|  |  |
| --- | --- |
| **Символ** | **Что обозначает в регулярном выражении** |
| **.** | Любой символ |
| **[...]** | Любой символ(диапазоны символов:a-z) из набора в скобках |
| **[^...]** | Любой символ не из набора в скобках |
| **\** | отменяет специальное значение следующего за ней символа Пр: \. Ξ . |
| **|** | Фрагмент справа или фрагмент слева. (А или Б) |
| **"\*?"** | Повторение фрагмента нуль или более раз (не жадное) |
| **"+?"** | Повторение фрагмента один или более раз (не жадное) |
| **"{m,n}?"** | Повторение предыдущего фрагмента от m до n раз включительно (не жадное) |
| **(...)** | Группировка |

Некоторые символьные классы можно заменить специальными метасимволами:

|  |  |
| --- | --- |
| **Символ** | **Описание** |
| \d | Соответствует цифре. Эквивалентно [0-9] |
| \D | Соответствует нецифровому символу. Эквивалентно [^0-9] |
| \s | Соответствует любому пробельному символу. Эквивалентно [ \f\n\r\t\v] |
| \S | Соответствует любому непробельному символу. Эквивалентно [^ \f\n\r\t\v] |
| \w | Соответствует любому буквенному символу, цифровому и знаку подчеркивания. Эквивалентно [[:word:]] |
| \W | Соответствует любому символу, кроме буквенного символа, цифрового или подчеркивания. Эквивалентно [^[:word:]] |

**Позиция внутри строки**

Следующие символы позволяют спозиционировать регулярное выражение относительно элементов текста: начала и конца строки, границ слова.

|  |  |  |  |
| --- | --- | --- | --- |
| **Представление** | **Позиция** | **Пример** | **Соответствие** |
| ^ | Начало строки | ^a | aaa aaa |
| $ | Конец строки | a$ | aaa aaa |
| \b | Граница слова | a\b | aaa aaa |
| \ba | aaa aaa |
| \B | Не граница слова | \Ba\B | aaa aaa |
| \G | Предыдущий успешный поиск | \Ga | aaa aaa (поиск остановился на 4-й позиции — там, где не нашлось a) |

## Управляющие символы

|  |  |  |  |
| --- | --- | --- | --- |
| **Представление** | **Символ** | **Обозначение** | **Расшифровка** |
| \t | Табуляция | HT | Horizontal tabulation |
| \v | Вертикальная табуляция | VT | Vertical tabulation |
| \r | Возврат каретки | CR | Carriage return |
| \n | Перевод строки | LF | Line feed |
| \f | Конец страницы | FF | Form feed |
| \a | Звонок | BEL | Bell character |
| \e | Escape-символ | ESC | Escape character |
| [\b] | Забой в квадратных скобок (иначе = граница слова). | BS | Backspace |
| \cA … \cZ | Ctrl+A … Ctrl+Z Эквивалентно \x01 … \x1A. |  |  |

**вантификация (поиск последовательностей)**

[*Квантификатор*](http://ru.wikipedia.org/wiki/%D0%9A%D0%B2%D0%B0%D0%BD%D1%82%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%82%D0%BE%D1%80) после символа, символьного класса или группы определяет, сколько раз предшествующее выражение может встречаться. Следует учитывать, что квантификатор может относиться более чем к одному символу в регулярном выражении, только если это символьный класс или группа.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Представление** | **Число повторений** | | **Пример** | | **Соответствие** |
| {*n*} | Ровно *n* раз | | colou{3}r | | colouuur |
| {*m*,*n*} | От *m* до *n* включительно | | colou{2,4}r | | colouur, colouuur, colouuuur |
| {*m*,} | Не менее *m* | | colou{2,}r | | colouur, colouuur, colouuuur и т. д. |
| {,*n*} | Не более *n* | | colou{,3}r | | color, colour, colouur, colouuur |
| **Представление** | **Число повторений** | **Эквив** | | **Пример** | **Соответствие** |
| \* | Ноль или более | {0,} | | colou\*r | color, colour и т. д. |
| + | Одно или более | {1,} | | colou+r | colour, colouur и т. д. (но не color) |
| ? | Ноль или одно | {0,1} | | colou?r | color, colour |

#### Жадная и ленивая квантификация

|  |  |  |  |
| --- | --- | --- | --- |
| **Жадный** | **Ленивый** | **Жадный** |  |
| \* | \*? | <a>.\*</a> | <a>txt1</a> txt2 <a>txt3</a> |
| + | +? | **Ленивый** |  |
| {*n*,} | {*n*,}? | <a>.\*?</a> | <a>txt1</a> txt2 <a>txt3</a> |

Также общей проблемой как жадных, так и ленив

### Группировка

#### Обозначение группы

Круглые скобки используются для определения области действия и [приоритета операций](http://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B8%D0%BE%D1%80%D0%B8%D1%82%D0%B5%D1%82_%D0%BE%D0%BF%D0%B5%D1%80%D0%B0%D1%86%D0%B8%D0%B8). Шаблон внутри группы обрабатывается как единое целое и может быть квантифицирован. Например, выражение (тр[ау]м-?)\* найдёт последовательность видатрам-трам-трумтрам-трум-трамтрум.

