ruby爬虫初探

2021-10-18 09:50:01 +0800

***写作本文的目的借博主关注的作家岳南的新浪博客介绍Ruby爬虫的使用。***

## 借助交互Ruby控制台改善[1]

Mechanize为编写爬取工具提供了强大的基础，有两个原因：其一，便于发送HTTP请求；其二，为搜索远程文档提供了强大的句法。我们已经见识到如何使用Mechanize轻松发送GET请求，下面探索如何使用它过滤大量文档，获取重要文本内容。我们可以在Ruby IRB（交互式Ruby shell）中手动探索爬取过程。

$ irb -r./scraper  
2.7.2 :001 > scraper=Scraper.new  
 => #<Scraper:0x090e36d0 @root="http://blog.sina.com.cn/s/articlelist\_1311...   
2.7.2 :002 > page=scraper.agent.get "#{scraper.root}#{1}.html"  
 =>   
#<Mechanize::Page  
...   
2.7.2 :003 > items=page / "a[title=\"\"]"  
 => [#<Nokogiri::XML::Element:0x6cca name="a" attributes=[#<Nokogiri::XML:...   
2.7.2 :004 > items[0].text  
 => "西南联大缘何大师辈出？"   
2.7.2 :005 > items[0].values[2]  
 => "http://blog.sina.com.cn/s/blog\_4e3308af01032aht.html"   
2.7.2 :006 > page1=scraper.agent.get(items[0].values[2])   
 =>   
#<Mechanize::Page  
...   
2.7.2 :007 > body = page1 / "div[id=sina\_keyword\_ad\_area2]"  
 => [#<Nokogiri::XML::Element:0xac08 name="div" attributes=[#<Nokogiri::XM...   
2.7.2 :008 > body.text.strip

第一行启动IRB,并使用-r开关加载当前目录中的爬虫库。如果你以前没用过IRB,有几件事要知道，了解这些事之后，你的生活会变得更轻松。IRB中有提示符，指明使用Ruby的版本和运行命令的序号。IRB的功能众多，这里所涉极少。命令序号可用于重放历史，还能用于控制作业，这与很多其他类型的shell相似。在IRB提示符后面可以输入Ruby代码，irb会立即执行，然后输出结果，在=>符号后面打印返回值。在试验Ruby的过程中，经常会遇到复杂的返回值，例如scraper.agent.get的返回值是一个Mechanize Ruby对象。这个对象很大，打印时要消耗大量资源。这里，我们简略了返回值的大部分。为了节省空间，以后讨论IRB时对复杂对象依然如此。上述命令中第二行将HTTP GET请求保存为一个page对象。获得这个对象之后如何从中提取信息呢？Mechanize提供了一个精巧的语法糖（syntatic sugar）　,搜索DOM结构特别方便，即“/”运算符。如第三行代码中如果查询找到了内容，那么将返回一个由Mechanize对象组成的数组。这里它将返回页面代码中 <a titile=　.../>”既“作家岳南”的博客相关网页内容。第四行代码显示其作家第一篇文章标题。第五行显示作家第一篇文章链接地址。第六行代码接着爬取第五行链接相关的html代码。第七行搜索正文相关的html位置。第八行显示正文内容并去皮。 [代码详见][scraper.rb]　“run.rb“　运行结果显示为：

作家岳南的新浪博客目录（截至　2021 Oct 8th）

西南联大缘何大师辈出？  
觉醒与混饨的年代  
1948：多少黄金与国宝运往台湾  
一段飘零的岁月一批远去的大师  
文人写就的文人的历史  
大师之后再无大师  
一位走过白银时代的骑士  
徐志摩该不该烧死？  
南渡北归：大师远去，再无大师  
王侯将相，就是有种  
李霖灿：从圣地归来  
秦始皇帝陵是否被盗？  
杨振宁晚年重返清华园的隐秘  
“九一八”夜晚,影星胡蝶在哪里？  
清华校长梅贻琦，一个传奇的存在  
抗战胜利，首都为何设在北平  
我的写作缘起  
在彼此忧患的眼睛里  
诗人穆旦与远征军  
地雷战：知识者未被引爆的命运  
微斯人，吾谁与归？  
一个遥远美丽的地方  
胡适：考上大学选啥专业最好？  
大师之后还有无大师？  
贺绍俊：重述中国现代知识分子的心…  
你是一部清华史  
梁鸿鹰：触摸一代文化英才的精神脉…  
西南联合大学的定胜糕与耗子肉  
为何影片与样板戏中的反面人物都姓…  
卢沟桥事变真相  
七七忆南渡  
1909：庚子赔款第一期留美生  
克林顿会见陕西老杨引发的官司  
何兆武点评西南联大教授  
大师远去，星辉依旧  
为什么这么多人关注西南联大？  
《南渡北归》:一部可以留下来的文…  
四十年前高考往事  
111年前京城的一场考试  
五台山佛光寺是如何发现的？  
张东荪教授一家的悲剧  
“护国娘娘”赛金花  
金岳霖晚年最后一位女友  
老铁匠砸壶  
诗歌：《南渡北归》增订本读后  
毛泽东与傅斯年夜谈北大五四运动  
陈总司令的位子该放在哪里？  
南渡北归永离别  
清华大学今犹在，不见当年梅贻琦  
庆祝清华大学建校110周年  
From 2021-04-22 14:18 to the 2021-11-11 17:14 作家岳南 have writen 50 articles 106856 characters about 2137 characters per article  
.................  
.................  
  
