雪晴数据网
xueqing.tv

用RSelenium+PhantomJS打造灵活强大的网络爬虫

陈堰平

自我介绍

- 雪晴数据网创始人
- 北理工大数据创新学习中心业界导师
- 微软最有价值专家

北京数据科学联合教育中心

Python学习汇报会

与部分参与汇报的学生合影

学术活动

2017年9月10日知识图谱与智能问答会议

行业活动

微软数据科学沙龙：知识图谱行业应用

正式开始

关于网络爬虫

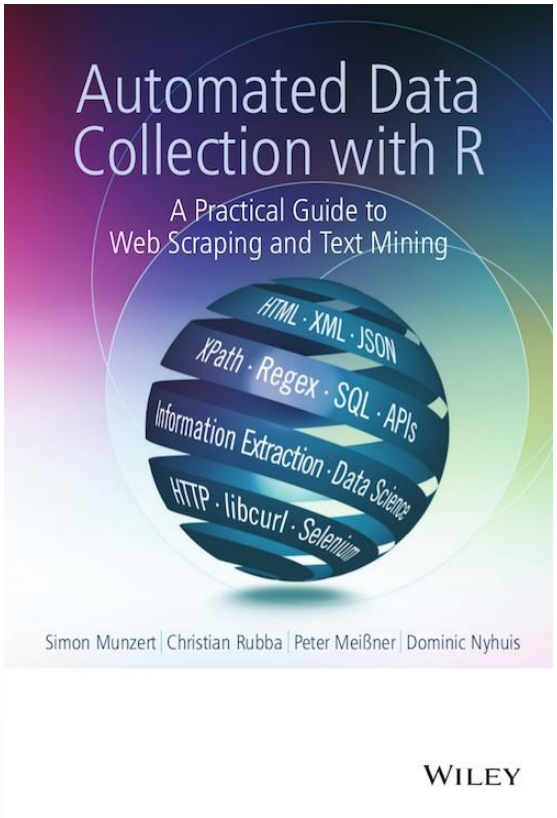
- 前端相关：html结构，css，javascript，ajax请求过程，h5，cookie，session
- 网络相关：request和response流程，http知识，代理proxy的使用
- 存储相关：sql，database，NoSQL，redis，文件读写
- 其他知识：Chrome调试，正则表达式，xpath，文件编码,分布式

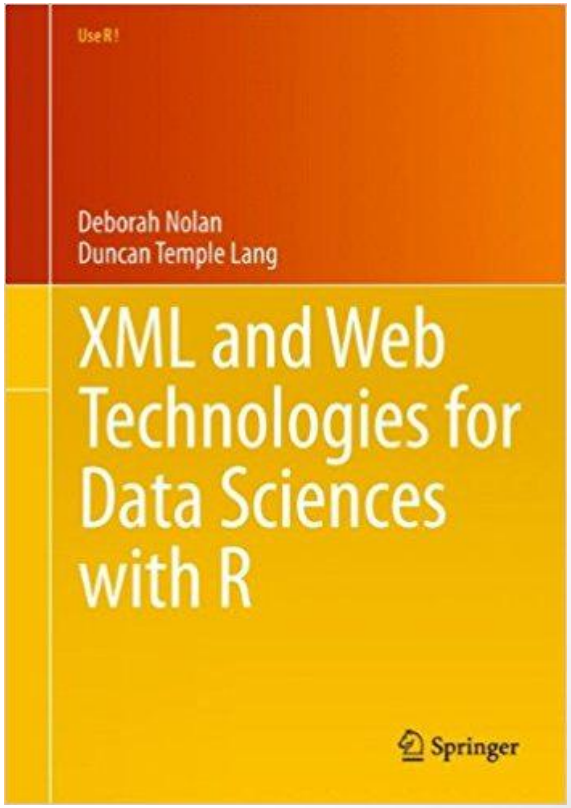
最少掌握什么？

- 前端相关：html
- 存储相关：文件读写
- 其他知识：Chrome调试，CSS选择器或xpath

跟爬虫有关的R包

- RCurl + XML
- rvest





如果数据不在网页代码

方法一：研究页面生成的过程，找到数据源的API

方法二：利用Selenium获取渲染后的页面代码

DEMO

http://www.cpppc.org:8082/efmisweb/ppp/projectLibrary/toPPPList.do

Selenium是什么？

- Selenium 是web自动化测试工具集
- Selenium支持：
  - Chrome, Firefox, Safari, Internet Explorer, PhantomJS 等浏览器
  - Windows, OS X, Linux, Android, iOS等操作系统
  - 多种编程语言
    http://docs.seleniumhq.org/download

Selenium有哪些工具？

- Selenium IDE 是firefox浏览器的一个插件。提供简单的脚本录制、编辑与回放功能
- Selenium Grid 是用来对测试脚步做分布式处理。现在已经集成到selenium server 中了。
- RC 和 WebDriver 更多应该把它看成一套规范，在这套规范里定义客户端脚步与浏览器交互的协议，以及元素定位与操作的接口。