#### [Обратная связь](http://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%80%D0%B0%D1%82%D0%BD%D0%B0%D1%8F_%D1%81%D0%B2%D1%8F%D0%B7%D1%8C)

Одно из применений группировки — повторное использование ранее найденных групп символов (*подстрок*, *блоков*, *отмеченных подвыражений*). При обработке выражения подстро́ки, найденные по шаблону внутри группы, сохраняются в отдельной области памяти и получают номер начиная с единицы. Каждой подстроке соответствует пара скобок в регулярном выражении. Квантификация группы не влияет на сохранённый результат, то есть сохраняется лишь первое вхождение. Обычно поддерживается до 9 нумерованных подстрок с номерами от 1 до 9, но некоторые интерпретаторы позволяют работать с бо́льшим количеством. Впоследствии в пределах данного регулярного выражения можно использовать обозначения от \1 до \9 для проверки на совпадение с ранее найденной подстрокой.

Например, регулярное выражение (та|ту)-\1 найдёт строку та-та или ту-ту, но пропустит строку та-ту.

Также ранее найденные подстро́ки можно использовать при замене по регулярному выражению. В таком случае в замещающий текст вставляются те же обозначения, что и в пределах самого выражения.

#### (?:шаблон) Группировка без обратной связи

Если группа используется только для группировки и её результат в дальнейшем не потребуется, то можно использовать группировку вида(?:*шаблон*). Под результат такой группировки не выделяется отдельная область памяти и, соответственно, ей не назначается номер. Это положительно влияет на скорость выполнения выражения, но понижает удобочитаемость.

#### (?>шаблон) Атомарная группировка

Атомарная группировка (вида (?>*шаблон*)), так же как группировка без обратной связи, не создаёт обратных связей. В отличие от неё, такая группировка запрещает возвращаться назад по строке, если часть шаблона уже найдена.

|  |  |  |
| --- | --- | --- |
| **Пример** | **Соответствие** | **Создаваемые группы** |
| a(bc|b|x)cc | abccaxcc  abccaxcc | abccaxcc  abccaxcc |
| a(?:bc|b|x)cc | нет |
| a(?>bc|b|x)cc | abccaxcc  но не abccaxcc: вариант x найден, остальные проигнорированы |
| a(?>x\*)xa | не найдётся axxxa: все x заняты, и нет возврата внутрь группы |

Атомарная группировка выполняется ещё быстрее, чем группировка без обратной связи, и как запрещает проверку любых других вариантов внутри группы, когда один вариант уже найден.

**Просмотр вперёд и назад**

Для поиска фрагмента текста, «просматривая» (но не включая в найденное) окружающий текст, который расположен до или после искомого фрагмента текста. Пр, найти имя тега HTML, не включая в результат поиска окружающие его угловые скобки или др.знаки

|  |  |  |  |
| --- | --- | --- | --- |
| **Представлен** | **Вид просмотра** | **Пример** | **Соответствие** |
| (?=*шаблон*) | Позитивный просмотр вперёд | Людовик(?=XVI) | ЛюдовикXV, ЛюдовикXVI, ЛюдовикXVIII, ЛюдовикLXVII, ЛюдовикXXL |
| (?!*шаблон*) | Негативный просмотр вперёд (с отрицанием) | Людовик(?!XVI) | ЛюдовикXV,ЛюдовикXVI,ЛюдовикXVIII, ЛюдовикLXVII, ЛюдовикXXL |
| (?<=*шаблон*) | Позитивный просмотр назад | (?<=Сергей) Иванов | Сергей Иванов, Игорь Иванов |
| (?<!*шаблон*) | Негативный просмотр назад (с отрицанием) | (?<!Сергей) Иванов | Сергей Иванов, Игорь Иванов |

#### Модификаторы

Модификаторы действуют с момента вхождения и до конца регулярного выражения или противоположного модификатора. Некоторые интерпретаторы могут применить модификатор ко всему выражению, а не с момента его вхождения.

