忍法:数据通灵术

杜亚磊

2017-02-19

爬虫步骤

- 下载网页

可能是最困难的部分，涉及登录和验证等。

- 解析网页

相对简单的过程，正则表达和Xpath足以应付各种情况。

- 存贮/分析

不在讨论之列。(:

软件工具

- 系统: Mac/Linux/Windows
- 浏览器: Chrome
- 编程语言: R/Python

网页

一个网页的源代码是包含HTML,CSS和JavaScript的纯文本。

- HTML: 标记语言，只有语法，没有变量和逻辑，不能称之为编程语言。
- CSS: 控制元素的展现形式
- JavaScript: 操作HTML中元素的增删改

一般来说，数据是在HTML元素中(否则你看不见它)。

HTML

HTML源码:

```
<html>
  <h1>Tital</h1>
  <p id="wth">What the hell?</p>
</html>
```

效果:

Tital

What the hell?

学习资料: HTML 教程

查看网页代码

网页: https://movie.douban.com/top250

你看到的网页

1

肖申克的救赎 / The Shawshank Redemption / 月黑高飞(港) / 刺激1995(台) [可播放]

导演: 弗兰克·德拉邦特 Frank Darabont 主演: 蒂姆·罗宾斯 Tim Robbins /...
1994 / 美国 / 犯罪 剧情

★★★★★ 9.6 788152人评价

“ 希望让人自由。 ”

查看网页代码

Chrome: `Ctrl(Command)` + `Alt(Option)` + `U`

丑陋的代码

```
<a href="https://movie.douban.com/subject/1292052/" class="">
    <span class="title">肖申克的救赎</span>
            <span class="title"> / The Shawshank Redemption</span>
        <span class="other"> / 月黑高飞(港)  /  刺激1995(台)</span>
</a>
```

下载网页

```
html_lines = readLines('https://movie.douban.com/top250')
doc = paste0(html_lines, collapse = '')
```

或者使用RCurl::getURL函数，Python中的标配是requests模块。

解析术之正则

```
title_lines = grep('class="title"', html_lines, value=T)
titles = gsub('.*>(.*?)<.*', '\\1', title_lines, perl=T)
```

Examples:

```
gsub('.*>(.*?)<.*', '\\1',
     '<span class="title">肖生克的救赎</span>', perl=T)
```

```
## [1] "肖生克的救赎"
```

学习资料:

[1] 正则表达式30分钟入门教程

[2] R语言中的正则表达式

[3] 正则表达式及R字符串处理之终结版

[4] Python中的re模块

解析术之XPath

```
library(xml2)
dom = read_html(doc)
title_nodes = xml_find_all(dom, './/span[@class="title"]')
xml_text(title_nodes)
```

```
[1] "肖申克的救赎"
[2] " / The Shawshank Redemption"
[3] "这个杀手不太冷"
[4] " / Léon"
[5] "霸王别姬"
[6] "阿甘正传"
......
```

学习资料: XPath 教程

解析术之CSS Selector

```
library(rvest)
read_html(doc) %>%
  html_nodes('.title') %>% # class="title"的标签
  html_text()
```

```
[1] "肖申克的救赎"
[2] " / The Shawshank Redemption"
[3] "这个杀手不太冷"
[4] " / Léon"
[5] "霸王别姬"
[6] "阿甘正传"
......
```

学习资料: CSS 元素选择器

你已经能抓取超过90%*的网站了

(Power: CSS Selectors ~= Xpath < Regular Expression)

[*] 随口一估计，不要太认真，它可能是99%

那还有什么问题?

那还有什么问题?

1.Ajax动态加载数据

那还有什么问题?

1.Ajax动态加载数据

2.数据分散在多个网页，难遍历

那还有什么问题?

1.Ajax动态加载数据

2.数据分散在多个网页，难遍历

3.登录验证

Ajax

赋予了网页动态变化的能力，无需重载页面即可获取并渲染新数据。

问题

数据通过Ajax加载后，通过JS添加进HTML中。直接下载网页得不到想要的数据！

学习资料: AJAX 教程

Ajax

解决办法

通过Chrome的Developer Tools寻找Ajax接口

因祸得福?

通常Ajax加载的数据都是JSON格式，爬到的数据方便解析

此处应有展示

https://wandergis.com/hospital-viz/index.htm

再来一个展示?

