



黑马程序员
www.itheima.com

传智播客旗下
高端IT教育品牌

数据提取方法




第二部分课程概要

1. 基础知识
2. **Json知识点复习**
3. 正则表达式的复习
4. **xpath和lxml**

数据提取

什么是数据提取？

简单的来说，数据提取就是从响应中获取我们想要的数据的过程




HuffPost

Lawsuit Accuses Donald Trump Of Illegally Destroying White House Records

HuffPost - 29 minutes ago

Auto-delete system aims to keep communications forever hidden from the public, a watchdog group claims. By Mary Papenfuss. Two watchdog groups have sued Donald Trump over White House records, accusing the president of illegally destroying ...




New York Times

Trump Says He Did Not Tape Comey Conversations

New York Times - 28 minutes ago

President Trump at a rally in Cedar Rapids, Iowa, on Wednesday. Credit Stephen Crowley/The New York Times. WASHINGTON - President Trump cleared up one of the capital's least suspenseful mysteries on Thursday, acknowledging that he did not record ...




CNN

Sniper hits ISIS target from over 2 miles away

CNN - 3 hours ago

Washington (CNN) A Canadian special operations sniper successfully hit an ISIS fighter from a record-breaking distance of more than two miles away while assisting Iraqi forces in the push to retake Mosul, according to Canadian Special Operations ...




ESPN

Bulls agree to send Jimmy Butler to Wolves; Zach LaVine, Kris Dunn to Chicago

ESPN - 1 hour ago

The Minnesota Timberwolves have an agreement in principle to acquire three-time All-Star Jimmy Butler from the Chicago Bulls, sources told ESPN's Marc Stein.



ABC News

Judges: 'Making a Murderer' confession improperly obtained

ABC News - 3 hours ago

FILE - In a Friday, March 3, 2006 file photo, Brendan Dassey is escorted out of a Manitowoc County Circuit courtroom, in Manitowoc, Wis.



sakila						
Info Tables Columns Indexes Triggers Views Stored Procedures Functions						
Name	Engine	V...	Row Format	Rows	Avg Row Length	Dat
actor	InnoDB	10	Compact	200	81	
address	InnoDB	10	Compact	603	163	
category	InnoDB	10	Compact	16	1024	
city	InnoDB	10	Compact	600	81	
country	InnoDB	10	Compact	109	150	
customer	InnoDB	10	Compact	599	136	
film	InnoDB	10	Compact	1000	196	
film_actor	InnoDB	10	Compact	5462	35	
film_category	InnoDB	10	Compact	1000	65	
film_text	InnoDB	10	Compact	1000	180	
inventory	InnoDB	10	Compact	4581	39	
language	InnoDB	10	Compact	6	2730	
payment	InnoDB	10	Compact	16086	98	
rental	InnoDB	10	Compact	16005	99	
staff	InnoDB	10	Compact	2	32768	
store	InnoDB	10	Compact	2	8192	

