Regular expressions

Regular expressions in text processing

Regular expressions are powerful tools for pattern matching and text processing used in various editors and filters, such as:

- Vi/Vim
- Emacs
- grep and egrep
- sed
- awk
- perl

Regular expressions in text processing

- *: Zero or more occurrences of the previous character.
  - go*l: Matches "gl", "gol", "gool", "gooool", and so on.
  - 12*34: Matches "1234", "12334", "123334", and so on.
- g*: Nothing or "g," "gg," "ggg," etc.
  - go*g: Matches "gg", "gog", "goog", "gooooog", and so on.
  - 12*g: Matches "12g", "12gg", "12ggg", and so on.
- .: A single character.
  - c.t: Matches "cat", "cet", "cft", but not "ct" or "catt".
  - a.b: Matches "abb", "acb", "aab", but not "ab" or "aabb".

Regular expressions in text processing

- .*: Nothing or any number of characters.
  - a.*b: Matches "ab", "aabb", "aXb", "aYb", "aZb", and so on.
  - go.*d: Matches "god", "good", "goooooooood", "goXd", "goYd", and so on.
- [abc]: Matches "a," "b," or "c."
  - [aeiou]: Matches any lowercase vowel.
  - [123]: Matches "1," "2," or "3."
- [1-3]: Matches a digit between 1 and 3.
  - [4-6]: Matches "4," "5," or "6."
  - [0-9]: Matches any digit.

Regular expressions in text processing

- [^Z]: Any character except "Z."
  - [^aeiou]: Matches any consonant.
  - [^0-9]: Matches any non-digit character.
- [^a-zA-Z]: A non-alphabetic character.
  - [^a-zA-Z0-9]: Matches any non-alphanumeric character.
  - [^xyz]: Matches any character except "x," "y," or "z."

Regular expressions in text processing

- ^DM: "DM" at the beginning of a line.
  - ^Start: Matches "Start" at the beginning of a line.
  - ^XYZ: Matches "XYZ" at the beginning of a line.
- sun$: "sun" at the end of a line.
  - run$: Matches "run" at the end of a line.
  - abc$: Matches "abc" at the end of a line.

Regular expressions in text processing

- g\+: One or more occurrences of the previous character.
  - go\+d: Matches "god", "good", "goooood", and so on.
  - a\+: Matches "a", "aa", "aaa", and so on.
- g\?: Zero or one occurrence of the previous character.
  - colou\?r: Matches "color" and "colour".
  - x\?y: Matches "y" and "xy".
- GIF\|JPEG: As above - GIF or JPEG.
  - The image format can be GIF\|JPEG: Matches "GIF" or "JPEG".
  - The file extension is jpg\|jpeg: Matches "jpg" or "jpeg".

Regular expressions in text processing

- wood\(cock\|house\): As above - woodcock or woodhouse.
  - The bird is wood\(cock\|house\): Matches "woodcock" or "woodhouse".
  - The furniture is made of wood\(cock\|house\): Matches "woodcock" or "woodhouse".
- \<pat: Pattern "pat" at the beginning of a word.
  - \<cat: Matches "catch", "caterpillar", but not "catapult".
  - \<in: Matches "in", "inside", but not "tin".
- pat\>: Pattern "pat" at the end of a word.
  - cat\>: Matches "cat", "scat", but not "catch".

Regular expressions in text processing

- \{m\}: m occurrences of the previous character.
  - a\{3\}: Matches "aaa" (three consecutive 'a' characters).
  - [0-9]\{5\}: Matches any five-digit number.
- \{m,\}: At least m occurrences of the previous character.
  - b\{2,\}: Matches "bb", "bbb", "bbbb", and so on.
  - x\{4,\}: Matches "xxxx", "xxxxx", "xxxxxx", and so on.

Regular expressions in text processing

- \{m,n\}: Between m and n occurrences of the previous character.
  - c\{2,4\}: Matches "cc", "ccc", and "cccc".
  - y\{3,5\}: Matches "yyy", "yyyy", and "yyyyy".
- \(exp\): exp and attaches tag \1, \2, etc. to exp.
  - \(\d\d\)-\(\d\d\)-\(\d\d\d\d\):
    - Matches a date in the format "dd-mm-yyyy".
    - Each part (day, month, year) is captured with a separate tag: \1, \2, \3.

Regular expressions in text processing

- \w: A word character (same as [a-zA-Z0-9_]).
  - \w\+: Matches one or more word characters (e.g., "word", "123", "abc_").
  - @\w\+: Matches email addresses like "example@example.com".
- \W: A non-word character (same as [^a-zA-Z0-9_]).
  - \W\+: Matches one or more non-word characters (e.g., "@", "$", "#").
  - \W\d\W: Matches a non-word character followed by a digit followed by another non-word character.

Regular expressions in text processing

- \d: A digit (same as [0-9]).
  - \d\{3}: Matches any three-digit number (e.g., "123", "456").
  - \d\+\.\d\{2}: Matches a decimal number with two decimal places (e.g., "10.50", "3.14").
- \D: A non-digit (same as [^0-9]).
  - \D\{2,\}: Matches two or more consecutive non-digit characters.
  - \D\S: Matches a non-digit character followed by a non-whitespace character.

Regular expressions in text processing

- \s: A whitespace character.
  - \s\+: Matches one or more whitespace characters (e.g., spaces, tabs, newlines).
  - \s\d\s: Matches a whitespace character before and after a digit.
- \S: A non-whitespace character.
  - \S\{4}: Matches any four consecutive non-whitespace characters.
- \t: A tab.
  - \t\w\+\t: Matches a word surrounded by tabs.
  - ^\t\d\+\t$: Matches a line that starts and ends with tabs, with digits in between.

Regular expressions in text processing

- [[:alpha:]] - An alphabetic character.
- [[:lower:]] - A lowercase alphabetic character.
- [[:upper:]] - An uppercase alphabetic character.
- [[:digit:]] - A numeric character.
- [[:alnum:]] - An alphanumeric character.
- [[:space:]] - A whitespace character, including form feed.
- [[:blank:]] - A space or tab.
- [[:punct:]] - A punctuation character (not a space, letter, digit, or control character).

Regular expressions in text processing

Examples

- [[:alpha:]]\+: Matches one or more alphabetic characters (e.g., "abc", "DEF").
- [[:lower:]]\{3\}: Matches exactly three lowercase alphabetic characters (e.g., "abc", "xyz").
- [[:upper:]][[:digit:]]: Matches an uppercase letter followed by a digit (e.g., "A1", "B9").
- [[:alnum:]]*: Matches zero or more alphanumeric characters (e.g., "123", "abc").

Regular expressions in text processing

- [[:space:]]: Matches any whitespace character, including spaces and tabs.
- [[:blank:]]\+: Matches one or more spaces or tabs.
- [[:print:]]{5,10}: Matches printable characters with a length between 5 and 10.
- [[:punct:]]: Matches punctuation characters like ".", "!", "?", etc.

References

1. Your Unix :The Ultimate Guide, Das, Sumitabha