

Python语言基础与应用

基本扩展模块 / 网络爬虫

陈斌 北京大学 gischen@pku.edu.cn



网络爬虫

- › 搜索引擎蜘蛛
- › requests库
- › Beautiful Soup
- › 爬虫的基本流程

搜索引擎蜘蛛

- › 爬虫是按照一定规则，自动地提取并保存网页中信息的程序

蜘蛛沿着网络抓取猎物

通过一个节点之后，顺着该节点的连线继续爬行到下一个节点，最终爬完整个网络的全部节点

- › 通过向网站发起请求获取资源，提取其中有用的信息



requests库

- › **Python实现的一个简单易用的HTTP库**
支持HTTP持久连接和连接池、SSL证书验证、cookies处理、流式上传等
- › **向服务器发起请求并获取响应，完成访问网页的步骤**
- › **简洁、容易理解，是最友好的网络爬虫库**

requests库

› http请求类型

`requests.request()`: 构造一个请求

`requests.get()`: 获取HTML网页

`requests.head()`: 获取HTML网页头信息

`requests.post()`: 提交POST请求

`requests.put()`: 提交PUT请求

`requests.patch()`: 提交局部修改请求

`requests.delete()`: 提交删除请求

`requests.options()`: 获取http请求

› 返回的是一个response对象

requests库

› response对象

包含服务器返回的所有信息，例如状态码、编码形式、文本内容等；也包含请求的request信息

.status_code: HTTP请求的返回状态

.text: HTTP响应内容的字符串形式

.content: HTTP响应内容的二进制形式

.encoding: (从HTTP header中)分析响应内容的编码方式

.apparent_encoding: (从内容中)分析响应内容的编码方式

requests库

```
>>> import requests
>>> r = requests.get("http://www.baidu.com")
>>> r.status_code
200
>>> r.text
'<!DOCTYPE html>\r\n<!--STATUS OK--><html> <head><meta http-equiv=content-type content=tex
t/html; charset=utf-8><meta http-equiv=X-UA-Compatible content=IE=Edge><meta content=always
name=referrer><link rel=stylesheet type=text/css href=http://sl.bdstatic.com/r/www/cache/b
dorz/baidu.min.css><title>ç\x99%â°;ä_\x80ä_\x8bî\x8cä°\xa0â°±ç\x9fÿé\x81\x93</title></hea

>>> r.encoding
'ISO-8859-1'
>>> r.apparent_encoding
'utf-8'
>>> r.encoding = "utf-8"
>>> r.text
'<!DOCTYPE html>\r\n<!--STATUS OK--><html> <head><meta http-equiv=content-type content=tex
t/html; charset=utf-8><meta http-equiv=X-UA-Compatible content=IE=Edge><meta content=always
name=referrer><link rel=stylesheet type=text/css href=http://sl.bdstatic.com/r/www/cache/b
dorz/baidu.min.css><title>百度一下, 你就知道</title></head> <body link=#0000cc> <div id=wrap
```

requests库

› 定制请求头

requests的请求接口有一个名为headers的参数，向它传递一个字典来完成请求头定制

› 设置代理

一些网站设置了同一IP访问次数的限制，可以在发送请求时指定proxies参数来替换代理，解决这一问题

```
proxies = {  
    "http": "http://10.10.10.10:1010",  
    "https": "http://10.11.10.14:1011",  
}  
r = requests.get(url, proxies = proxies)
```


Beautiful Soup

› 页面解析器

使用requests库下载了网页并转换成字符串后，需要一个解析器来处理HTML和XML，解析页面格式，提取有用的信息

› 解析器类型

解析器	使用方法	优势
python标准库	BeautifulSoup(markup, "html.parser")	- Python的内置标准库 - 文档容错能力强
lxml HTML解析器	BeautifulSoup(markup, "lxml")	- 速度快 - 文档容错能力强
lxml XML解析器	BeautifulSoup(markup, ["lxml-xml"]) BeautifulSoup(markup, "xml")	- 速度快 - 唯一支持XML的解析器
Html5lib	BeautifulSoup(markup, "html5lib")	- 最好的容错性 - 以浏览器的方式解析文档 - 生成HTML5格式的文档