东风标致-经销商列表

http://dealer.peugeot.com.cn/

省级列表

```
library(rvest)
```

```
## Loading required package: xml2
```

```
get_options <- function(url){
  nodes = read_html(url) %>%
  html_nodes('option[value]')
  list(options = nodes %>% html_text(),
       values = nodes %>% html_attr('value'))
}
# get_options('http://www.peugeot.com.cn/')
```

```
[1] "北京市"          "天津市"          "河北省"          "山西省"
[1] "3361"            "3362"            "3363"            "3364"
```

城市列表

```
get_city <- function(pid){
  url = sprintf('http://dealer.peugeot.com.cn/ajax.php?pid=%s&action=
  get_options(url)
}
# get_city('3361')
```

```
$options
[1] "北京市"

$values
[1] "3392"
```

经销商列表

```
get_jxs <- function(cid){
  url = sprintf('http://dealer.peugeot.com.cn/ajax.php?cid=%s&action=
  get_options(url)
}
# get_city('3392')
```

```
$options
 [1] "北京鹏翰贸易有限公司"
 [2] "北京金泰凯迪汽车销售服务有限公司"
 ...

$values
 [1] "BJPHMYYXGS/"          "BJJTKDQCXSFWYXGS/"    "BJHJKMQCMYFZYXGS/"
 [4] "BJBLJJQCXSFWYXZRGS/" "BJJRYQCXSFWYXGS/"     "BJBRXZQCXSFWYXGS/"
 ...
```

经销商位置

```
get_location <- function(jid){
  url = 'http://dealer.peugeot.com.cn/dealer/%s'
  url = sprintf(url, jid)
  lines = readLines(url)
  line = grep('map.centerAndZoom', lines, value=T)
  gsub('.*BMap.Point((.*)),.*', '\\1', line, perl=T)
}
# get_location('BJPHMYYXGS')
```

```
(116.258447, 39.953778)
```

```
res = list()
main_list = get_options('http://www.peugeot.com.cn/')
for (i in 1:length(main_list)){          # 省列表
  province = main_list$options[i]
  pid = main_list$values[i]
  city_list = get_city(pid)

  for (i in 1:length(city_list)){        # 城市列表
    city = city_list$options[i]
    cid = city_list$values[i]
    jxs_list = get_jxs(cid)

    for (i in 1:length(jxs_list)) {      # 经销商列表
      jsx_name = jxs_list$options[i]
      jxs_id = jxs_list$values[i]
      loction = get_location(jxs_id)
      # save
      res = list(res, list(province=province,
                           city = city,
                           jsx_name = jsx_name,
                           loction = loction)
                 )
    }
  }
}
```

汇总

```
as.data.frame(do.call(rbind, res))
```

```
  province   city                                jsx_name                 loction
1   北京市 北京市                  北京鹏翰贸易有限公司 (116.258447, 39.953778)
2   北京市 北京市 北京金泰凯迪汽车销售服务有限公司 (116.425208, 40.101336)
......
```

新渠道:手机APP

利用Charles抓包。

- PC端开启代理服务
- 手机端通过PC代理上网
- 查看手机端请求

Show Charles

验证

与东风标致的类似，逐步分析每个请求，保存Cookie。

- 简单的登录，一个Post请求即可
- 复杂的登录，访问A页面获取Cookie/Token，带着Cookie/Token访问B,balabala...

花样很多，情况很复杂。

工具

1.phantomjs

2.pygoose

3.Selenium with Python 通过Chrome/FireFox浏览器打开网页，执行js文件。所得即所见。

- 使用css selector和Xpath进行元素筛选
- 模拟表单填写,按钮点击等操作
- 截图

4.curl & uncurl/curl2r

Selenium with Python: 案例

微信公众号抓取:

1. 通过Selenium调用FireFox登录web版微信: https://wx.qq.com/
2. 截图，扫码->登录
3. 点击切换到公众号面板
4. 搜集公众号数据，定时刷新页面. 如此循环...

curl & uncurl[1] / curl2r[2]案例

我的微博首页

http://weibo.com/u/2214298737/home?topnav=1&wvr=5

[1] 爬虫利器

[2] curl2r

江湖险恶，保护好自己的 Cookie

最后一问

限制频率怎么办?

限制频率怎么办?

慢点爬...

限制频率怎么办?

慢点爬...

切换代理访问

蟹蟹

本幻灯片由 R 包 **xaringan** 生成；

查克拉来自于 remark.js、**knitr**、以及 R Markdown。