关于RSelenium

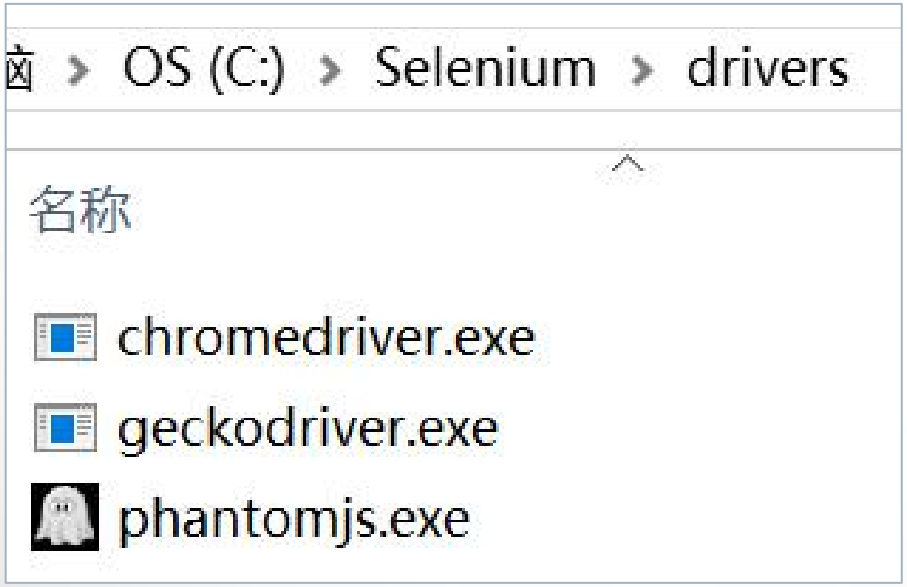
- 项目页面 https://github.com/johndharrison/RSelenium
- 问题反馈 https://github.com/johndharrison/RSelenium/issues?state=open
- CRAN主页 http://cran.r-project.org/web/packages/RSelenium

基本操作

- 初始化浏览器
- DOM交互
- 框架和窗口

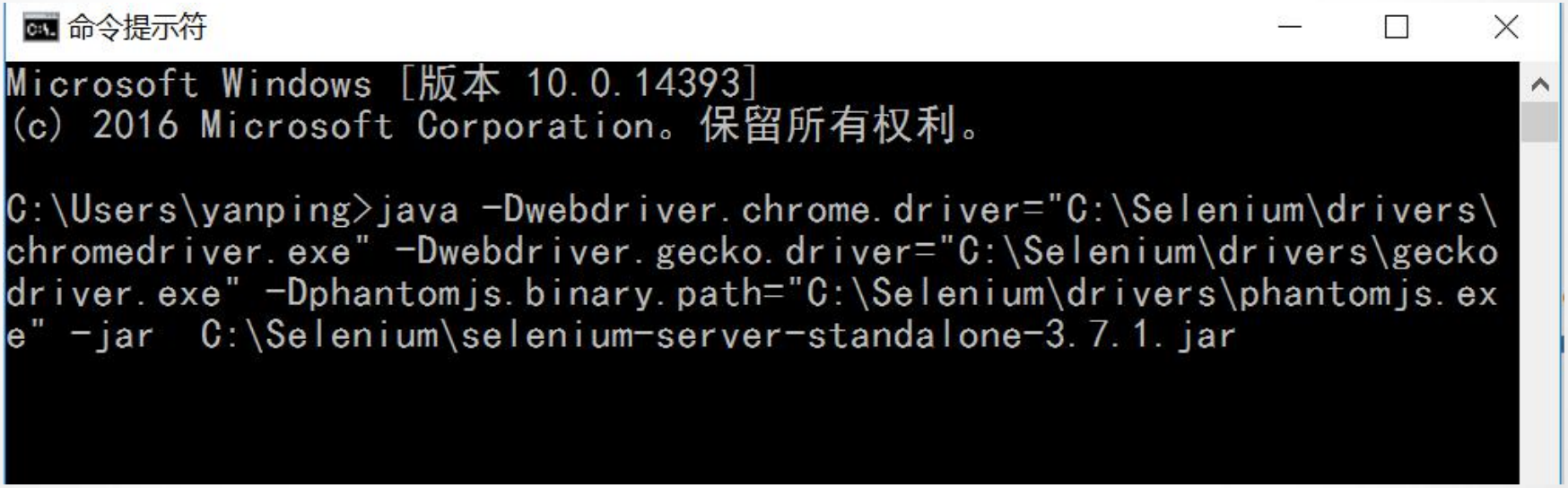
准备工作

- 下载好浏览器驱动
  - Chrome驱动 http://npm.taobao.org/mirrors/chromedriver
  - Firefox驱动 https://github.com/mozilla/geckodriver/releases
  - PhantomJS http://phantomjs.org/download.html



准备工作

- 启动Selenium Server



初始化浏览器

```
library(RSelenium)
remDr <- remoteDriver(
                    browserName = "chrome",
                    remoteServerAddr = ip,
                    port = port)
remDr$open(silent = TRUE)
remDr$navigate('http://baidu.com')
```

浏览页面

```
remDr$Navigate("http://xueqing.tv")
remDr$goBack()
remDr$goForward()
remDr$refresh()
remDr$getTitle()
remDr$getCurrentUrl()
remDr$getStatus()
remDr$getAllCookies()
```

DOM交互

- RSelenium有若干函数来寻找DOM（文档对象模型）元素和锚元素
  - id
  - class
  - css selector
  - name
  - tag name
  - link text (锚元素)
  - partial link text (锚元素)

高级操作

- Javascript注入
- 与shiny apps交互
- 远程驱动

Demo

- 全国PPP综合信息平台
- 通过关键词搜索微信公众号文章
- 航班历史信息的抓取
- 截图保存HTML5幻灯片
- 登录微博，输入验证码，搜索关键词

请关注我们的微信公众号，获取演讲视频和案例