数据分类

非结构化的数据：html等

处理方法：正则表达式、xpath

结构化数据：json，xml等

处理方法：转化为python数据类型

```
{ "size": 10, "list": [ { "imgurl": "http://cms-bucket.nosdn.127.net/catchpic/e2/e2837ffc625829b0255c3499c4814094.jpg", "has_378", "time": "2017-06-23 10:34:03", "title": "美官员称朝再次测试火箭引擎 同日美拦截测试失败", "bucket.nosdn.127.net/catchpic/4/4d/4d2c124f3b9522ce465367d094538713.jpeg", "has_9379", "time": "2017-06-23 10:31:08", "title": "谈朝核谈南海也谈萨德 中美安全对话结束首日", "bucket.nosdn.127.net/catchpic/3/3c/3cf0634e3e70b598cad4aef7724512dd.jpg", "has_380", "time": "2017-06-23 10:29:35", "title": "炸掉 建国 清真寺 伊斯兰国 已穷途末路?", "bucket.nosdn.127.net/catchpic/c/c1/c18cd4480d8da18e427d395ac568bb80.jpg", "has_381", "time": "2017-06-23 10:28:33", "title": "韩新任外长与美国务卿通话 美称尊重韩 萨德", "bucket.nosdn.127.net/catchpic/d/d9/d9a6d6d901c1bcd683f8f4e5786fb2fe.jpg", "has_382", "time": "2017-06-23 10:24:24", "title": "美国务院:美国尚未决定禁止其公民赴朝旅游", "bucket.nosdn.127.net/d88d67225dfa4f8392153c9a3aa20b6120170623102150.png", "has_366", "time": "2017-06-23 10:21:54", "title": "美媒:一名美国男子被控向中国传递绝密文件", "bucket.nosdn.127.net/4c8731bcf052490bbbe4a98ad4098dc420170623101809.inea". "has
```

```
<!-- head:start -->
<meta charset="UTF-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge" />
<title>正则表达式re模块 | Python爬虫课程讲义</title>
<meta content="text/html; charset=utf-8" http-equiv="Content-Type">
<meta name="description" content="">
<meta name="generator" content="GitBook 2.6.7">
<meta name="author" content="BigCat">
```

```
<bookstore>
<book category="COOKING">
  <title lang="en">Everyday Italian</title>
  <author>Giada De Laurentiis</author>
  <year>2005</year>
  <price>30.00</price>
</book>
<book category="CHILDREN">
  <title lang="en">Harry Potter</title>
  <author>J K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
</book>
<book category="WEB">
  <title lang="en">Learning XML</title>
  <author>Erik T. Ray</author>
  <year>2003</year>
  <price>39.95</price>
</book>
</bookstore>
```

点击查看源网页

数据提取之JSON

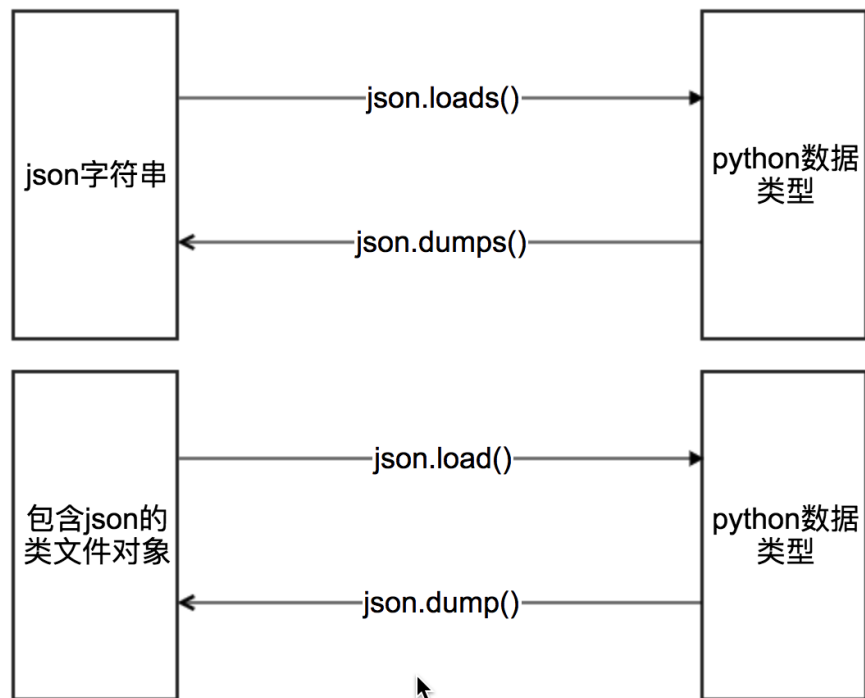
由于把json数据转化为python内建数据类型很简单，所以爬虫中，如果我们能够找到返回json数据的URL，就会尽量使用这种URL

JSON(JavaScript Object Notation) 是一种轻量级的数据交换格式，它使得人们很容易的进行阅读和编写。同时也方便了机器进行解析和生成。适用于进行数据交互的场景，比如网站前台与后台之间的数据交互。