|  |  |  |
| --- | --- | --- |
| **Синтаксис** | **Описание** | |
| (?i) | Включает | чувствительность выражения к регистру символов ([англ.](http://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *case insensitivity*) |
| (?-i) | Выключает |
| (?s) | Включает | режим соответствия точки символам переноса строки и возврата каретки |
| (?-s) | Выключает |
| (?m) | Символы ^ и $ вызывают соответствие только | после и до символов новой строки |
| (?-m) | с началом и концом строки |
| (?x) | Включает | режим без учёта пробелов между частями регулярного выражения и позволяет использовать # для комментариев |
| (?-x) | Выключает |

Группы-модификаторы можно объединять в одну группу: (?i-sm). Такая группа включает режим i и выключает режим s, m. Если использование модификаторов требуется только в пределах группы, то нужный шаблон указывается внутри группы после модификаторов и двоеточия. Например, (?-i)(?i:tv)set найдёт TVset, но не TVSET.

Некоторых случаях предпочтительно представление символов по их коду.

|  |  |  |
| --- | --- | --- |
| **Представление** | **Пояснение** | **Кодировка** |
| \0*n* | *n* — восьмеричное число от 0 до 377 | 8-битная |
| \x*dd* | *d* — шестнадцатеричная цифра |
| \u*dddd* | 16-битная (Юникод) |

### Символьные классы [POSIX](http://ru.wikipedia.org/wiki/POSIX)

Многие диапазоны символов зависят от выбранных настроек локализации. POSIX стандартизовал объявление некоторых классов и категорий символов, как показано в следующей таблице.

|  |  |  |
| --- | --- | --- |
| **POSIX-класс** | **Эквивалент** | **Значение** |
| [:upper:] | [A-Z] | Символы верхнего регистра |
| [:lower:] | [a-z] | Символы нижнего регистра |
| [:alpha:] | [[:upper:][:lower:]] | Буквы |
| [:digit:] | [0-9], т. е. \d | Цифры |
| [:xdigit:] | [[:digit:]A-Fa-f] | Шестнадцатеричные цифры |
| [:alnum:] | [[:alpha:][:digit:]] | Буквы и цифры |
| [:word:] | [[:alnum:]\_], т. е. \w | Символы, образующие «слово» |
| [:punct:] | [-!"#$%&'()\*+,./:;<=>?@[\\\]\_`{|}~] | Знаки пунктуации |
| [:blank:] | [ \t] | Пробел и табуляция |
| [:space:] | [[:blank:]\v\r\n\f], т. е. \s | Пробельные символы |
| [:cntrl:] | [\x00-\x1F\x7F] | Управляющие символы |
| [:graph:] | [\x21-\x7E] | Печатные символы |
| [:print:] | [\x20-\x7E], т. е. [[:graph:] ] | Печатные символы с пробелом |

Использование класса возможно лишь **внутри квадратных скобок** пример частой ошибки —

^[:upper:]il+$ вместо ^**[**[:upper:]**]**il+$

#### Примеры шаблонов

r"\b\w+\b"

Соответствует слову из букв и знаков подчеркивания.

r"[+-]?\d+"

Соответствует целому числу. Возможно, со знаком.

r"\([+-]?\d+\)"

Число, стоящее в скобках. Скобки используются в самих регулярных выражениях, поэтому они экранируются"\".

r"[a-cA-C]{2}"

Соответствует строке из двух букв "a", "b" или "c". Например, "Ac", "CC", "bc".

r"aa|bb|cc|AA|BB|CC"

Строка из двух одинаковых букв.

r"([a-cA-C])\1"

Строка из двух одинаковых букв, но шаблон задан с использованием групп

r"aa|bb".

Соответствует "aa" или "bb"

r"a(a|b)b"

Соответствует "aab" или "abb"

r"^(?:\d{8}|\d{4}):\s\*(.\*)$"

Соответствует строке, которая начинается с набора из восьми или четырех цифр и двоеточия. Все, что идет после двоеточия и после следующих за ним пробелов, выделяется в группу с номером 1, тогда как набор цифр в группу не выделен.

r"(\w+)=.\*\b\1\b"

Слова слева и справа от знака равенства присутствуют. Операнд "\1" соответствует группе с номером 1, выделенной с помощью скобок.

r"(?P<var>\w+)=.\*\b(?P=var)\b"

То же самое, но теперь используется именованная группа var.

r"\bregular(?=\s+expression)".

Соответствует слову "regular" только в том случае, если за ним после пробелов следует "expression"

r"(?<=regular )expression"

Соответствует слову "expression", перед которым стоит "regular" и один пробел.