Starting shortly

Please wait!

ActivityInfo

Learning Regular Expressions for validation rules and quality data
Presentation outline

Overview

- What is a regular expression?
- Basic Syntax
- Groups
- More questions
What is a regular expression?
Definition
REGULAR EXPRESSIONS

“A sequence of characters that specifies a search pattern”

- Wikipedia
Literal and Meta- characters

REGULAR EXPRESSIONS

Example:

Literal characters, match only “A”, “B”, “D”

Metacharacter, meaning “match any character”

Matches:
- ABCD
- ABXD
- AB3D
- AB9D
- AB . D
Escaping meta characters

REGULAR EXPRESSIONS

Example:

Literal characters, match only “A”, “B”, “.” and “D”

Matches:

AB\.D = AB.D
Basic Syntax
Quantifiers

Syntax

. Match any character

[0123456789] Matches any of the characters between the brackets

[0-9] Matches any of the character ranges

[A-Z] Matches any uppercase letter

[0-9A-Za-z] Matches any letter or number

[^abc] Matches any character NOT in range
Quantifiers

Syntax

Literal characters, match only “A” or “B”

Metacharacter, meaning “match any character”

Quantifier, meaning “maybe” (“zero” or “once”)

Matches:

AB
AB
AB
Quantifiers

Syntax

Literal characters, match only “A” or “B”

Metacharacter, meaning “match any character”

Quantifier, meaning “at least once”

Matches:

AB#
ABCXD
ABX122:
Quantifiers

Syntax

Literal characters, match any digit

Metacharacters, meaning “match any digit”

Quantifier, meaning “At least once”

Matches:
0
123
1253232
Quantifiers

Syntax

Literal characters, match any digit

[0-9]{6}

Metacharacters, meaning “match any digit”

Quantifier, meaning “Match 6 times”

Matches:

123456
932311
235235
234234
Quantifiers

Syntax

Literal characters, match any digit

Metacharacters, meaning “match any digit”

Quantifier, meaning “Between 3 and 6 times”

Matches:

- 123
- 1334
- 123456
- 94389
## Character class shortcuts

<table>
<thead>
<tr>
<th>Shortcut</th>
<th>Description</th>
<th>Equivalent class</th>
</tr>
</thead>
<tbody>
<tr>
<td>\d</td>
<td>Digits</td>
<td>[0-9]</td>
</tr>
<tr>
<td>\w</td>
<td>Alphanumeric and _</td>
<td>[A-Za-z0-9_]</td>
</tr>
<tr>
<td>\s</td>
<td>Whitespace</td>
<td>[ ]</td>
</tr>
</tbody>
</table>
Character class shortcuts - Inverse

<table>
<thead>
<tr>
<th>Shortcut</th>
<th>Description</th>
<th>Equivalent class</th>
</tr>
</thead>
<tbody>
<tr>
<td>\D</td>
<td>NOT Digits</td>
<td>[^0-9]</td>
</tr>
<tr>
<td>\W</td>
<td>NOT Alphanumeric nor _</td>
<td>[^A-Za-z0-9_]</td>
</tr>
<tr>
<td>\S</td>
<td>NOT Whitespace</td>
<td>[^ ]</td>
</tr>
</tbody>
</table>
Phone number

EXCEPTIONS

Dutch Mobile numbers:

06-47205389

06-\[0-9\]{8}

https://www.activityinfo.org/support/docs/regex/test.html
Email addresses

**EXAMPLES**

alex@bedatadriven.com

```
.+@.+\..+
```

https://www.activityinfo.org/support/docs/regex/test.html
Groups

SYNTAX

Literal characters, matches only “06”

Group is matched as a whole

Quantifier applies to the group as a whole

Matches:

06
0606
060606
Disjunctions

Either the pattern on the left, or on the right

(06) | (07)

Groups matched as a whole

Matches:

06
07
Assertions - Word boundaries

SYNTAX

\b safe

Requires a word boundary

Matches:

safe
she is safe

 Doesn’t match:

failsafe
vouchsafe
 Assertions - Beginning of input

SYNTAX

Must match at beginning of input

Matches:
- safe
- safety

Doesn't match:
- failsafe

She is safe
Look-behind assertions

**SYNTAX**

Must be preceded by

```
(?<=@) .+
```

Matches:

```
bob@google.com
```

Doesn't match:

```
google.com
```