那么问题来了：哪里能找到返回json的url呢？

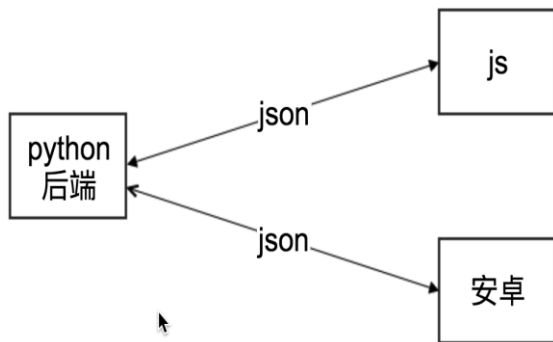
- 1、使用chrome切换到手机页面
- 2、抓包手机app的软件

数据提取之JSON



具有read()或者write()方法的对象就是类文件对象
f = open("a.txt","r") f就是类文件对象

数据提取之JSON



JSON	Python
object	dict
array	list
string	unicode
number (int)	int, long
number (real)	float
true	True
false	False
null	None

Json在数据交换中起到了一个载体的作用，承载着相互传递的数据

正则表达式复习

正则表达式的定义：

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。

常用正则表达式的方法：

`re.compile`（编译）

`pattern.match`（从头找一个）

`pattern.search`（找一个）

`pattern.findall`（找所有）

`pattern.sub`（替换）

正则表达式复习

语法	说明	表达式实例	完整匹配的字符串
字符			
一般字符	匹配自身	abc	abc
.	匹配任意除换行符"\n"外的字符。 在DOTALL模式中也能匹配换行符。	a.c	abc
\	转义字符，使后一个字符改变原来的意思。 如果字符串中有字符*需要匹配，可以使用*或者字符集[*]。	a\.c a\\c	a.c a\c
[...]	字符集（字符类）。对应的位置可以是字符集中任意字符。 字符集中的字符可以逐个列出，也可以给出范围，如[abc]或[a-c]。第一个字符如果是^则表示取反，如[^abc]表示不是abc的其他字符。 所有的特殊字符在字符集中都失去其原有的特殊含义。在字符集中如果要使用]、-或^，可以在前面加上反斜杠，或把]、-放在第一个字符，把^放在非第一个字符。	a[bcd]e	abe ace ade
预定义字符集（可以写在字符集[...]中）			
\d	数字：[0-9]	a\d c	a1c
\D	非数字：[^0-9]	a\D c	abc
\s	空白字符：[<空格>\t\r\n\f\v]	a\s c	a c
\S	非空白字符：[^0-9]	a\S c	abc
\w	单词字符：[A-Za-z0-9_]	a\w c	abc
\W	非单词字符：[^A-Za-z0-9_]	a\W c	a c
数量词（用在字符或(...)之后）			
*	匹配前一个字符0或无限次。	abc*	ab abccc
+	匹配前一个字符1次或无限次。	abc+	abc abccc
?	匹配前一个字符0次或1次。	abc?	ab abc
{m}	匹配前一个字符m次。	ab{2}c	abbc

python中原始字符串r的用法

```
In [45]: r'\nab'=="\\nab"  
Out[45]: True
```

r'\nab'不就是"\\nab" 么，为什么匹配不到呢？

```
In [43]: re.match(r"\nab", "\\nab").group()  
-----  
AttributeError                                Traceback (most recent call last)  
<ipython-input-43-e9a0eb35a667> in <module>()  
----> 1 re.match(r"\nab", "\\nab").group()  
  
AttributeError: 'NoneType' object has no attribute 'group'  
  
In [44]: re.match(r"\\nab", "\\nab").group()  
Out[44]: '\\nab'  
In [46]: re.match(r"\nab", "\nab").group()  
Out[46]: '\nab'
```

