chars

/cutu.*cutu/ — cutu appears twice must.

⑨ dollar ($) — matches end of line

- /dog$/
  - a black dog ✓
  - dog banks ✗
  - a dog is banking ✗

⑩ /b and /B ~ non-boundary

/\B the \B/

- other
- father and mother

↓

word-boundary

the

/\b the \b/

- the apple is red (space আছে)
- father ✗ (space নেই)

এখানে the এর আগে ও পরে অবশ্যই অন্য character (not space) থাকা লাগবে

Ex:
- the apple ✗ (পরে space)
- the ✗ (আগে পরে নাই)
- then ✗ (আগে নাই)
- mthe ✗ (পরে নাই)

* in case of digit

/\b 99 \b/ matches if 99 is not with any digit, underscore or letter

Ex:

999x a99x

$99✓ -99✓ _99x

(11) Pipe symbol (|) or disjuction operator:

* /cat|dog/ matches

cat or dog

(12) parenthesis operator ( )

/gupp(y|ies)

- guppy or guppies

(13) Aliases for common sets of chars

\d - [0-9] any digit

\D - [^0-9] not digit

\w - [a-z A-Z 0-9_] num, dig, underscore

\W - [^\w]   \w এ যা আছে ঐগুলো ছাড়া

\s - [␣\r\t\n\f]   whitespace (space, newline, tab)

\S - [^\s] - not space

(14) Counting operators : (যে char, or expression কে নির্দেশ করে, always ঐ char এর পরে বসবে)

* - 0 or more occurences
\+ - 1 or more
? - exactly 0 or 1

] previously discussed

{n} - n occurences
{n, m} - n to m "
{n, } - at least n occurences
{, m} - maximum m "

N.B: sequences – string

(15) Backslashed chars

* – star sign
\. – full stop
\? – ques mark
\n – newline
\t – tab

Operator precedence:

counters > sequences > disjunction

Ex: a) /the*/ – theeee ✓
– thethe ✗

b) /the|any/ – the on any ✓
– thany ✗
– theny ✗

For self-understanding, visit: regex101.com