Regular Expressions

Regular expressions are a language for string patterns.

RegEx is integral to many programming languages:
- Perl
- Python
- Javascript
- PHP
- C/C++
- grep (in Unix)
- Java (java.util.regex.*)
  and many more

RegEx is often the best way to validate String input.

A RegEx is compared with a String, looking for a “match”.

Regular Expression Notation

Any single character is a regex.

The period (\.) is a regex. (A period matches any single character.)

If $R_1$ and $R_2$ are regexes, then so is
  - $R_1R_2$ matches $R_1$ followed immediately by $R_2$
  - $R_1\|R_2$ matches either $R_1$ or $R_2$
  - $R_1$ while forming a group

Examples

```
(red|blue) fish
(0|1|2|3)(A|B|C)\.
```

Write a regEx to match any possible digital time.
Regular Expression Repetition

If \( R \) is a regEx, then so is

- \( ^{\ast} \) matches \( R \) repeated zero or more times
- \( ^{+} \) matches \( R \) repeated one or more times
- \( ^{?} \) matches \( R \) optionally
- \( ^{\{n\}} \) matches \( R \) repeated \( n \) times (for positive integer \( n \))
- \( ^{\{m,n\}} \) matches \( R \) repeated at least \( m \) times and at most \( m \) times (for positive integers \( m \leq n \))

Examples

\[-7(\{0|1|2|3|4|5|6|7|8|9\})^{*}\]
\[(\{0|1|2|3|4|5|6|7|8|9\})^{\{3\}}-(\{0|1|2|3|4|5|6|7|8|9\})^{\{4\}}\]

Write a regEx to match any possible Social Security Number

Regular Expression Escape Characters

Some characters are used in regEx as metasymbols.

\[. \ + \ * \ ? \ ( \ ) \ { \ } \ [ \ ] \ \^ \ \$ \ | \]

To use metasymbols as literals they must be escaped (prefixed with a backslash - \)

Some characters have no printable symbol.

\n (\n) -- matches the new line character
\t (\t) -- matches the tab character
\r (\r) -- matches the carriage return character
\v (\v) -- matches the vertical tab character
\f (\f) -- matches the form feed character

Examples

This is a single line.\n
\[(0|1|2|3|4|5|6|7|8|9)^{\ast}\]
Regular Expression Character Classes

Within a regEx square brackets enclose a character class. Each character class matches one character.

Individual characters and contiguous character ranges are permitted within square brackets.

(MetasyMBOLS don't apply so escaping metASYMBOLS is not allowed.)

Examples
- [Dab]
- [c-e][A-Z5-7!]
- [a-zA*]*

Regular Expression Character Classes with ^

The one meta-symbol within character classes is ^.
A class that begins [^ matches any character except those given by the pattern inside the character class.

Examples
- [^a]
- [^0-2]
Regular Expression Location Anchors

These two metasymbols match only by position

-- matches only at the beginning of a string

-- matches only at the end of a string

Examples

^This must be the complete line.$

(^Now)!($)

Regular Expression Abbreviations

There are several pattern abbreviations. Here are some

-- matches a single digit (same as [0-9])

-- matches a single non-digit character (same as [^0-9])

-- matches a single white space character
  (same as [ \n\t\r\f\v])

-- matches a single character that is not white space
  (same as [^ \n\t\r\f\v])

-- matches a single character that is alphanumeric
  (same as [a-zA-Z0-9])

-- matches any single character that is not alphanumeric
  (same as [^a-zA-Z0-9])