R七种武器之网络爬虫RCurl 第2周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

关注炼数成金企业微信

- 提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

本周内容

- curl部分参数设置
- getBinaryURL()函数
- 批量下载的实现
- 结合XML，RCurl更强大

curl部分参数设置

- verbose：输出访问的交互信息
- httpheader：设置访问信息报头
- .encoding=“UTF-8”” GBK”
- debugfunction,headerfunction,curl
- .params：提交的参数组
- dirlistonly：仅读取目录

  ftp.wcc.nrcs.usda.gov/data/snow/snow_course/table/history/idaho/
- followlocation：支持重定向

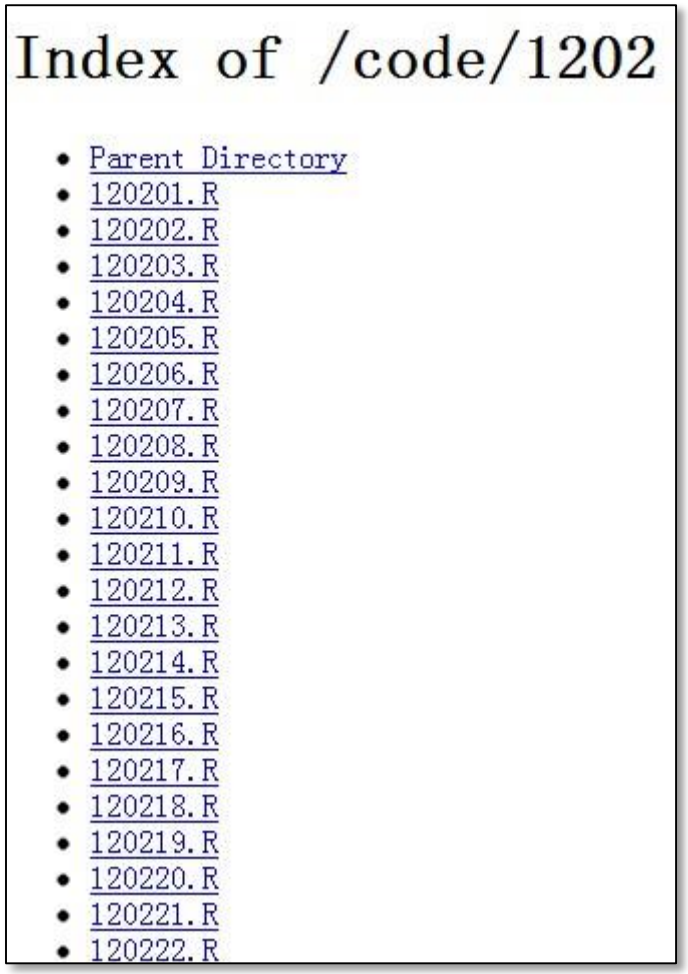
  http://www.sina.com
- maxredirs：最大重定向次数

getBinaryURL()下载文件

- temp<- getBinaryURL(url)
- note <- file("hellodata.xls", open = "wb")
- writeBin(temp,note)
- close(note)

批量下载文件的一个例子

- http://rfunction.com/code/1202/
- 所需函数：
- getURL()
- strsplit()
- lapply()
- paste()
- getBinaryURL()
- Sys.sleep()


Index of /code/1202

- Parent Directory
- 120201.R
- 120202.R
- 120203.R
- 120204.R
- 120205.R
- 120206.R
- 120207.R
- 120208.R
- 120209.R
- 120210.R
- 120211.R
- 120212.R
- 120213.R
- 120214.R
- 120215.R
- 120216.R
- 120217.R
- 120218.R
- 120219.R
- 120220.R
- 120221.R
- 120222.R

XML简介

- 网页解析工具

- 表格
- 网页节点
- 对标准 XML 文件的解析函数 xmlParse
- 对html的解析函数 htmlTreeParse

- 缺点：windows下对中文的支持不理想

下载表格

- http://www.bioguo.org/AnimalTFDB/BrowseAllTF.php?spe=Mus_musculus

- wp <- getURL(url)
- doc <- htmlParse(wp, asText = TRUE)
- tables <- readHTMLTable(doc)

```
$table1
      No.          Ensembl ID   Gene ID           Symbol                 Family
1       1 ENSMUSG00000029313      17355             Aff1                   AF-4
2       2 ENSMUSG00000031189      14266             Aff2                   AF-4
3       3 ENSMUSG00000037138      16764             Aff3                   AF-4
4       4 ENSMUSG00000049470      93736             Aff4                   AF-4
5       5 ENSMUSG00000046532      11835               Ar     Androgen receptor
6       6 ENSMUSG00000021359      21418          Tcfap2a                   AP-2
7       7 ENSMUSG00000025927      21419          Tcfap2b                   AP-2
8       8 ENSMUSG00000028640      21420          Tcfap2c                   AP-2
9       9 ENSMUSG00000042477     332937          Tcfap2e                   AP-2
```

抓取地震数据

- 国家地震科学数据共享中心
- http://data.earthquake.cn/datashare/datashare_more_quickdata_new.jsp

- wp <- getURL(url)
- doc <-htmlParse(wp, asText = TRUE)
- tables <- readHTMLTable(doc,header=F)
- 参数which

XPath

- 斜杠（/）作为路径内部的分割符。
- /：表示选择根节点
- //：表示选择任意位置的某个节点
- @： 表示选择某个属性
- *表示匹配任何元素节点。
- @*表示匹配任何属性值。
- node()表示匹配任何类型的节点。

试一试

- http://www.w3school.com.cn/example/xmle/books.xml

- doc <- xmlParse(url)
- 添加谓语条件
- getNodeSet(doc,'/bookstore/book[1]')
- ...
- 多个并列路径
- getNodeSet(doc,"//book/title | //book/price")
- ...

抓取电影团购

- http://t.dianping.com/guangzhou?q=%E7%94%B5%E5%BD%B1

- getURL()
- htmlParse()
- getNodeSet()
- sapply()
- paste()

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间