考古纪实会终结盗墓文学吗  
[转载]魂牵梦绕诺奖梦：祝贺莫言荣…  
辅仁－－台湾最好的私立大学  
[转载]读者是不能欺骗的  
大学排名涵有什么信息  
2012年兩岸清華大學學術硏討會結語  
天涯也有江南信，梅破知春近。  
中秋节前，大陆作家代表团来台访问  
中国考古探秘丛书写作缘起与新版  
拒人门外的“黑大个儿”  
[转载]格言摘录  
《如果我的心是一朵莲花——林徽因时…  
国立清华大学毕业生  
国立清华大学校長 陳力俊 博士  
国立清华大学迎来诺贝尔奖  
昆明市經貿文化參訪團访问台湾清华…  
江苏赠新竹清华大学图书 邀台生赴…  
[转载]“花园村”“新貌”  
[转载]老家  东河  
憶清華名師演講會致詞  
第三十五屆全國比較文學會議  
清華大學清華建校一百零一年  
遠方的祝福 - 卓以玉的詩與畫  
[转载]客人造访！！  
岳南：严肃的文学更有营养  
台北中正纪念堂巡礼  
2012年04月01日  
[转载]孙萌的诗：2011年5月2日在敦…  
[转载]孙萌的诗（9首）  
[转载]《不为繁华易素心.民国文人…  
[转载]《不为繁华易素心.民国文人…  
[转载]新书《不为繁华易素心.民国…  
2012国窖1573 封藏大典  
[转载]细数西拐庄之最  
[转载]突飞猛进、漂亮大气的诸城市…  
春节家乡行  
國立清華大學校長陳力俊院士  
12.28梁任公來臺百年紀念會  
同学发来的名信片  
驻校台湾国立清华大学  
作家富了，文学“富”了没有？  
[转载]2011第六届中国作家富豪榜震…  
文學的寂寞與期望  
《南渡北归》装帧谈屑  
[转载]12元人民币在中国旅游能干什…  
[转载]岳南老师与华文粉丝  
民族使命感成就大师  
[转载]带时钟的日历，很实用  
岳南谈新作《南渡北归》再现大师风…  
[转载]学者岳南印象

## 爬取指定关键字文章

岳南是一位从事考古纪实文学写作的作家，其作品《风雪定陵》《日暮东陵》《万世法门》　《银雀山竹简》《南渡北归－三部曲》都是我们耳熟悉能详的作品，他的新浪博客从2011十月开通至今已经十年，累计博客近400篇，几十万字，从程序可以分析出它的写作字数从最初的平均八百字到现在的平均两千字增长，这和他多年累积的读书量有关，这可以从他会复粉丝的文字中找到佐证。匆匆浏览一番博客目录不难发现作家当前的写作重点－－既近代民国学术大师们的治学经历（胡适、傅思年、陈寅恪、…..）及他们在特殊岁月中为中国学术氛围留下的宝贵遗产。虽然在外人看来有些孤芳自赏，但是作为学者内行们因该能从中体味获得许多治学的方法与途径。相信作为一个有志于学习甚至是文学写作的人会从他的文章中获取不少的灵感。下面这段代码可以通过关键字正则匹配而快速爬取需要浏览的文章主体：

正则搜索关键字

def run   
 scrape()  
 character =0  
 articles =0  
 @pages.each do |page|   
 rows =(page /"a[title=\"\"]") #爬取博文目录  
 finds = rows.grep(/东陵/) #正则匹配关键字  
 finds.each do |find|  
 puts find.text  
 @agent.get(find.values[2]) do |page1|  
 body =(page1 /"div[id=sina\_keyword\_ad\_area2]")#爬取博文正文  
 puts body.text.strip.gsub(/\n\n/,'')  
 end  
 end  
 end  
end

通过输入感兴趣的文字既可

# 参考文献

[1] 安道. GitHub实践[M]. 北京: 人民邮电出版社, 2017.