Введение в регулярные выражения

Регулярные выражения (regular expressions, regex, regexp) – формальный язык поиска и осуществления манипуляций с подстроками в тексте, основанный на использовании метасимволов. Впервые появились в системе UNIX. Поддержка есть в .NET, Java, JavaScript, Perl, PHP, Python и др.

Шаблон (pattern) – строка-образец, состоит из символов и метасимволов и задает правило поиска. Синтаксис шаблонов в разных языках программирования в основном одинаковый.
Строка замены – используется для манипуляций с текстом, может содержать в себе спецсимволы.

Результатом работы с регулярным выражением может быть:

- проверка наличия искомого образца в заданном тексте;
- определение подстроки текста, соответствующей образцу;
- определение групп символов, соответствующих отдельным частям образца;
- удаление найденных подстрок;
- замена найденных подстрок по определенному шаблону.

Составление шаблонов

Классы символов

Обозначение класса символов	Пример
[...] – любой символ, указанный в []	[А-ЯЁ] – любая заглавная буква русского алфавита
[^...] – любой символ кроме указанных в []	[^0-9] – символ кроме цифр
. – любой символ кроме \n	
\w – любая буква, цифра, нижнее подчеркивание	
\W – любой символ кроме \w	
\s – пробельный символ (\t \n \r \v \f)	
\S – непробельный символ	
\d – цифра	
\D – не цифра	

Квантификация (поиск последовательностей)

Квантификатор определяет, сколько повторов символа искать, указывается после требуемого символа

Обозначение квантификатора	Пример
? – 0 или 1 повторение	\.**?** – 0 или 1 точка
* – 0 и более повторений	x\d***** – 0 и более цифр после символа x 123 x018 x xyz
+ – 1 и более повторений	\w+ – 1 и более буквенно-цифровых символов (поиск слов) привет_мир, hello, 123 x123 \s+ – 1 и более пробельных символов привет_мир, hello, 123 x123
{n} – ровно n повторений	\d**{2}** – 2 цифры
{m,n} – от m до n повторений	\d**{1,4}** – от 1 до 4 цифр
{m,} – не менее m повторений	\s**{2,}** – 2 и более пробельных символа привет_мир, hello, 123 x123
{,n} – не более n повторений	:[a-z]**{,5}** – после двоеточия не более 5 букв латинского алфавита :test :1234test :pwd

Якоря (привязки, позиция внутри строки)

Обозначение якоря	Пример
^ – начало строки	^\+7 – ищет строки, начинающиеся с +7
$ – конец строки	[.?!]**$** – ищет . или ? или ! В конце строки
\G – совпадение начинается там, где закончилось предыдущее	\G\(\d\) – ищет идущие подряд цифры в () (1)(3)(5)[7](9)

\b – граница (bound) слова (между \w и \W)	\w+ – 1 и более буквенно-цифровых символов (поиск слов) привет_мир, hello, 123 x123 \s+ – 1 и более пробельных символов привет_мир, hello, 123 x123
\B – не граница слова	\bТест\B – слова, начинающиеся на Тест Тест, тестировщик, Тестирование

Группировка

Обозначение	Пример
(выражение) – захватывает подстроку, соответствующую выражению. Нумерация групп с 1 (0 – все совпадение с шаблоном)	(\d{3};){2} – два повтора группы из трех цифр и ; 1234;567;890;12 (\w)\1 – два одинаковых символа \w (\1 – символы 1ой группы) Hello. address
(?<имя> выражение) или (?'имя' выражение) - выделяет именованную группу	x\d* – 0 и более цифр после символа x 123 x018 x xyz
(выражение1 | выражение2) – совпадение с 1 или 2 выражением	

Просмотр вперед и назад

Обозначение	Пример
(?=шаблон) – позитивный просмотр вперёд	Лев(?=XVI) – ищет Лев перед XVI ЛевXV, ЛевXVI, ЛевXVIII, ЛевLXVII, ЛевXXL
(?!шаблон) – негативный просмотр вперёд (с отрицанием)	Лев(?!XVI) – ищет Лев не перед XVI ЛевXV, ЛевXVI, ЛевXVIII, ЛевLXVII, ЛевXXL
(?<=шаблон) – позитивный просмотр назад	(?<=Сергей)Иванов – ищет Иванов после «Сергей » Сергей Иванов, Игорь Иванов
(?<!шаблон) – негативный просмотр назад (с отрицанием)	(?<!Сергей)Иванов – ищет Иванов не после «Сергей » Сергей Иванов, Игорь Иванов

Чувствительность к регистру

Обозначение	Пример
(?i) – без учета регистра	(?i)\a\w* - ищет слова, начинающиеся с буквы а в любом регистре Арбуз, банан, ананас, яблоко
(?-i) – с учетом регистра (используется по умолчанию)	

Подстановки

Обозначение	Пример
$*число* - замещает часть строки, соответствующую группе *число*	Шаблон: \b(\w+)(\s)(\w+)\b Шаблон замены: $3$2$1 Было: "one two" Стало: "two one"
${*имя*} - замещает часть строки, соответствующую именованной группе *имя*	Шаблон: \b(?<word1>\w+)(\s)(?<word2>\w+)\b Шаблон замены: ${word2} ${word1} Было: "one two" Стало: "two one"