

## 唯心不易

好读书，不求甚解，稍稍看些历史

首页

Python

算法

机器学习

音频

管理

随笔 - 29 文章 - 0 评论 - 72

Python 正则表达式入门（中级篇）

# Python 正则表达式入门（中级篇）

初级篇链接：<http://www.cnblogs.com/chuxiuhong/p/5885073.html>

上一篇我们说在这一篇里，我们会介绍子表达式，向前向后查找，回溯引用。到这一篇开始前除了回溯引用在一些场合不可替代以外，大部分情况下的正则表达式你应该都会写了。

## 1.子表达式

子表达式的概念特别好理解。其实它就是将几个字符的组合形式看做一个大的“字符”。不好理解？举个栗子：我们要匹配类似IP地址这种形式的字符（暂且不考虑数值范围的合理性，这个留作学完之后的思考题

公告

昵称：唯心不易

园龄：1年6个月

粉丝：72

关注：2

+加关注

吧)。形如192.168.1.1这样的地址我们怎么写表达式呢？

```
答案一 \d+.\d+.\d+.\d+
```

不好，一个是太繁琐，另一个是连位数都控制不了

```
答案二 \d+{1,3}.\d+{1,3}.\d+{1,3}.\d+{1,3}
```

一般般，复杂但是起码能把位数控制在合理范围

```
答案三 (\d+{1,3}\.){3}\d+{1,3}\.
```

利用子表达式，将 123. 这种数字加小数点看做一个整体字符，对其规定重复匹配的次數，既简洁，效果又好。所以只要你将几个字符组合用圆括号括起来，那么你就可以把一个圆括号内的内容当做一个字符，外面可以加我们之前讲过的所有元字符来控制匹配。

## 2.向前向后查找

现在，我们终于来到了向前向后查找这一块。为什么说终于来到这了呢？还记得我们在初级篇最开始的例子吗？

```
假如你在写一个爬虫，你得到了一个网页的HTML源码。其中有一段html
<html><body><h1>hello world</h1></body></html>
你想要把这个hello world提取出来
```

<	2018年3月						>
日	一	二	三	四	五	六	
25	26	27	28	1	2	3	
4	5	6	7	8	9	10	
11	12	13	14	15	16	17	
18	19	20	21	22	23	24	
25	26	27	28	29	30	31	
1	2	3	4	5	6	7	

搜索

常用链接

[我的随笔](#)

[我的评论](#)

[我的参与](#)

[最新评论](#)

[我的标签](#)

我的标签

python(23)

```
import re

key = r"<html><body><h1>hello world</h1></body></html>"#这段是你要匹配的文本
p1 = r"(?<=<h1>).+?(?=</h1>)"#这是我们写的正则表达式规则，你现在可以不理解啥意思
pattern1 = re.compile(p1)#我们在编译这段正则表达式
matcher1 = re.search(pattern1, key)#在源文本中搜索符合正则表达式的部分
print matcher1.group(0)#打印出来
```

这个正则表达式

```
p1 = r"(?<=<h1>).+?(?=<h1>)"
```

看到 `(?<=<h1>)` 和 `(?=<h1>)` 了吗？第一个`?<=`表示在被匹配字符前必须得有 `<h1>`，后面的`?=`表示被匹配字符后必须有 `<h1>`

简单来说，就是你要匹配的字符是XX，但必须满足形式是AXXB这样的字符串，那么你就可以这样写正则表达式

```
p = r"(?<=A)XX(?=B)"
```

匹配到的字符串就是XX。并且，向前查找向后查找不需要必须同时出现。如果你愿意，可以只写满足一个条件。

所以你也不需要记住哪个是向前查找，哪个是向后查找。只要记住`?<=`后面跟着的是前缀要求，`?=`后面跟的是后缀要求。

[算法\(12\)](#)

[PyQt4\(6\)](#)

[机器学习\(3\)](#)

[网络\(2\)](#)

[音乐检索\(2\)](#)

[正则表达式\(2\)](#)

[django\(2\)](#)

[java\(2\)](#)

[kd树\(1\)](#)

[更多](#)

随笔分类

[PyQt入门学习笔记\(5\)](#)

随笔档案

[2018年1月 \(2\)](#)

[2017年10月 \(1\)](#)

[2017年8月 \(1\)](#)

[2017年5月 \(2\)](#)

[2017年4月 \(1\)](#)

[2016年12月 \(3\)](#)

[2016年11月 \(4\)](#)

[2016年10月 \(8\)](#)

[2016年9月 \(7\)](#)

最新评论

本质上来说，向前查找和向后查找其实是匹配整个字符串，即AXXB，但返回时仅仅返回一个XX。也就是说，如果你愿意，完全可以避开向前向后查找的方式，直接匹配带有前后缀的字符串，然后做字符串切片处理。

### 3.回溯引用

不同于前面的向前向后查找，这一条有时候你未必绕的过去。在有些情况下，你还必须得用到回溯引用，所以你如果想拥有在实际应用中使用正则表达式，回溯引用是你应该了解和掌握的。

