搜博主文章

使用java开源工具httpClient及jsoup抓取解析网页数据

转载 2016年04月13日 09:52:04 标签:httpClient / jsoup

3257



使用java开源工具httpClient及jsoup抓取解析网页数据

今天做项目的时候遇到这样一个需求,需要在网页上展示今日黄历信息,数据格式如下

 $\overline{\odot}$

抓取。

公历时间: 2016年04月11日 星期一

• 农历时间:猴年三月初五

• 天干地支: 丙申年 壬辰月 癸亥日

人 1 地文:內中午 工成方 关次口

• 宜:求子 祈福 开光 祭祀 安床

忌:玉堂(黄道)危日,忌出行

主要包括公历/农历日期,以及忌宜信息的等。但是手里并没有现成的数据可供使用,怎么办呢? 革命前辈曾经说过,没有枪,没有炮,敌(wang)人(luo)给我们造!网络上有很多现成的在线 万年历应用可供使用,虽然没有现成接口,但是我们可以伸出手来,自己去拿。也就是所谓的数据

这里介绍两个使用的工具, httpClient以及jsoup,简介如下:

HttpClient是Apache Jakarta Common下的子项目,用来提供高效的、最新的、功能丰富的支持HTTP协议的客户端编程工具包,并且它支持HTTP协议最新的版本和建议。HttpClient已经应用在很多的项目中,比如Apache Jakarta上很著名的另外两个开源项目Cactus和HTMLUnit都使用了HttpClient。

httpClient使用方法如下:

- 1. 创建HttpClient对象。
- 2. 创建请求方法的实例,并指定请求URL。
- 3. 调用HttpClient对象的execute(HttpUriRequest request)发送请求,该方法返回一个HttpResponse。
- 4. 调用HttpResponse相关方法获取相应内容。
- 5. 释放连接。

jsoup 是一款 Java 的 HTML 解析器,可直接解析某个 URL 地址、HTML 文本内容。它提供了一套非常省力的 API,可通过 DOM,CSS 以及类似于 jQuery 的操作方法来取出和操作数据。

需要更多信息可以参见官网

httpClient:http://hc.apache.org/httpcomponents-client-5.0.x/index.html jsoup:http://jsoup.org/

接下来我们直接上代码,这里我们抓取2345在线万年历的数据 http://tools.2345.com/rili.htm 首先我们定义一个实体类Almanac来存储黄历数据

Almanac.java

复制代码

```
1 package com.likx.picker.util.bean;
2
3 /**
4 * 万年历工具实体类
5 *
6 * @author 溯源blog
7 * 2016年4月11日
8 */
9 public class Almanac {
```

加入CSDN,享受更精准的内容推荐,与500万程序员共同成长!



联系我们



请扫描二维码联系客服

webmaster@csdn.net

2 400-660-0108

♣ QQ客服 ●客服论坛

关于 招聘 广告服务 協百度 ◎1999-2018 CSDN版权所有 京ICP证09002463号

经营性网站备案信息

网络110报警服务

中国互联网举报中心

北京互联网违法和不良信息举报中心

博主最新文章

更多文章

在Dubbo中开发REST风格的远程调用(R ESTful Remoting)

JavaScript防http劫持与XSS

无源码程序反编译修改文字

安全框架Shiro和Spring Security比较

非常详细的 Docker 学习笔记

文章分类

java	31篇
架构设计	5篇
javascript	19篇
数据库	25篇
jquery	8篇
服务器	22篇
展工	T 🗸

文章存档

2018年4月		1篇
2018年3月		1篇
20103/3		工/冊
2018年2月		5篇
2017年10月		C ***
2017—10/7		6篇
2017年9月		2篇
2017年8月		4篇
1 -73	ロエン コード	• /115
	展开~	