在python正则表达中**尽可能的使用原始字符串**，待匹配的字符串中看到什么就在正则表达式写什么，就不会出现问题

python中原始字符串r的用法

原始字符串定义(raw string): 所有的字符串都是直接按照字面的意思来使用，没有转义特殊或不能打印的字符，**原始字符串往往针对特殊字符而言。**

```
[In [2]: a = 'a\nb']  
  
[In [3]: a  
Out[3]: 'a\nb']  
  
[In [4]: len(a)  
Out[4]: 3  
  
[In [5]: print(a)  
a  
b
```

特殊符号，换行符

```
[In [8]: a = r"a\nb"]  
  
[In [9]: a  
Out[9]: 'a\\nb']  
  
[In [10]: len(a)  
Out[10]: 4  
  
[In [11]: print(a)  
a\nb
```

原始字符串，表示\n本身

'\n'长度为1，r'\n'长度为2

python中原始字符串r的用法

windows下不使用原始字符串会出现的问题：

```
In [14]: f = open("C:\Users\Frank\Desktop\test_r.txt", "r")
File "<ipython-input-14-023d1c11c893>", line 1
    f = open("C:\Users\Frank\Desktop\test_r.txt", "r")
          ^
SyntaxError: (unicode error) 'unicodeescape' codec can't decode bytes in position 2-3: truncated \UXXXXXXXX escape

In [15]: f = open(r"C:\Users\Frank\Desktop\test_r.txt", "r")

In [16]: f.read()
Out[16]: '123\nwege\nwqegw'
```

正则表达式复习

问题来了：

1. `re.compile` 该如何使用？
2. 如何非贪婪的去匹配内容？
3. `re.findall(r"a.*bc","a\nbc",re.DOTALL)` 和
`re.findall(r"a(.*)bc","a\nbc",re.DOTALL)` 的区别？
不分组时匹配的是全部，分组后匹配的是组内的内容

正则表达式爬虫练习

1. 通过正则表达式匹配内涵段子的段子
<http://neihanshequ.com/>
2. 获取<http://36kr.com/>网站首页的所有新闻

XPATH和LXML类库

为什么要学习XPATH和LXML类库：

lxml是一款高性能的 Python HTML/XML 解析器，我们可以利用 XPath，来快速的定位特定元素以及获取节点信息

什么是XPATH：

XPath (XML Path Language) 是一门在 HTML/XML 文档中查找信息的语言，可用在 HTML/XML 文档中对元素和属性进行遍历。

W3School官方文档：<http://www.w3school.com.cn/xpath/index.asp>

认识XML

数据格式	描述	设计目标
XML	Extensible Markup Language (可扩展标记语言)	被设计为传输和存储数据，其焦点是数据的内容。
HTML	HyperText Markup Language (超文本标记语言)	显示数据以及如何更好显示数据。

