**参考** [**http://www.runoob.com/regexp/regexp-syntax.html**](http://www.runoob.com/regexp/regexp-syntax.html)

**正则表达式可视化**[**https://regexper.com/**](https://regexper.com/)

[**http://blog.csdn.net/laokaizzz/article/details/43342285**](http://blog.csdn.net/laokaizzz/article/details/43342285)

**为什么使用正则表达式？**

典型的搜索和替换操作要求您提供与预期的搜索结果匹配的确切文本。虽然这种技术对于对静态文本执行简单搜索和替换任务可能已经足够了，但它缺乏灵活性，若采用这种方法搜索动态文本，即使不是不可能，至少也会变得很困难。

通过使用正则表达式，可以：

* 测试字符串内的模式。  
  例如，可以测试输入字符串，以查看字符串内是否出现电话号码模式或信用卡号码模式。这称为数据验证。
* 替换文本。  
  可以使用正则表达式来识别文档中的特定文本，完全删除该文本或者用其他文本替换它。
* 基于模式匹配从字符串中提取子字符串。  
  可以查找文档内或输入域内特定的文本。

例如，您可能需要搜索整个网站，删除过时的材料，以及替换某些 HTML 格式标记。在这种情况下，可以使用正则表达式来确定在每个文件中是否出现该材料或该 HTML 格式标记。此过程将受影响的文件列表缩小到包含需要删除或更改的材料的那些文件。然后可以使用正则表达式来删除过时的材料。最后，可以使用正则表达式来搜索和替换标记。

# 正则表达式 - 语法

正则表达式是由普通字符（例如字符 a 到 z）以及特殊字符（称为"元字符"）组成的文字模式。

## 普通字符

普通字符包括没有显式指定为元字符的所有可打印和不可打印字符。这包括所有大写和小写字母、所有数字、所有标点符号和一些其他符号。

## 非打印字符

非打印字符也可以是正则表达式的组成部分。下表列出了表示非打印字符的转义序列：

|  |  |
| --- | --- |
| **字符** | **描述** |
| \cx | 匹配由x指明的控制字符。例如， \cM 匹配一个 Control-M 或回车符。x 的值必须为 A-Z 或 a-z 之一。否则，将 c 视为一个原义的 'c' 字符。 |
| \f | 匹配一个换页符。等价于 \x0c 和 \cL。 |
| \n | 匹配一个换行符。等价于 \x0a 和 \cJ。 |
| \r | 匹配一个回车符。等价于 \x0d 和 \cM。 |
| \s | 匹配任何空白字符，包括空格、制表符、换页符等等。等价于 [ \f\n\r\t\v]。 |
| \S | 匹配任何非空白字符。等价于 [^ \f\n\r\t\v]。 |
| \t | 匹配一个制表符。等价于 \x09 和 \cI。 |
| \v | 匹配一个垂直制表符。等价于 \x0b 和 \cK。 |

## 特殊字符

所谓特殊字符，就是一些有特殊含义的字符，如上面说的"\*.txt"中的\*，简单的说就是表示任何字符串的意思。如果要查找文件名中有\*的文件，则需要对\*进行转义，即在其前加一个\。ls \\*.txt。

许多元字符要求在试图匹配它们时特别对待。若要匹配这些特殊字符，必须首先使字符"转义"，即，将反斜杠字符 (\) 放在它们前面。下表列出了正则表达式中的特殊字符：

|  |  |
| --- | --- |
| **特别字符** | **描述** |
| $ | 匹配输入字符串的结尾位置。如果设置了 RegExp 对象的 Multiline 属性，则 $ 也匹配 '\n' 或 '\r'。要匹配 $ 字符本身，请使用 \$。 |
| ( ) | 标记一个子表达式的开始和结束位置。子表达式可以获取供以后使用。要匹配这些字符，请使用 \( 和 \)。 |
| \* | 匹配前面的子表达式零次或多次。要匹配 \* 字符，请使用 \\*。 |
| + | 匹配前面的子表达式一次或多次。要匹配 + 字符，请使用 \+。 |
| . | 匹配除换行符 \n 之外的任何单字符。要匹配 . ，请使用 \. 。 |
| [ | 标记一个中括号表达式的开始。要匹配 [，请使用 \[。 |
| ? | 匹配前面的子表达式零次或一次，或指明一个非贪婪限定符。要匹配 ? 字符，请使用 \?。 |
| \ | 将下一个字符标记为或特殊字符、或原义字符、或向后引用、或八进制转义符。例如， 'n' 匹配字符 'n'。'\n' 匹配换行符。序列 '\\' 匹配 "\"，而 '\(' 则匹配 "("。 |
| ^ | 匹配输入字符串的开始位置，除非在方括号表达式中使用，此时它表示不接受该字符集合。要匹配 ^ 字符本身，请使用 \^。 |
| { | 标记限定符表达式的开始。要匹配 {，请使用 \{。 |
| | | 指明两项之间的一个选择。要匹配 |，请使用 \|。 |

