

Starting  
shortly

Please  
wait!

# ActivityInfo

Learning Regular Expressions for validation rules  
and quality data



ActivityInfo

# 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”

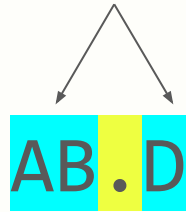
- [Wikipedia](#)

# Literal and Meta- characters

## REGULAR EXPRESSIONS

Example:

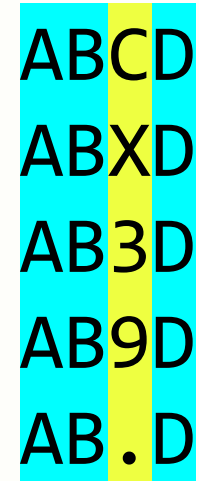
Literal characters, match  
only "A", "B", "D"



AB.D

Metacharacter, meaning  
"match any character"

Matches:



ABCD  
ABXD  
AB3D  
AB9D  
AB.D

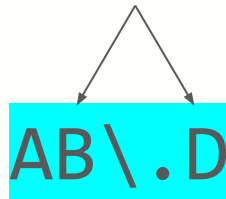


# Escaping meta characters

## REGULAR EXPRESSIONS

Example:

Literal characters, match  
only "A", "B", "." and "D"



AB\\.D

=

AB.D

Matches:

# Basic Syntax

# Quantifiers

## Syntax

<code>.</code>	Match any character
<code>[0123456789]</code>	Matches any of the characters between the brackets
<code>[0-9]</code>	Matches any of the character ranges
<code>[A-Z]</code>	Matches any uppercase letter
<code>[0-9A-Za-z]</code>	Matches any letter or number
<code>[^abc]</code>	Matches any character NOT in range



# Quantifiers

## Syntax

Literal characters, match  
only "A" or "B"

AB.?

Metacharacter, meaning  
"match any character"

Quantifier, meaning  
**"maybe"**  
("zero" or "once")

Matches:

ABC  
ABX  
AB



# Quantifiers

## Syntax

Literal characters, match  
only "A" or "B"

AB.+

Metacharacter, meaning  
"match any character"

Quantifier, meaning  
**"at least once"**

Matches:

AB#  
ABCXD  
ABX122:



# Quantifiers

## Syntax

Literal characters, match  
any digit

[0-9]+

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

[0-9]{3,6}

Metacharacters,  
meaning  
“match any digit”

Quantifier, meaning  
“Between 3 and 6 times”

Matches:

123  
1334  
123456  
94389

# Character class shortcuts

## SYNTAX

Shortcut	Description	Equivalent class
<code>\d</code>	Digits	<code>[0-9]</code>
<code>\w</code>	Alphanumeric and <code>_</code>	<code>[A-Za-z0-9_]</code>
<code>\s</code>	Whitespace	<code>[ ]</code>

# Character class shortcuts - Inverse

## SYNTAX

Shortcut	Description	Equivalent class
<code>\D</code>	NOT Digits	<code>[^0-9]</code>
<code>\W</code>	NOT Alphanumeric nor <code>_</code>	<code>[^A-Za-z0-9_]</code>
<code>\S</code>	NOT Whitespace	<code>[^ ]</code>

# Phone number

## EXAMPLES

Dutch Mobile numbers:

06-47205389

06-[0-9]{8}

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



ActivityInfo



# Email addresses

## EXAMPLES

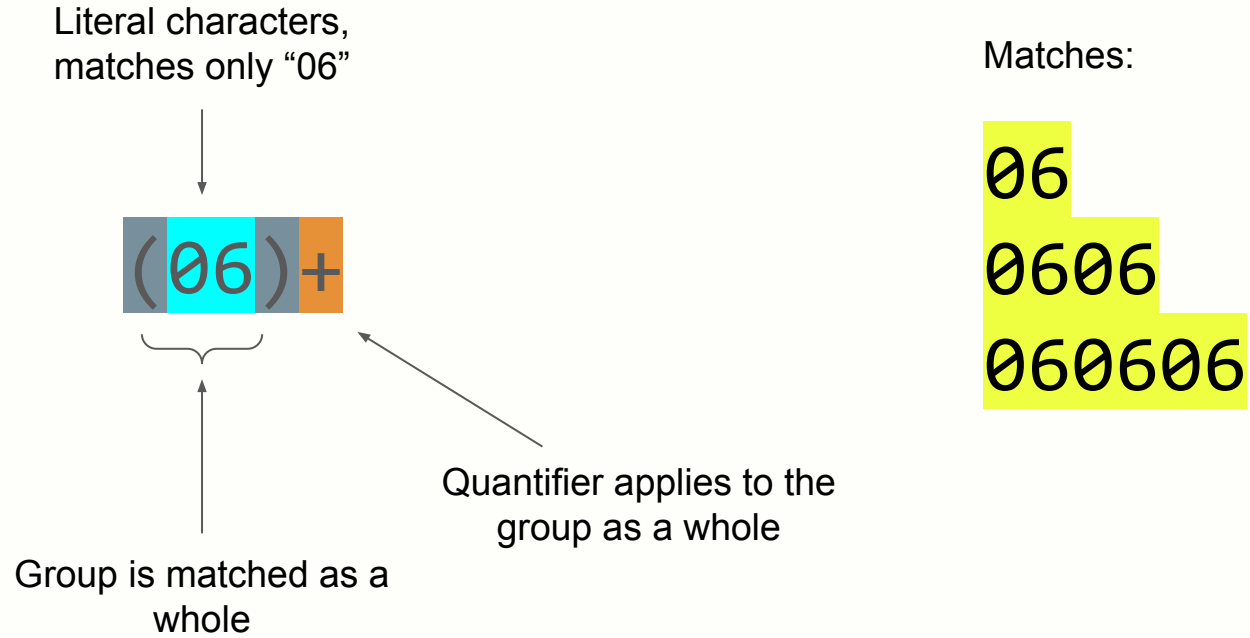
alex@bedatadriven.com

.+@.+\.+.+

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

# Groups

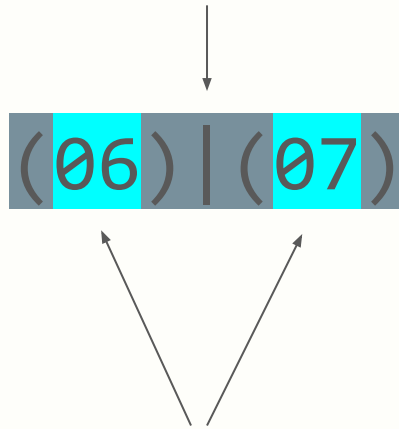
## SYNTAX



# Disjunctions

## SYNTAX

Either the pattern on the left, or on the right



Groups matched as a whole

Matches:

06  
07

# Assertions - Word boundaries

## SYNTAX

Requires a word  
boundary

↓  
**\b**safe

Matches:

safe

she is safe

Doesn't match:

failsafe

vouchsafe

# Assertions - Beginning of input

## SYNTAX

Must match at beginning  
of input

↓  
^ safe

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**