Синтаксис регулярных выражений

Traffic Monitor component knowledge base

Exported on 11/05/2019

1 Table of Contents

При написании регулярного выражения можно пользоваться обозначениями, приведенными в таблице:

Представление	Значение	Эквивалент	Комментарий	Пример
\w	любая буква	[А-яA-zЁё]	Будет ловиться буква как кириллического, так и латинского алфавитов вне зависимости от регистра. При этом не ловятся буквы с диакритиками (Àà, Èè, Ùù, Éé, Çç, å) и особые буквы некоторых алфавитов типа норвержских æ, ø. Буквы других алфавитов (арабского, китайского) ловиться не будут. Цифры и нижнее подчёркивание не ловятся.	Регулярное выражение "Пет\w" будет ловить слова "Петя", "Пете", "Петю", "Петр" и т.д. Регулярное выражение "\wоля" будет ловить слова "Коля", "коля", "Толя", "толя" и т.д.
\W	не буква	[^А-яA-zЁё]	Соответственно помимо небуквенных знаков типа -!,.?%;№ ловит ещё и буквы с диакритиками и нетипичные буквы алфавитов, составленных на основе латиницы или кириллицы (например, корякскую "ƕ"), а также буквы не кириллического и не латинского алфавита.	
\d	любая цифра	[0-9]		
\D	не цифра	[^0-9]	в том числе пробел и перенос строки	

Представление	Значение	Эквивалент	Комментарий	Пример
\s	пробел	( )	ловит не только пробел, но и табуляцию	
\S	не пробел	[^]	ловит в том числе и табуляцию	
[ ]	символьный класс		Внутри скобок задаются символы, один из которых может встретиться в перехватываемой строке	Регулярное выражение "Пет[яюе]" поймает и "Петя", и "Пете", и "Петю". [0-9] соответствует всем цифрам от нуля до девяти. [А-Яа-я] соответствует всем буквам русского алфавита, кроме "Ёё". Также возможна запись [А-я]. [А-Г] соответствует русским заглавным буквам от "А" до "Г" (иначе можно было бы это записать так: [АБВГ], а [А-о] - всем русским заглавным буквам, кроме "Ё", а также русским строчным буквам от "а" до "о", за исключением "ё". Внутри квадратных скобок можно задавать одновременно и записанный с помощью дефиса символьный класс, и ряд обычных символов: [A-zА-я0-9Ёё!,%] будет реагировать на все буквы русского и английского алфавитов, на цифры, а также на восклицательный знак, запятую и знак процента.
[^]	исключение из набора			[^0-9] – не цифра
*	ноль или более	{0,}	квантификатор ставится после символа, группы или символьного класса. Является жадным	[0-9]* – любое количество цифр, в том числе и ни одной цифры.

Представление	Значение	Эквивалент	Комментарий	Пример
+	один или более	{1,}	квантификатор ставится после символа, группы или символьного класса. Является жадным	\w+ – минимум одна буква
()	группировка		заключённое в скобки выражение рассматривается как единое целое	O(хо)+ – После "О" идет как минимум одно "хо": поймается и "Охо", и "Охохохохохохо".
{}	количество повторений		ставится после символа, группы или символьного класса	[0-9]{9} – девять любых цифр подряд. [0-9]{2,11} – от двух до одиннадцати любых цифр подряд. \w{0,20} – не больше двадцати букв.
?	ноль или один	{0,1}	квантификатор ставится после символа, группы или символьного класса. Альтернативой ему может служить {0,1}	Петр(ушка)? – реагирует как на "Петр", так и на "Петрушка".
|	или		в ситуации, когда выбор идет между несколькими односимвольными альтернативами, вместо | можно использовать [] Ср. "Пет(я|ю)" и "Пет[яю]"	Регулярное выражение "паспорт (номер|№)" поймает как строку "паспорт номер", так и "паспорт №"
.	любой символ			(.*) – все, что угодно (буквы, символы, пунктуационные знаки, разделители любого рода) любое количество раз, в том числе и ноль раз

Представление	Значение	Эквивалент	Комментарий	Пример
\	экранирование		экранировать следует те символы, которые являются служебными при написании регулярных выражений: .\()-[]+?	Регулярное выражение "[кч]то?" ловит слова "кто", "что", "кт" и "чт". Регулярное выражение "[кч]то\?" ловит только "кто?" и "что?", но не, скажем, "что!" Внутри символьного класса экранировать требуется только дефис.
\r\n\t	возврат каретки, новая строка, табуляция		Поскольку разные текстовые редакторы могут маркировать перенос строки различным образом, надежнее задавать перенос строки в регулярных выражениях так: [\r\n]{1,2}	Регулярное выражение "бутылка(\t)рома" поймает строку вида "бутылка+табуляция+рома", но не "бутылка+пробел+рома".

2 Про жадность квантификаторов:

В случае, когда в объекте защиты порог встречаемости для тестового объекта больше одного, следует помнить, что квантификаторы "*" и "+" являются "жадными", то есть они пытаются съесть максимально длинный кусок строки, на который они могут налезть.
Например, мы создали регулярное выражение вида

```
\((.*)\)
```

Оно предназначено для того, чтобы отлавливать информацию, данную в круглых скобках. Но в строке типа *Петя (он был очень любопытен) решил посмотреть, что находится в заброшенном доме (хотя мама просила его туда не лазить).* наше выражение увидит только одну группу в скобках, а именно: *(он был очень любопытен) решил посмотреть, что находится в заброшенном доме (хотя мама просила его туда не лазить).* То есть оно среагирует на самую первую открывающую скобку и на самую последнюю закрывающую. Если бы мы ввели ограничение на то, что в скобках могут быть только символы русского алфавита и пробелы, этого бы не произошло. Так, регулярное выражением вида

```
\([А-я ]+\)
```

поймало бы обе группы в скобках.

3 Пример регулярного выражения:

```
(^|([\r\n]{1,2}))Весьма конфиденциально[\r\n]{1,2}Экз\. ?[№]? ?[0-9]{1,4}[\r\n]{1,2}(веселый|ВЕСЕЛЫЙ) ['"].{0,10}['"][\r\n]{1,2}Степень веселья по шкале [0-9]{1,2}
```