Регулярные выражения

R класс 03.03.2015

Регулярные выражения

формальный язык поиска и осуществления манипуляций с подстроками в тексте, основанный на использовании метасимволов

(символов-джокеров, *англ.* wildcard characters)

Регулярные выражения

Базовые регулярные выражения POSIX

basic regular expressions (BRE)

Расширенные регулярные выражения POSIX

extended regular expressions (ERE)

Регулярные выражения, совместимые с Perl

(Perl compatible regular expression (PCRE))

Примеры регулярных выражений

Знак Йены

\u00A5

Email

^([\w\-\.]+)@((\[([0-9]{1,3}\.){3}[0-9]{1,3}\])|(([\w\-]+\.)+)([a-zA-Z]{2,4}))$

Дата

^((0[1-9])|(1[0-2]))\/(\d{2})$

Телефон

^(\+[1-9][0-9]*(\([0-9]*\)|-[0-9]*-))?[0]?[1-9][0-9\-]*$

Номер карты VISA

^(\d{4}[-]){3}\d{4}|\d{16}$

Время в 24-часовом формате

([0-1][0-9]|2[0-3]):[0-5][0-9]

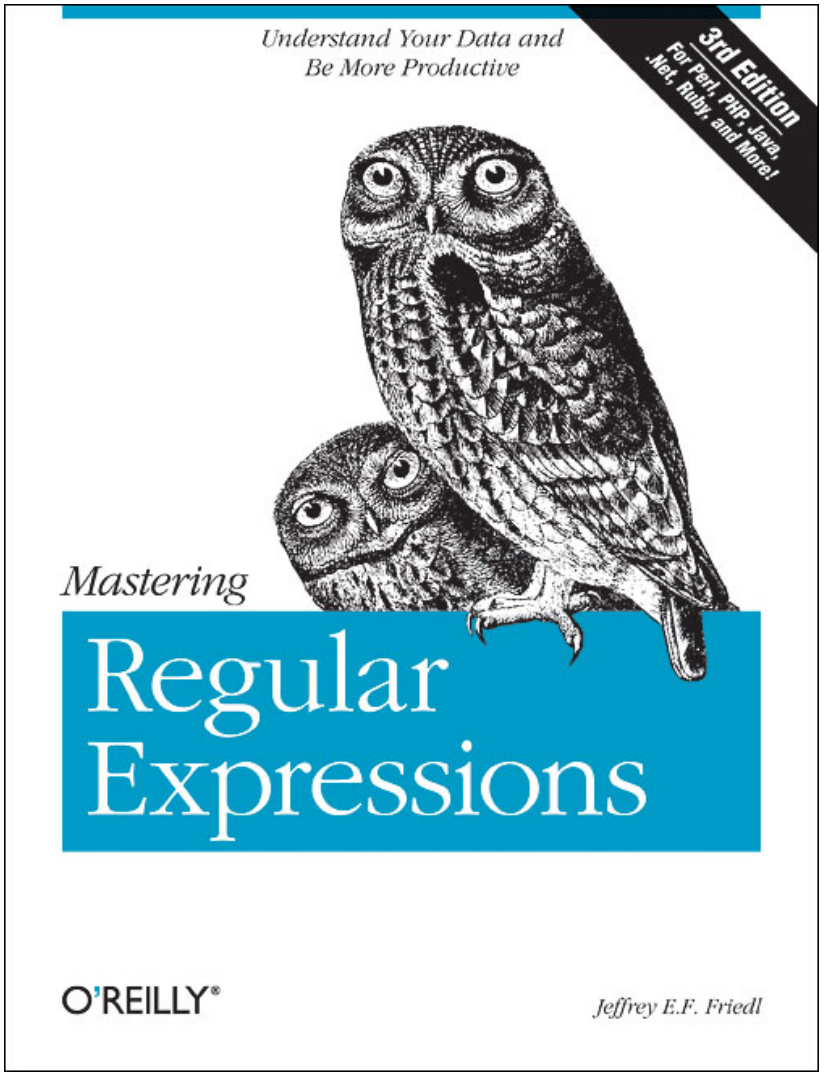
Примеры регулярных выражений

Справочные материалы

- https://ru.wikibooks.org/wiki/Регулярные_выражения
- http://www.regular-expressions.info
- http://www.rexegg.com
- http://stackoverflow.com
- http://www.books.ru/books/regulyarnye-vyrazheniya-3-e-izdanie-592346/
- http://www.amazon.com/Mastering-Regular-Expressions-Jeffrey-Friedl/dp/0596528124/ref=sr_1_fkmr1_1?ie=UTF8&qid=1425021802&sr=8-1-fkmr1&keywords=Regular+expression+Friddle

Справочные материалы

Регулярные выражения, третья редакция
Джефри Фриддл.



Регулярные выражения в реальном времени

- http://regex101.com
- http://regexr.com

Отладка

- http://debuggex.com

RegEx

Regular Expression Basics

.	Any character except newline
a	The character a
ab	The string ab
a|b	a or b
a*	0 or more a's
****	Escapes a special character

Regular Expression Quantifiers

*****	0 or more
+	1 or more
?	0 or 1
{2}	Exactly 2
{2, 5}	Between 2 and 5
{,5}	No more 5
{2,}	2 or more

RegEx

Regular Expression Character Classes

[ab-d]	One character of: a, b, c, d
[^ab-d]	One character except: a, b, c, d
\d	One digit
\D	One non-digit
\s	One whitespace
\S	One non-whitespace
\w	One word character
\W	One non-word character

Regular Expression Assertions