XML 的节点关系

节点的概念：每个XML的标签我们都称之为节点

节点

`<book>`

节点

`<title>Harry Potter</title>`

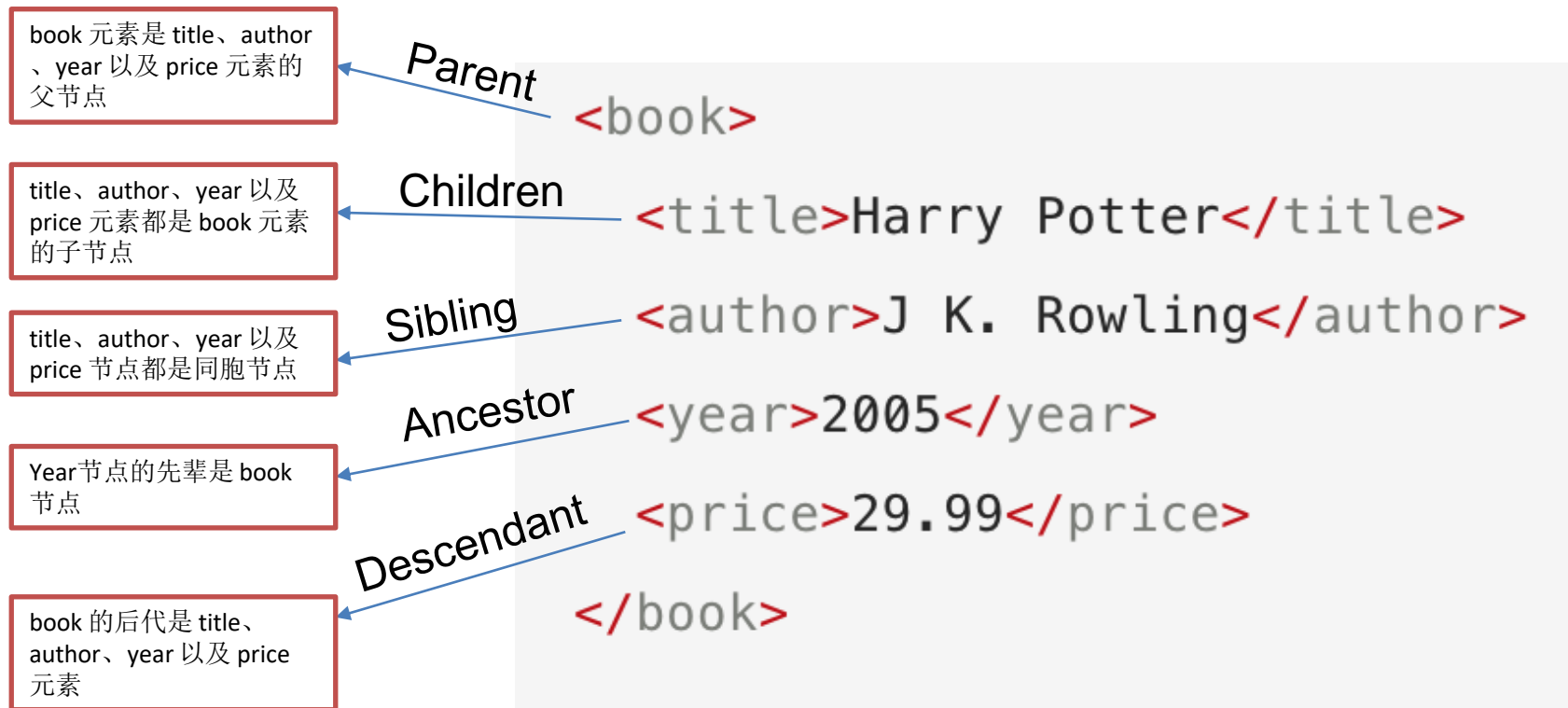
`<author>J K. Rowling</author>`

`<year>2005</year>`

`<price>29.99</price>`

`</book>`

XML 的节点关系



XPATH节点选择

常用节点选择工具：

- Chrome插件 XPath Helper
- 开源的XPath表达式编辑工具:XMLQuire(XML格式文件可用)
- Firefox插件 XPath Checker

节点选择语法

XPath 使用路径表达式来选取 XML 文档中的节点或者节点集。这些路径表达式和我们在常规的电脑文件系统中看到的表达式非常相似。

表达式	描述
nodename	选取此节点的所有子节点。
/	从根节点选取。
//	从匹配选择的当前节点选择文档中的节点，而不考虑它们的位置。
.	选取当前节点。
..	选取当前节点的父节点。
@	选取属性。

使用chrome插件选择标签时候，选中时，选中的标签会添加属性class="xh-highlight"

节点选择语法

查找某个特定的节点或者包含某个指定的值的节点

路径表达式	结果
/bookstore/book[1]	选取属于 bookstore 子元素的第一个 book 元素。
/bookstore/book[last()]	选取属于 bookstore 子元素的最后一个 book 元素。
/bookstore/book[last()-1]	选取属于 bookstore 子元素的倒数第二个 book 元素。
/bookstore/book[position()<3]	选取最前面的两个属于 bookstore 元素的子元素的 book 元素。
//title[@lang]	选取所有拥有名为 lang 的属性的 title 元素。
//title[@lang='eng']	选取所有 title 元素，且这些元素拥有值为 eng 的 lang 属性。
/bookstore/book[price>35.00]	选取 bookstore 元素的所有 book 元素，且其中的 price 元素的值须大于 35.00。
/bookstore/book[price>35.00]/title	选取 bookstore 元素中的 book 元素的所有 title 元素，且其中的 price 元素的值须大于 35.00。