博主热门文章

几种常见的载入中、loading页面效果的 实现方法总结

27490

零基础写java网络爬虫

21541

Java 日期转毫秒和毫秒转日期

15208

IntelliJ Idea编译报错,解决方法

/* 宜e.g. 求子 祈福 开光 祭祀 安床*/

/* 忌 e.g. 玉堂(黄道)危日,忌出行*/

```
2018/4/15
    13
            private String should;
            private String avoid;
    14
    15
    16
            public String getSolar() {
    17
    18
            }
    19
            public void setSolar(String date) {
     10
    <u>1</u>1
    22
     !3
            public String getLunar() {
     25
 \Box
     16
            }
    27
    28
    29
    30
    31
    32
    33
    34
    35
    36
    37
    38
    39
    40
    41
    42
    43
    44
    45
    46
    47
```

return solar;

return lunar;

this.solar = date;

```
public void setLunar(String lunar) {
           this.lunar = lunar;
       }
       public String getChineseAra() {
           return chineseAra;
       public void setChineseAra(String chineseAra) {
           this.chineseAra = chineseAra;
       }
       public String getAvoid() {
           return avoid;
       }
       public void setAvoid(String avoid) {
           this.avoid = avoid;
48
       public String getShould() {
49
           return should;
50
       }
51
       public void setShould(String should) {
52
           this.should = should;
53
54
       }
55
       public Almanac(String solar, String lunar, String chineseAra, String should,
56
57
               String avoid) {
           this.solar = solar;
58
           this.lunar = lunar;
59
60
           this.chineseAra = chineseAra;
           this.should = should;
61
           this.avoid = avoid;
62
63
       }
64 }
```

然后是抓取解析的主程序,写程序之前需要在官网下载需要的jar包

AlmanacUtil.java

复制代码

复制代码

```
package com.likx.picker.util;
import java.io.IOException;
import java.text.SimpleDateFormat;
import java.util.Calendar;
import java.util.Date;
import org.apache.http.HttpEntity;
import org.apache.http.ParseException;
import org.apache.http.client.ClientProtocolException;
import org.apache.http.client.methods.CloseableHttpResponse;
import org.apache.http.client.methods.HttpGet;
import org.apache.http.impl.client.CloseableHttpClient;
import org.apache.http.impl.client.HttpClients;
import org.apache.http.util.EntityUtils;
import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
```



关于 招聘 广告服务 百度 ©1999-2018 CSDN版权所有 京ICP证09002463号

▲ QQ客服 ● 客服论坛

经营性网站备案信息 网络110报警服务 中国互联网举报中心

北京互联网违法和不良信息举报中心

```
* @version 1.0 
* @author 溯源blog
* <STRONG>创建时间</STRONG> : 2016年4月11日 下午14:15:44
* <STRONG>修改历史</STRONG> :
* 修改人
                         修改时间
                                                     修改内容
*
public class AlmanacUtil {
    * 单例工具类
    */
   private AlmanacUtil() {
   }
    * 获取万年历信息
    * @return
    */
   public static Almanac getAlmanac(){
       String url="http://tools.2345.com/rili.htm";
       String html=pickData(url);
       Almanac almanac=analyzeHTMLByString(html);
       return almanac;
   }
   /*
    * 爬取网页信息
   private static String pickData(String url) {
       CloseableHttpClient httpclient = HttpClients.createDefault();
       try {
          HttpGet httpget = new HttpGet(url);
          CloseableHttpResponse response = httpclient.execute(httpget);
          try {
              // 获取响应实体
              HttpEntity entity = response.getEntity();
              // 打印响应状态
              if (entity != null) {
                  return EntityUtils.toString(entity);
              }
          } finally {
              response.close();
          }
      } catch (ClientProtocolException e) {
          e.printStackTrace();
      } catch (ParseException e) {
          e.printStackTrace();
       } catch (IOException e) {
          e.printStackTrace();
      } finally {
          // 关闭连接,释放资源
          try {
              httpclient.close();
          } catch (IOException e) {
              e.printStackTrace();
          }
       }
       return null;
      使用jsoup解析网页信息
   private static Almanac analyzeHTMLByString(String html){
       String solarDate,lunarDate,chineseAra,should,avoid=" ";
       Document document = Jsoup.parse(html);
       //公历时间
       solarDate=getSolarDate();
       //农历时间
       Element eLunarDate=document.getElementById("info_nong");
       lunarDate=eLunarDate.child(0).html().substring(1,3)+eLunarDate.html().substring(11);
       //天干地支纪年法
       Element eChineseAra=document.getElementById("info_chang");
       chineseAra=eChineseAra.text().toString();
       //宜
       should=getSuggestion(document,"yi");
       //己
```



