小象学院

Parser

主　讲: 杨真

小象学院

Part 1 基础

- 环境搭建
- HTML 基础
- 第一个10行代码的爬虫
- 内容抽取及解析
- HTTP 协议
- POSTMAN 工具详解

Part 2 爬虫

- 网站结构分析
- 抓取方案
- 多线程并行及排重
- 用 MySQL 信息存储

Part 3 进阶

- 网站服务结构
- Cookie 及 登录处理
- 控制抓取的节奏
- 日志
- 守护进程

Part 4 实战

- 网站结构分析
- 网页抓取方案
- 数据提取
- 存储方案

课时大纲

- Regular Expression（正则表达式）
- LXML（基于DOM树的提取）

正则表达式

正则表达式是对字符串操作的一种逻辑公式，就是用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑

在爬虫的解析中，经常会将正则表达式与 Dom 选择器结合使用。正则表达式适用于字符串特征比较明显的情况，但是同样的正则表达可能在HTML源码里多次出现；而 Dom 选择器可以通过 class 及 id 来精确找到 DOM 块，从而缩小查找的范围

常用规则

\	转意符，例如 \?
^	字符串起始
$	字符串结束
*	匹配前面子表达式0次或多次
+	匹配前面子表达式1次或多次
?	匹配前面子表达式0次或1次
{n,m}	匹配至少n次，最多m次
.	匹配除 \n 之外的单个字符
(pattern)	匹配并获取这个匹配，例如匹配ab(cd)e正则表达式只返回 cd
[xyz]	字符集合，匹配任意集合里的字符
[^xyz]	排除集合里的字符，不能匹配
\d	匹配一个数字，等价 [0-9]

贪婪模式

？该字符紧跟在任何一个其他限制符（*,+,?，{n}，{n,}，{n,m}）后面时，匹配模式是非贪婪的。

非贪婪模式尽可能少的匹配所搜索的字符串

默认的贪婪模式则尽可能多的匹配所搜索的字符串

贪婪模式

```
<a
href='https://www.google.com/search?q=regular+expression&oq=regular+
expression&aqs=chrome..69i57.3239j0j7&sourceid=chrome&ie=UTF-
8'>Google Test 1</a><a
href='https://www.google.com/search?q=regular+expression&oq=regular+
expression&aqs=chrome..69i57.3239j0j7&sourceid=chrome&ie=UTF-
8'>Google Test 2</a>
```

对于上面的字符串，贪婪模式将匹配整个字符串，

```
re.findall('https://www.google.com/search\?q=.*UTF-8', s)
```

对于上面的字符串，而非贪婪模式才是我们想要的，只返回一个链接

```
re.findall('https://www.google.com/search\?q=.*?UTF-8', s)
```

小象学院

DOM 操作器 XPATH

简介

表达式	描述
nodename	选取此节点的所有子节点，tag 或 * 选择任意的tag
/	从根节点选取，选择直接子节点，不包含更小的后代（例如孙、从孙）
//	从当前路径选择文档中的节点，而不考虑它们的位置，包含所有后代
.	选取当前节点
..	选取当前节点的父节点
@	选取属性

属性选择

在DOM树，以路径的方式查询节点

通过 @ 符号来选取属性

```
<a rel="nofollow" class="external text" href="http://google.ac">google<wbr
/>.ac</a>
```

rel class href 都是属性，可以通过 "//*[@class='external text']" 来选取对应元素

= 符号要求属性完全匹配，可以用 contains 方法来部分匹配，例如

“//*[contains(@class, ‘external’)]” 可以匹配，而

"//*[@class='external']" 则不能

属性选择

and 和 or 运算符：

选择 p 或者 span 或者 h1 标签的元素

```
soup = tree.xpath('//td[@class="editor bbsDetailContainer"]//*[self::p or self::span
or self::h1]')
```

选择 class 为 editor 或者 tag 的元素

```
soup = tree.xpath('//td[@class="editor" or @class="tag"]')
```

小象学院

只为遇见明天更优秀的你！