PYTHON
编程基础

正则表达式的基础语法

概述

- 通过正则表达式可以定义一些匹配规则，只要满足匹配规则即认为匹配成功，从而实现模糊匹配。
- 正则表达式中既可以包含普通字符，也可以包含由特殊字符指定的匹配模式。
- 在实际应用正则表达式进行匹配时，正则表达式中的普通字符需要做精确匹配，而特殊字符指定的匹配模式则对应了用于模糊匹配的规则。

部分匹配模式

匹配模式	描述
.（点）	匹配换行外的任一字符。例如，对于正则表达式“ab.c”，其与“abdc”和“ab1c”匹配，但与“acdb”、“abc”和“ab12c”不匹配。
^（插入符）	匹配字符串开头的若干字符。例如，对于正则表达式“^py”，其与“python”匹配，但与“puppy”不匹配。
$	匹配字符串末尾的若干字符。例如，对于正则表达式“py$”，其与“puppy”匹配，但与“python”不匹配。
[]	字符集合，对应位置可以是该集合中的任一字符。既可以依次指定每一个字符，如[0123456789]；也可以通过短横线“-”指定一个范围，如[0-9]。在字符序列前加^表示取反，如[^0-9]表示匹配不在0至9之间的字符

部分匹配模式

匹配模式	描述
*	匹配前一个模式0次或多次。例如，对于正则表达式“a[0-9]*c”，其与“ac”、“a0c”和“a01c”匹配，但与“abc”不匹配
+	匹配前一个模式1次或多次。例如，对于正则表达式“a[0-9]+c”，其与“a0c”和“a01c”匹配，但与“ac”和“abc”不匹配
?	匹配前一个模式0次或1次。例如，对于正则表达式“a[0-9]?c”，其与“ac”和“a0c”匹配，但与“a01c”和“abc”不匹配
{m}	匹配前一个模式m次。例如，对于正则表达式“a[0-9]{1}c”，其与“a0c”匹配，但与“ac”、“a01c”和“abc”不匹配
{m,n}	匹配前一个模式m至n次；省略n则匹配前一个模式m次至无限次。例如，对于正则表达式“a[0-9]{1,2}c”，其与“a0c”和“a01c”匹配，但与“ac”和“abc”不匹配

部分匹配模式

匹配模式	描述
|	“A|B”表示匹配A或B中的任一模式即可。例如，对于正则表达式“a[b|d]c”，其与“abc”和“adc”匹配，但与“ac”、“aac”和“abbc”不匹配。
(…)	用()括起来的内容表示一个分组。在匹配完成后，可以获取每个分组在字符串中匹配到的内容。例如，对于正则表达式“(.*?)abc”，其与“123abc456abc”匹配结果为“123”和“456”；而对于正则表达式“(.*)abc”，其与“123abc456abc”匹配结果为“123abc456”。“*?”与“*”的区别在于：“*?”每次匹配尽可能少的字符；而“*”每次会匹配尽可能多的字符
\	转义符，使后面一个字符改变原来的含义。例如，在正则表达式中要精确匹配字符$，则需要写成“\$”；要精确匹配字符^，则需要写成“\^”

特殊序列

正则表达式中还提供了特殊序列以表示特殊的含义，其由“\”和一个字符组成。“\”后面的字符可以是数字，也可以是部分英文字母。

特殊序列	描述
\number	number表示一个数字，\number用于引用同一编号的分组中的模式（分组编号从1开始）。例如，对于正则表达式“([0-9])abc\1”，其中的“\1”就表示引用第1个分组中的模式“[0-9]”，即等价于“([0-9])abc[0-9]”，匹配以一个数字开头、一个数字结尾、中间是abc的字符串
\A	匹配字符串开头的若干字符，同匹配模式中的^
\b	单词边界符，即\b两边的字符应该一个是非单词字符、另一个是单词字符，或者一个是单词字符、另一个是空字符（即字符串的开头或末尾）。例如，对于正则表达式“\bfoo\b”，其与“foo”、“foo.”、“(foo)”和“bar foo baz”匹配，但与“foobar”、“foo3”和“foo_bar”不匹配
\B	非单词边界符，与\b功能相反

特殊序列

特殊序列	描述
\d	匹配任一数字字符，等价于[0-9]
\D	与\d作用相反，匹配任一非数字字符，等价于[^0-9]
\s	匹配任一空白字符
\S	与\s作用相反，匹配任一非空白字符
\w	匹配包含数字和下划线在内的任一可能出现在单词中的字符
\W	与\w作用相反，即匹配\w不匹配的那些特殊字符
\Z	匹配字符串末尾的若干字符，同匹配模式中的$

特殊序列

提示

由于Python的字符串中使用“\”作为转义符，如果要在字符串中使用字符“\”，则需要写作“\\”。

因此，当进行“\bfoo\b”的匹配时，实际编写代码时要写作'\\bfoo\\b'，这样会造成代码编写时容易出错且代码可读性较差。

我们通常在用于表示正则表达式的字符串前加上一个字符r，使得后面的字符串忽略转义符。例如，对于字符串'\\bfoo\\b'，我们可以写作r'\bfoo\b'。