正则表达式

Regular Expression, 正则表达式, 一种使用表达式的方式对字符串进行匹配的语法规则.

我们抓取到的网页源代码本质上就是一个超长的字符串, 想从里面提取内容.用正则再合适不过了.

正则的优点: 速度快, 效率高, 准确性高 正则的缺点: 新手上手难度有点儿高.

不过只要掌握了正则编写的逻辑关系, 写出一个提取页面内容的正则其实并不复杂

正则的语法: 使用元字符进行排列组合用来匹配字符串 在线测试正则表达式https://tool.oschina.net/regex/

元字符: 具有固定含义的特殊符号 常用元字符:

```
1    .       匹配除换行符以外的任意字符
2    \w      匹配字母或数字或下划线
3    \s      匹配任意的空白符
4    \d      匹配数字
```

```
\n      匹配一个换行符
\t      匹配一个制表符

^       匹配字符串的开始
$       匹配字符串的结尾

\W      匹配非字母或数字或下划线
\D      匹配非数字
\S      匹配非空白符
a|b     匹配字符a或字符b
()      匹配括号内的表达式，也表示一个组
[...]       匹配字符组中的字符
[^...]      匹配除了字符组中字符的所有字符
```

量词: 控制前面的元字符出现的次数

```
*       重复零次或更多次
+       重复一次或更多次
?       重复零次或一次
{n}       重复n次
{n,}      重复n次或更多次
{n,m}       重复n到m次
```

贪婪匹配和惰性匹配

```
.*      贪婪匹配
.*?     惰性匹配
```

这两个要着重的说一下. 因为我们写爬虫用的最多的就是这个惰性匹配.

先看案例

```
str: 玩儿吃鸡游戏，晚上一起上游戏，干嘛呢? 打游戏啊
reg: 玩儿.*?游戏

此时匹配的是: 玩儿吃鸡游戏

reg: 玩儿.*游戏
此时匹配的是: 玩儿吃鸡游戏，晚上一起上游戏，干嘛呢? 打游戏


str: <div>胡辣汤</div>
reg: <.*>
结果: <div>胡辣汤</div>


str: <div>胡辣汤</div>
reg: <.*?>
结果:
```

```
18        <div>
19        </div>
20
21    str: <div>胡辣汤</div><span>饭团</span>
22    reg: <div>.*?</div>
23    结果:
24        <div>胡辣汤</div>
```

所以我们能发现这样一个规律: *.? 表示尽可能少的匹配, .*表示尽可能多的匹配, 暂时先记住这个规律. 后面写爬虫会用到的哦