

网络爬虫技术

Fibears

2016年3月29日

Outline

- 1. 什么是网络爬虫技术？
- 2. 掌握爬虫技术需要哪些基础知识？
- 3. 如何开发网页爬虫程序？
- 4. 介绍 rvest库(R) 和 urllib2库(Python)
- 5. 应用案例

什么是网络爬虫技术？

网络爬虫技术

- 网络爬虫，又被称为网页蜘蛛，网络机器人，它是指按照一定的规则，自动抓取万维网(World Wide Web)信息的程序或脚本。
- 网络爬虫的实质就是利用抽象化的程序模拟人登录网页的过程。

自动化抓取数据的一般逻辑

准备好对应的Http请求 (Http Request)

获得返回的响应 (Http Response)
获得Http Response中的网页源码

人访问网页

请求 (访问) —— 响应
(显示网页内容)

自动化抓取

通过程序模拟访问URL地址, 并获得其所返回的内容 (HTML 源码, Json格式的字符串等)



网页: 输入访问地址

网页: 显示相应内容

用途

- 利用爬虫程序批量地从网络中获取数据
- 利用爬虫程序模拟登录网页，比如快速预约讲座

那么，问题来了。。。

- 学习爬虫需要哪些基础知识呢？
- 当前有哪些实用的爬虫工具呢？
- 作为一名初学者，怎样才能快速掌握这门技术呢？

基础知识

- URL (Uniform Resource Locator)
- HTTP (HyperText Transfer Protocol)
- HTML (HyperText Markup Language)
- XPath
- COOKIE

统一资源定位符URL

- 简单说，URL就是所谓的网址。在万维网上，每一个信息资源都有形式统一且唯一存在的地址，该地址就叫URL。
- 它通常由三部分组成：模式或协议、服务器或IP地址、路径和文件名
- `http://www.baidu.com/`

超文本传输协议HTTP

- 超文本传输协议是互联网上应用最为广泛的一种网络协议。HTTP协议是基于请求响应模式的，客户端向服务器发送一个请求，服务器则以一个状态行作为响应，响应的内容通常为网页源码。
- HTTP协议中定义了操作资源的八种不同方法，其中最基本的方法有4种：GET，POST，PUT，DELETE，它们分别表示这个资源执行查、改、增、删4个操作。

沟通协议——语言

响应方



发起方

你吃了嘛？

吃过了。

你呢？

我也吃了。

发起方



响应方

沟通内容



超文本标记语言HTML

- 浏览网页时，我们通过浏览器所看到的网站是由HTML, CSS和JavaScript代码所构成的。HTML是一种建立网页文件的语言，它通过标记式的指令(Tag)，将影像、声音、图片、文字等信息显示出来。
- HTML语言使用标记对的方法编写文件，通常使用“<标志名>内容</标志名>”来表示标志的开始和结束。

HTML的常用标记符号

- `<html>...</html>`: 表示网页的开始与结束
- `<body>...</body>`: 网页的主体部分
- `<p>...</p>`: 创建一个段落
- `
...</br>`: 创建一个回车换行
- `<div>...</div>`: 用于排版大块的HTML段落
- `<h1></h1>...<h6></h6>`: 不同层级的标题
- ``: 处理图像, 将文件名赋值给标志对的src属性
- `...`: 链接标记符
- `...`: 创建一个新标签

- CSS是Cascading Style Sheets的英文缩写,即层叠样式表,它是一种标记语言,不需要经过编译过程,可以直接由浏览器执行。它主要用于美化网页,比如定义网页的背景颜色、字体的类型等。
- JavaScript是一种基于对象和事件驱动的客户端脚本语言。它可以实现Ajax异步请求过程、实现与用户之间的交互过程等。

XPath

- XPath是一门在XML文档中查找信息的语言，在XPath语境中，XML文档被视作节点树，节点树的根节点也被称作文档节点。XPath将节点树中的节点(Node)分为七类：元素、属性、文本、命名空间、处理指令、注释和文档节点。XPath使用路径表达式来选取XML文档中的节点或者节点集。
- 简单说，我们可以利用XPath从网页中提取出目标数据。

Cookie

- Cookie是指某些网站为了辨别用户身份、进行session跟踪而储存在用户本地终端上的数据。
- Cookie由服务器端生成，并发送给User-Agent（一般是浏览器），浏览器随后会将Cookie的信息保存到某个目录下的文本文件内，下次请求同一网站时就发送该Cookie给服务器。

所需的工具

爬虫工具：R & Python

- R语言中用于网页抓取的库有RCurl和rvest
- Python中用于网页抓取的库有urllib*和Requests
- R/Python+selenium2(webdriver+selenium): 主要用于处理动态网页
- Python: Pyspider框架、Scrapy框架

其他工具

- Firefox浏览器以及Firebug、HTTPFox插件
- XPath解析器和正则表达式

如何快速上手

- 首先，学习一门语言最重要的是带着目的去学习。当你有了一定的目标后，你在学习的过程中将会事半功倍。
- 其次，你需要先了解下该语言的一些基本语法，比如如何使用条件语句和循环语句、有哪几种常用的数据格式、如何读写数据。
- 接下来，你就可以一边阅读相关软件库的说明文档和案例，一边试着完成自己的项目。
- 最后就是勤练习、多提问、多思考。。。。

如何开发网络爬虫程序

- 爬虫程序的运行逻辑是模拟访问URL地址，然后获取服务器返回的响应文件(HTML源码或JSON格式的字符串)。
- 因此，在开发爬虫程序前，我们必须先搞清楚网页的运行逻辑，然后再一步步利用程序将其实现。

谢谢！