

## Regular expressions in Perl

This document presents a tabular **summary** of the regular expression (regexp) syntax in Perl, then illustrates it with a collection of annotated **examples**.

### Metacharacters

char	meaning
<b>^</b>	beginning of string
<b>\$</b>	end of string
<b>.</b>	any character except newline
<b>*</b>	match 0 or more times
<b>+</b>	match 1 or more times
<b>?</b>	match 0 or 1 times; <i>or</i> : shortest match
<b> </b>	alternative
<b>()</b>	grouping; “storing”
<b>[]</b>	set of characters
<b>{ }</b>	repetition modifier
<b>\</b>	quote or special

To present a metacharacter as a data character standing for itself, precede it with `\` (e.g. `\.` matches the full stop character `.` only).

In the table above, the characters themselves, in the first column, are links to descriptions of characters in my *The ISO Latin 1 character repertoire - a description with usage notes*. Note that the physical appearance (glyph) of a character may vary from one device or program or font to another.

### Repetition

<i>a</i> *	zero or more <i>a</i> 's
<i>a</i> +	one or more <i>a</i> 's
<i>a</i> ?	zero or one <i>a</i> 's (i.e., optional <i>a</i> )
<i>a</i> { <i>m</i> }	exactly <i>m</i> <i>a</i> 's
<i>a</i> { <i>m</i> ,}	at least <i>m</i> <i>a</i> 's
<i>a</i> { <i>m</i> , <i>n</i> }	at least <i>m</i> but at most <i>n</i> <i>a</i> 's
<i>repetition</i> ?	same as <i>repetition</i> but the <i>shortest</i> match is taken

Read the notation *a*'s as “occurrences of strings, each of which matches the pattern *a*”. Read *repetition* as any of the repetition expressions listed above it. Shortest match means that the shortest string matching the pattern is taken. The default is “greedy matching”, which finds the longest match. The *repetition*? construct was introduced in Perl version 5.

## Special notations with `\`

Single characters	“Zero-width assertions”
<code>\t</code> tab	<code>\b</code> “word” boundary
<code>\n</code> newline	<code>\B</code> not a “word” boundary
<code>\r</code> return (CR)	
<code>\xhh</code> character with hex. code <i>hh</i>	

### Matching

<code>\w</code>	matches any <i>single</i> character classified as a “word” character (alphanumeric or “_”)
<code>\W</code>	matches any non-“word” character
<code>\s</code>	matches any whitespace character (space, tab, newline)
<code>\S</code>	matches any non-whitespace character
<code>\d</code>	matches any digit character, equiv. to <code>[0-9]</code>
<code>\D</code>	matches any non-digit character

## Character sets: specialities inside [...]

**Different meanings** apply inside a character set (“character class”) denoted by [...] so that, **instead of** the normal rules given here, the following apply:

[ <i>characters</i> ]	matches any of the characters in the sequence
[ <i>x-y</i> ]	matches any of the characters from <i>x</i> to <i>y</i> (inclusively) in the ASCII code
[\-]	matches the hyphen character “-”
[\n]	matches the newline; other single character denotations with \ apply normally, too
[^ <i>something</i> ]	matches any character <i>except</i> those that [ <i>something</i> ] denotes; that is, immediately after the leading “[”, the circumflex “^” means “not” applied to all of the rest

## Examples

expression	matches...
abc	abc (that exact character sequence, but anywhere in the string)
^abc	abc at the <i>beginning</i> of the string
abc\$	abc at the <i>end</i> of the string
a b	either of a and b
^abc abc\$	the string abc at the beginning or at the end of the string
ab{2,4}c	an a followed by two, three or four b’s followed by a c
ab{2,}c	an a followed by at least two b’s followed by a c
ab*c	an a followed by any number (zero or more) of b’s followed by a c
ab+c	an a followed by one or more b’s followed by a c
ab?c	an a followed by an optional b followed by a c; that is, either abc or ac
a.c	an a followed by any single character (not newline) followed by a c
a\.c	a.c exactly
[abc]	any one of a, b and c
[Aa]bc	either of Abc and abc
[abc]+	any (nonempty) string of a’s, b’s and c’s (such as a, abba, acbabcacaa)
[^abc]+	any (nonempty) string which does <i>not</i> contain any of a, b and c (such as defg)
\d\d	any two decimal digits, such as 42; same as \d{2}
\w+	a “word”: a nonempty sequence of alphanumeric characters and low lines (underscores), such as foo and 12bar8 and foo_1
100\s*mk	the strings 100 and mk optionally separated by any amount of white space (spaces, tabs, newlines)
abc\b	abc when followed by a word boundary (e.g. in abc! but not in abcd)
perl\b	perl when <i>not</i> followed by a word boundary (e.g. in perlert but not in perl stuff)