^	Start of string
$	End of string
\b	Word boundary
\B	Non-word boundary
(?=...)	Positive lookahead
(?!...)	Negative lookahead
(?<=...)	Positive lookbehind
(?<!...)	Negative lookbehind
(?(...)|)	Conditional

/foo(?=bar)/
foobar foobaz

/foo(?!bar)/
foobar foobaz

/(?<=foo)bar/
foobar foobaz

/(?<=!not)foo/
not foo but foo

RegEx

Regular Expression Groups

(...)	Capturing group
(?P<Y>...)	Capturing group named Y
(?:...)	Non-capturing group
\Y	Match the Y'th captured group
(?P=Y)	Match the named group Y
\g{Y}	Match the named or numbered group Y
(?#...)	Comment

Regular Expression Flags

i	Ignore case
m	^ and $ match start and end of line
g	global. All matches
(?igm)	Set flags within regex

RegEx

Regular Expression Posix Classes	
[:alnum:]	Letters and digits
[:alpha:]	Letters
[:blank:]	Space or tab only
[:digit:]	Decimal digits
[:graph:]	Visible characters, except space
[:lower:]	Lowercase letters
[:print:]	Visible characters
[:punct:]	Visible punctuation characters
[:space:]	Whitespace
[:upper:]	Uppercase letters
[:word:]	Word characters

Regular Expression Special Characters	
\n	Newline
\r	Carriage return
\t	Tab
\0	Null character

Регулярные выражения в R

Поиск

grep(match, x, perl=TRUE, value=FALSE)

Числовой вектор длины x или менее

grep(match, x, perl=TRUE, value=TRUE)

Строковый вектор длины x или менее

grepl(match, x, perl=TRUE)

Логический вектор длины x

regexpr(match, x, perl=TRUE)

Числовой вектор длины x, значения — длина match

Числовой вектор длины x, значения — количество совпадений match

gregexpr(match, x, perl=TRUE)

Список из числовых векторов. Размер списка = x.
Функция аналогична regexpr только ищет все вхождения match.

regmatches(x, m)

Строковый вектор длиной x или список из строковых векторов размером x.
Возвращает значения соответствующие регулярному выражению

Регулярные выражения в R

Замена

sub(match, replace, x, perl=TRUE)
gsub(match, replace, x, perl=TRUE)

Строковый вектор длины x.
gsub выполняет замену всех совпадений в строке с регулярным выражением, а не только первого.

Примеры в regex101 и в R