Fundamentals of Regex

Regular expressions are extremely useful in extracting information from text such as code, log files, spreadsheets, or even documents.

Fields of application range from validation to parsing/replacing strings, passing through translating data to other formats and web scraping.

Anchors — ^ and $

^The matches any string that **starts with The**

end$ matches a string that **ends with end**

^The end$ **exact string match** (starts and ends with **The end**)

roar matches any string that **has the text roar in it**

Quantifiers — * + ? and {}

abc***** matches a string that has **ab followed by zero or more c**

abc**+** matches a string that has **ab followed by one or more c**

abc**?** matches a string that has **ab followed by zero or one c**

abc**{2}** matches a string that has **ab followed by 2 c**

abc**{2,}** matches a string that has **ab followed by 2 or more c**

abc**{2,5}** matches a string that has **ab followed by 2 up to 5 c**

a**(bc)*** matches a string that has **a followed by zero or more copies of the sequence bc**

a**(bc){2,5}** matches a string that has **a followed by 2 up to 5 copies of the sequence bc**

OR operator — | or []

a(b|c) matches a string that has **a followed by b or c** *(and captures **b or c**)*

a[bc] same as previous, *but without capturing b or c*

Character classes — \d \w \s and .

\d matches a **single character** that is a **digit**

\w matches a **word character** (alphanumeric character plus underscore)

\s matches a **whitespace character** (includes tabs and line breaks)

. matches **any character**

\d, \w and \s also present their negations with \D, \W and \S respectively.

For example, \D will perform the inverse match with respect to that obtained with \d.

\D matches a **single non-digit character**

In order to be taken literally, you must escape the characters ^.[$()|*+?{\with a backslash \ as they have special meaning.

\$\d matches a string that has a **$ before one digit**

Bracket expressions—[]

[abc] matches a string that has **either an a or a b or a c** -> is the **same as a|b|c**

[a-c] same as previous

[a-fA-F0-9] a string that represents **a single hexadecimal digit, case insensitively** ->

[0-9]% a string that has a character **from 0 to 9 before a % sign**

[^a-zA-Z] a string that has **not a letter from a to z or from A to Z.** In this case the ^ is used as **negation of the expression**

\babc**\b** performs a **"whole words only" search**

\Babc**\B** matches only if the pattern is **fully surrounded by word** characters

1. 0 or 11 or 101

   0 | 11 | 101
2. only 0s

   0*
3. all binary strings

   (0|1)*
4. all binary strings except empty string

   (0|1)(0|1)*
5. begins with 1, ends with 1

   1 | (0|1)*|1
6. ends with 00

   (0|1)*00
7. contains at least three 1s

   (0|1)*1(0|1)*1(0|1)*1
8. contains at least three consecutive 1s

   (0|1)*111(0|1)*
9. contains the substring 110

   (0|1)*110(0|1)*
10. doesn't contain the substring 110

    (0|10)*1*
11. contains at least two 0s but not consecutive 0s

    (1*011*(0+011*))*
12. has at least 3 characters, and the third character is 0

    (0|1)(0|1)0(0|1)*
13. number of 0s is a multiple of 3

    1*|(1*01*01*01)*
14. starts and ends with the same character

    1(0|1)*1|0(0|1)*0
15. odd length

    (0|1)((0|1)(0|1))*
16. starts with 0 and has odd length, or starts with 1 and has even length

    0((0|1)(0|1))*|1(0|1)((0|1)(0|1))*
17. length is at least 1 and at most 3

    (0|1)|(0|1)(0|1)|(0|1)(0|1)(0|1)

How to Find or Validate an Email Address

^[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}$

Matching a Valid Date

(0[1-9]|1[012])[- /.](0[1-9]|[12][0-9]|3[01])[- /.](19|20)\d\d

IP Addresses

\b\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}\b

Match an *American Express Credit Card Number* which always begin with 34 or 37 and totals 15 digits.

/3[47]\d{13}/

Match a full U.S. Phone Number: **+1-(555)-555-5555**

/\+1-\(\d{3}\)-\d{3}-\d{4}/

References

- https://www.princeton.edu/~mlovett/reference/Regular-Expressions.pdf
- https://medium.com/factory-mind/regex-tutorial-a-simple-cheatsheet-by-examples-649dc1c3f285