节点选择语法

选择未知节点

通配符	描述
*	匹配任何元素节点。
@*	匹配任何属性节点。
node()	匹配任何类型的节点。

在下面的表格中，我们列出了一些路径表达式，以及这些表达式的结果：

路径表达式	结果
/bookstore/*	选取 bookstore 元素的所有子元素。
//*	选取文档中的所有元素。
html/node()/meta/@*	选择html下面任意节点下的meta节点的所有属性
//title[@*]	选取所有带有属性的 title 元素。

节点选择语法

选取若干路径

路径表达式	结果
<code>//book/title //book/price</code>	选取 book 元素的所有 title 和 price 元素。
<code>//title //price</code>	选取文档中的所有 title 和 price 元素。
<code>/bookstore/book/title //price</code>	选取属于 bookstore 元素的 book 元素的所有 title 元素，以及文档中所有的 price 元素。

xpath的更多语法:

[https://msdn.microsoft.com/zh-cn/library/ms256039\(v=vs.80\).aspx](https://msdn.microsoft.com/zh-cn/library/ms256039(v=vs.80).aspx)

lxml库

- 使用入门：
 - 导入lxml 的 etree 库
`from lxml import etree`
 - 利用etree.HTML，将字符串转化为Element对象
 - Element对象具有xpath的方法
`html = etree.HTML(text)`
- lxml 可以自动修正 html 代码

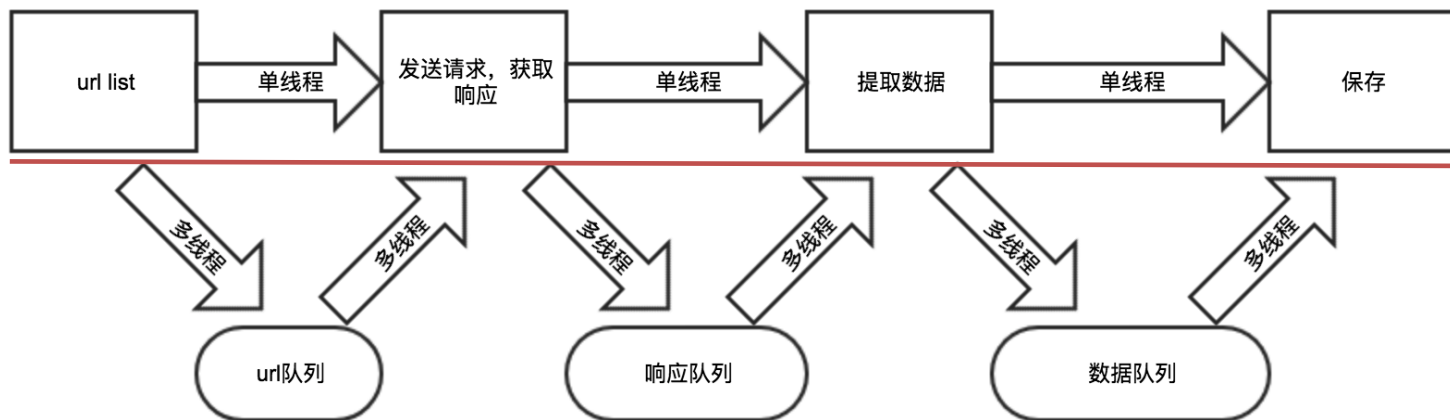
动手

用XPath来做一个简单的爬虫，爬取某个贴吧里的所有帖子，获取每个帖子的标题，连接和帖子中图片

动手

1、爬取糗事百科段子，页面的URL是
<http://www.qiushibaike.com/8hr/page/1>

2、动手把上述的爬虫改为多线程爬虫





Thank You!

改变中国 IT 教育，我们正在行动

www.itcast.cn