Beautiful Soup

› 搜索方法

`find_all(name, attrs, recursive, string, **kwargs)`

返回文档中符合条件的所有tag，是一个列表

`find(name, attrs, recursive, string, **kwargs)`

相当于`find_all()`中`limit = 1`，返回一个结果

name: 对标签名称的检索字符串

attrs: 对标签属性值的检索字符串

recursive: 是否对子节点全部检索，默认为True

string: `<>...</>` 中检索字符串

****kwargs:** 关键词参数列表

爬虫的基本流程

分析网页结构



爬虫的基本流程

爬取页面

通过requests库向目标站点发送请求，若对方服务器正常响应，能够收到一个response对象，它包含了服务器返回的所有信息

```
import requests
url = "http://news.qq.com/"
r = requests.get(url, timeout = 30)
print(r.text)
```

```
Python 3.6.2 Shell
File Edit Shell Debug Options Window Help
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <meta name="renderer" content="webkit" />
  <meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
  <title>新闻中心_腾讯网</title>
  <meta name="keywords" content="新闻 新闻中心 事实派 新闻频道,时事报道">
  <meta name="description" content="腾讯新闻,事实派。新闻中心,包含有时政新闻、国内新闻、国际新闻、社会新闻、时事评论、新闻图片、新闻专题、新闻论坛、军事、历史、的专业时事报道门户网站">
  <meta name="author" content="skeetershi" />
  <meta name="applicable-device" content="pc,mobile">
  <link rel="alternate" media="only screen and (max-width: 640px)" href="https://xw.qq.com/">
  <link rel="stylesheet" type="text/css" href="//mat1.gtimg.com/news/skeetershi/news_index/css/index.min_080409.css">
  <!-- 2016.7.13 jackiejiang add house.com use-->
  <script src="//mat1.gtimg.com/house/js/h5rewrite.js"></script>
  <!--2014.3.20 byAustinjin-->
  <!--20140225 menshen-->
  <script type="text/javascript" src="//js.aq.qq.com/js/aq_common.js"></script><!--[if !IE]>[xGv00]135a509a6ac85759a2a10161f645f1ba<![endif]-->
  <script>
    !function() {var i=function(i) {if(i&&i.indexOf("/a/")>0) {var t=i.split("/") [2].split("qq.com") [0].split(".");e=i.split("/a/") [1].split(".html") [0].replace(/[" 0-9]/g,""),n="",o="",s="";if(o=3==t.length?t[0]+t[1]:t[0],s=o.indexOf("_")>0&&o.split("_") [0].length<3?(o.split("_") [0]+o.split("_") [1]).substring(0,3).toUpperCase():o.substring(0,3).toUpperCase(),o.indexOf("house")>0&&(s="HOU"),o.indexOf("auto")>0&&(s="AUT"),/cd/fj|cq|hb|gd|zj|ln|js/.test(o)&&(s=o.toUpperCase()+"C"),/henan/.test(o)&&(s="HNC"),/xian/.test(o)&&(s="XAC"),/hn/.test(o)&&(s="HNN"),/2016/.test(o)&&(s="BCP"),undefined!=typeof o&&undefined!=typeof e) {if(o.split(".").length>2)for(var a=o.split(".").length,l=0;a>l;l++)n+=o.split(".")[l].a-2>l&&(n+="");else n=o.split(".")[0];return"https://xw.qq.com/"+o+"/"+s+"/"+t+e+"00"}},t=[news:"news",photo:"photo",video:"video",view:"huati",finance:"finance",ent:"ent",sports:"sports",digi:"digi",mil:"mil",lady:"lady",auto:"auto",games:"games",house:"house",astro:"astro",cul:"cul",fashion:"fashion",2014:"shi",2016:"2016",edu:"edu",nba:"nba",stock:"stock",health:"health",focus:"fx",lv:"visit",like:"shin
```


爬虫的基本流程

› 解析页面

- HTML代码-网页解析器
- Json数据-json模块, 转换成Json对象
- 二进制数据-以wb形式写入文件, 再做进一步处理

此处使用bs4进行解析

```
Python 3.6.2 Shell
File Edit Shell Debug Options Window Help
标题: 央视连视中美贸易摩擦: 最新清单的分析
链接: https://news.qq.com/a/20180719/000993.htm
标题: 国际快讯: 日本欧盟抱团逼自贸 这是逼美国的反?
链接: https://news.qq.com/omn/STO2018071900109600
标题: 未来房价能否平稳、房价会大涨? 多部门权威回应
链接: https://news.qq.com/omn/20180719/20180719A027F800
标题: 教育部: 参与购买、代写学位论文将开除学籍
链接: https://news.qq.com/a/20180719/001325.htm
标题: 北京监狱管理局: 李天一仍在服刑, 未减刑或假释
链接: http://ent.qq.com/a/20180718/040549.htm
标题: 葛农喝止小偷转杀凶犯致死判 其子: 更工作难起京
链接: https://news.qq.com/omn/20180719/20180719A00E1M00
标题: 巴西众议院弹劾总统案获通过 罗塞夫政变承认落败
链接: http://news.qq.com/a/20180418/023091.htm
标题: 国际快讯: 日本和欧盟抱团逼自贸 这是逼美国的反?
链接: http://news.qq.com/omn/20180719/20180719001096.html
标题: 中国科技公司赴美最疯狂上市潮, 小米之后美团拼多多蓄势待发
链接: http://news.qq.com/omn/20180716/20180716A1CB3H.html
标题: 新京报评论: 央行与财政部专家争鸣, 有莫大公共价值
链接: http://news.qq.com/zt/template/?id=FIN2018071702306600
标题: 北京监狱管理局: 李天一仍在服刑, 并未减刑或假释
链接: http://news.qq.com/omn/20180718/20180718040549.html
标题: 大公司要闻速览: 谷歌遭欧盟43亿欧元重罚, 阿里150亿投分众
链接: http://news.qq.com/omn/20180718/TBC2018071803687600

from bs4 import BeautifulSoup
soup = BeautifulSoup(r.text, 'lxml')
for news in soup.find_all('div', class_ = 'text'):
    info = news.find('a')
    if len(info) > 0:
        title = info.get_text()
        link = str(info.get('href'))
        print('标题: ' + title)
        print('链接: ' + link + '\n')
```