## 限定符

限定符用来指定正则表达式的一个给定组件必须要出现多少次才能满足匹配。有\*或+或?或{n}或{n,}或{n,m}共6种。

正则表达式的限定符有：

|  |  |
| --- | --- |
| **字符** | **描述** |
| \* | 匹配前面的子表达式零次或多次。例如，zo\* 能匹配 "z" 以及 "zoo"。\* 等价于{0,}。 |
| + | 匹配前面的子表达式一次或多次。例如，'zo+' 能匹配 "zo" 以及 "zoo"，但不能匹配 "z"。+ 等价于 {1,}。 |
| ? | 匹配前面的子表达式零次或一次。例如，"do(es)?" 可以匹配 "do" 或 "does" 中的"do" 。? 等价于 {0,1}。 |
| {n} | n 是一个非负整数。匹配确定的 n 次。例如，'o{2}' 不能匹配 "Bob" 中的 'o'，但是能匹配 "food" 中的两个 o。 |
| {n,} | n 是一个非负整数。至少匹配n 次。例如，'o{2,}' 不能匹配 "Bob" 中的 'o'，但能匹配 "foooood" 中的所有 o。'o{1,}' 等价于 'o+'。'o{0,}' 则等价于 'o\*'。 |
| {n,m} | m 和 n 均为非负整数，其中n <= m。最少匹配 n 次且最多匹配 m 次。例如，"o{1,3}" 将匹配 "fooooood" 中的前三个 o。'o{0,1}' 等价于 'o?'。请注意在逗号和两个数之间不能有空格。 |

匹配中文字符的正则表达式： [\u4e00-\u9fa5]  
评注：匹配中文还真是个头疼的事，有了这个表达式就好办了  
匹配双字节字符(包括汉字在内)：[^\x00-\xff]  
评注：可以用来计算字符串的长度（一个双字节字符长度计2，ASCII字符计1）  
匹配空白行的正则表达式：\n\s\*\r  
评注：可以用来删除空白行  
匹配HTML标记的正则表达式：<(\S\*?)[^>]\*>.\*?</>|<.\*? />  
评注：网上流传的版本太糟糕，上面这个也仅仅能匹配部分，对于复杂的嵌套标记依旧无能为力  
匹配首尾空白字符的正则表达式：^\s\*|\s\*$  
评注：可以用来删除行首行尾的空白字符(包括空格、制表符、换页符等等)，非常有用的表达式  
匹配Email地址的正则表达式：\w+([-+.]\w+)\*@\w+([-.]\w+)\*\.\w+([-.]\w+)\*  
评注：表单验证时很实用  
匹配网址URL的正则表达式：[a-zA-z]+://[^\s]\*  
评注：网上流传的版本功能很有限，上面这个基本可以满足需求  
匹配帐号是否合法(字母开头，允许5-16字节，允许字母数字下划线)：^[a-zA-Z][a-zA-Z0-9\_]$  
评注：表单验证时很实用  
匹配国内电话号码：\d-\d|\d-\d  
评注：匹配形式如 0511-4405222 或 021-87888822  
匹配腾讯QQ号：[1-9][0-9]  
评注：腾讯QQ号从10000开始  
匹配中国邮政编码：[1-9]\d(?!\d)  
评注：中国邮政编码为6位数字  
匹配身份证：\d|\d  
评注：中国的身份证为15位或18位  
匹配ip地址：\d+\.\d+\.\d+\.\d+  
评注：提取ip地址时有用  
匹配特定数字：  
^[1-9]\d\*$　 　 //匹配正整数  
^-[1-9]\d\*$ 　 //匹配负整数  
^-?[1-9]\d\*$　　 //匹配整数  
^[1-9]\d\*|0$　 //匹配非负整数（正整数 + 0）  
^-[1-9]\d\*|0$　　 //匹配非正整数（负整数 + 0）  
^[1-9]\d\*\.\d\*|0\.\d\*[1-9]\d\*$　　 //匹配正浮点数  
^-([1-9]\d\*\.\d\*|0\.\d\*[1-9]\d\*)$　 //匹配负浮点数  
^-?([1-9]\d\*\.\d\*|0\.\d\*[1-9]\d\*|0?\.0+|0)$　 //匹配浮点数  
^[1-9]\d\*\.\d\*|0\.\d\*[1-9]\d\*|0?\.0+|0$　　 //匹配非负浮点数（正浮点数 + 0）  
^(-([1-9]\d\*\.\d\*|0\.\d\*[1-9]\d\*))|0?\.0+|0$　　//匹配非正浮点数（负浮点数 + 0）  
评注：处理大量数据时有用，具体应用时注意修正  
匹配特定字符串：  
^[A-Za-z]+$　　//匹配由26个英文字母组成的字符串  
^[A-Z]+$　　//匹配由26个英文字母的大写组成的字符串  
^[a-z]+$　　//匹配由26个英文字母的小写组成的字符串  
^[A-Za-z0-9]+$　　//匹配由数字和26个英文字母组成的字符串  
^\w+$　　//匹配由数字、26个英文字母或者下划线组成的字符串