联系我们



请扫描二维码联系客服

webmaster@csdn.net

2 400-660-0108

▲ QQ客服 ● 客服论坛

关于 招聘 广告服务 百度 ©1999-2018 CSDN版权所有

经营性网站备案信息 网络110报警服务

京ICP证09002463号

中国互联网举报中心

北京互联网违法和不良信息举报中心

```
return almanac;
       }
       /*
        * 获取忌/宜
        */
       private static String getSuggestion(Document doc,String id){
           Element element=doc.getElementById(id);
           Elements elements=element.getElementsByTag("a");
凸
           StringBuffer sb=new StringBuffer();
           for (Element e : elements) {
              sb.append(e.text()+" ");
          }
           return sb.toString();
\bigcirc
      }
        * 获取公历时间,用yyyy年MM月dd日 EEEE格式表示。
        * @return yyyy年MM月dd日 EEEE
       private static String getSolarDate() {
           Calendar calendar = Calendar.getInstance();
           Date solarDate = calendar.getTime();
           SimpleDateFormat formatter = new SimpleDateFormat("yyyy年MM月dd日 EEEE");
           return formatter.format(solarDate);
       }
   }
   复制代码
```

为了简单明了我把抓取解析抽象成了几个独立的方法,

其中pickData()方法使用httpClient来抓取数据到一个字符串中(就是在网页上点击查看源代码看到的HT ML源码),

analyzeHTMLByString()方法来解析抓取到的字符串,getSuggestion方法把抓取方法类似的宜忌数据抽象到了

一起,另外因为公历时间可以很容易的自己生成就没有在网页上爬取。

然后下面是一个测试类简单测试下效果:

AlmanacUtilTest.java

```
复制代码
```

```
package com.likx.picker.util.test;

public class AlmanacUtilTest {

   public static void main(String args[]){
        Almanac almanac=AlmanacUtil.getAlmanac();
        System.out.println("公历时间: "+almanac.getSolar());
        System.out.println("农历时间: "+almanac.getLunar());
        System.out.println("天干地支: "+almanac.getChineseAra());
        System.out.println("宜: "+almanac.getShould());
        System.out.println("忌: "+almanac.getAvoid());
    }
}
```

复制代码

运行结果如下:





另外最近博客一直没怎么更新,因为最近考虑到技术氛围的原因,离开了对日外包行业,前往 一家互联网公司就职。说一下最近的感受,那就是一个程序员最核心的竞争力不是学会了多少框架, 掌握多少种工具(当然这些对于程序员也不可或缺),而是扎实的基础以及快速学习的能力,比如今天 这个项目,从对httpClient,jsoup工具一无所知到编写出Demo代码总计大概1个多小时,在之前对于 我来说是不可想象的,在技术氛围浓厚的地方快速get技能的感觉,非常好。

当然本例只是一个非常浅显的小例子,网页上内容也很容易抓取,httpClient及jsoup工具更多强大 的地方没有体现到,比如httpClient不仅可以发送get请求,而且可以发送post请求,提交表单,传送 文件,还比如jsoup最强大的地方在于它支持仿jquery的选择器。本例仅仅使用了最简单的document.getElem entById()

匹配元素,实际上jsoup的选择器异常强大,可以说它就是java版的jquery,比如这样:

复制代码

```
Elements links = doc.select("a[href]"); // a with href
Elements pngs = doc.select("img[src$=.png]");
  // img with src ending .png
Element masthead = doc.select("div.masthead").first();
  // div with class=masthead
Elements resultLinks = doc.select("h3.r > a"); // direct a after h3
```

复制代码

另外还有很多强大的功能水平有限就不——列举了,感兴趣的可以参照官网文档,也欢迎交流指正。新技 能get起来!