我们还是从最开始的例子来说。

你原本要匹配 `<h1></h1>` 之间的内容，现在你知道HTML有多级标题，你想把每一级的标题内容都提取出来。你也许会这样写：

```
p = r"<h[1-6]>.*?</h[1-6]>"
```

这样一来，你就可以将HTML页面内所有的标题内容全部匹配出来。即 `<h1></h1>` 到 `<h6></h6>` 的内容都可以被提取出来。但是我们之前说过，写正则表达式困难的不是匹配到想要的内容，而是尽可能的不匹配到不想要的内容。在这个例子中，很有可能你就会被下面这样的用例玩坏。

比方说

```
<h1>hello world</h3>
```

发现后面的 `</h3>` 了吗？我们不管是怎么写出来这样的标题的，但实实在在的是我们的正则表达式同样

1. Re:Python 正则表达式入门（初级篇）

写的很好啊

--执小白

2. Re:银行家算法学习笔记

可以可以，简明易懂

--icelee

3. Re:卷积神经网络提取特征并用于SVM

list(map())就可以了，主要是python2和3的差别

--隐泊浮生

4. Re:卷积神经网络提取特征并用于SVM

@月明塘 遇到了同样的问题，请问你解决了吗？...

--隐泊浮生

5. Re:PyQt4入门学习笔记（一）

支持。对于了解PYQT的基本使用不错。

--豪门百里

#### 阅读排行榜

1. Python 正则表达式入门（初级篇）

(63754)

2. IDEA上安装和使用

checkstyle,findbugs,visualVM,PMD插件

(9863)

3. 用树莓派从零开始做一个家庭监控(9008)

4. 卷积神经网络提取特征并用于SVM(8826)

5. pycharm连接mysql数据库(8662)

会把这里的hello world匹配出来。这时候就是回溯引用的重要作用。下面就是一个示例：

```
import re

key = r"<h1>hello world</h3>"
p1 = r"<h([1-6])>.*?</h\1>"
pattern1 = re.compile(p1)
m1 = re.search(pattern1, key)
print m1.group(0) #这里是会报错的，因为匹配不到，你如果将源字符串改成</h1>
结尾就能看出效果
```

看到 `\1` 了吗？原本那个位置应该是 `[1-6]`，但是我们写的是`\1`，我们之前说过，转义符 `\` 干的活就是把特殊的字符转成一般的字符，把一般的字符转成特殊字符。普普通通的数字1被转移成什么了呢？在这里1表示第一个子表达式，也就是说，它是动态的，是随着前面第一个子表达式的匹配到的东西而变化的。比方说前面的子表达式内是 `[1-6]`，在实际字符串中找到了1，那么后面的`\1`就是1，如果前面的子表达式在实际字符串中找到了2，那么后面的`\1`就是2。

类似的，`\2`,`\3`,.....就代表第二个第三个子表达式。

所以回溯引用是正则表达式内的一个“动态”的正则表达式，让你根据实际的情况变化进行匹配。

中级篇就到这里，其实正则表达式还有很多细节还没有写出来，也有很多元字符我没有交代，但掌握了纲要，懂得原理之后剩下的就类似于查表构造这种活了。

建议看到这的朋友看看《正则表达式必知必会》，初级篇和这篇中有几个例子也是取材于此。

## 评论排行榜

1. 用树莓派从零开始做一个家庭监控(20)
2. 卷积神经网络提取特征并用于SVM(13)
3. 听歌识曲--用python实现一个音乐检索器(12)
4. Python 正则表达式入门（初级篇）(8)
5. 不到一百行实现一个命令词识别(7)

## 推荐排行榜

1. Python 正则表达式入门（初级篇）(14)
2. 不到一百行实现一个命令词识别(12)
3. 听歌识曲--用python实现一个音乐检索器(12)
4. 银行家算法学习笔记(7)
5. 用树莓派从零开始做一个家庭监控(7)

苍生苦难，不知伊于胡底

标签: python , 正则表达式

好文要顶

关注我

收藏该文







唯心不易

关注 - 2

粉丝 - 72

+加关注

1

« 上一篇: [PyQt4入门学习笔记（二）](#)  
» 下一篇: [基于傅里叶变换和PyQt4开发一个简单的频率计数器](#)  
posted @ 2016-09-25 23:09 唯心不易 阅读(7621) 评论(0) 编辑 收藏

[刷新评论](#) [刷新页面](#) [返回顶部](#)

发表评论

昵称:

评论内容:



提交评论

[退出](#) [订阅评论](#)

[Ctrl+Enter快捷键提交]

最新IT新闻:

- 专访寺库任冠军：纯奢侈品电商没有未来，不担心巨头入局
  - 谷歌大脑发布神经网络的「核磁共振」，并公开相关代码
  - Scale推出传感器融合标注API，为自动驾驶技术更快注入数据燃料
  - 出道三年，马化腾在鬼畜界终于迎来艺术巅峰
  - 3月7日这一夜，黑客耍了币圈的所有人
- » 更多新闻...

最新知识库文章:

- 写给自学者的入门指南
  - 和程序员谈恋爱
  - 学会学习
  - 优秀技术人的管理陷阱
  - 作为一个程序员，数学对你到底有多重要
- » 更多知识库文章...

---

Copyright ©2018 唯心不易

---