本文版权归作者及博客园所有,转载请注明作者及原文出处

溯源blog http://www.cnblogs.com/lkxsnow/

2 目前您尚未登录,请登录或注册后进行评论

HttpClient+Jsoup 抓取网页信息



🤪 qq_25821067 2017年04月16日 12:55 🕮 8931

利用HttpClient和Jsoup技术抓取网页信息。HttpClient是支持HTTP协议的客户端编程工具包,并且它支持HTTP协议。 jsoup 是 一款基于 Java 平台的 网页html解析器...

加入CSDN,享受更精准的内容推荐,与500万程序员共同成长!



联系我们



请扫描二维码联系客服

webmaster@csdn.net

2 400-660-0108

▲ QQ客服 ● 客服论坛

关于 招聘 广告服务 酒 百度 ©1999-2018 CSDN版权所有 京ICP证09002463号

经营性网站备案信息 网络110报警服务 中国互联网举报中心

北京互联网违法和不良信息举报中心

使用Eclipse+httpClient+Jsoup读取网页数据-初级

本人最近几天学习使用HttpClient包读取



🥥 zhongmingzhao1234 2014年11月13日 22:35 👊 1542



大数据工程师为啥2018年又火热起来了?

是从未受冷落还是一直这么强?大数据工程师的这份资料你需要看看.....





联系我们

学习网络爬虫必备,HttpClient+JSOUP

2014年08月14日 16:50 1.37MB 下载



HttpClient+jsoup实现网页数据抓取和处理



🌑 java_zys 2016年06月18日 10:48 🕮 2177

介绍一种简单的网页抓取和处理方案



dreamer2020 2014年10月26日 23:01 🕮 1859

©1999-2018 CSDN版权所有

招聘

请扫描二维码联系客服

2 400-660-0108

广告服务

webmaster@csdn.net

■ QQ客服 ● 客服论坛

百度

京ICP证09002463号

经营性网站备案信息

网络110报警服务

中国互联网举报中心

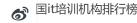
北京互联网违法和不良信息举报中心

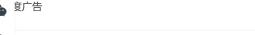
网络爬虫利器:fiddle+httpclient+jsoup

利用jsoup和httpclient来进行网站的爬取

前段日子帮同学写一个网络爬虫,

it培训机构排名









建议:事先定义一个线程池进行线程托管,推荐线程数20需定义:pool、worker、task、queue等参数(在此并不进行线程的讨 论)一、请求模拟 定义默认的一个closeableHttp...

httpClient+jsoup 抓取网页数据

2017年04月18日 17:25 2.82MB 下载



Java爬虫学习:利用HttpClient和Jsoup库实现简单的Java爬虫程序

本文介绍了如何利用HttpClient和Jsoup库的配合实现简单的Java爬虫程序



johnson_moon 2017年11月06日 16:37 🕮 274

Jsoup HttpClient 抓取网络上的图片



🌍 sxsj333 2011年07月27日 13:20 🕮 7212

package com.th.spider.test; import java.io.BufferedOutputStream; import java.io.FileOutputStream; ...

我是如何分分钟采集别人的WORDPRESS博客的

本文来源地址: http://www.xujh.top/2017/08/23/1708/ 最近我的博客新开了,一直在申请Google AdSense,结果申请一次 被拒一次,google发邮件说我...



50万码农评论:英语对于程序员有多重要?

不背单词和语法,一个公式学好英语



Java 爬虫工具Jsoup解析

🎁 Smile_Miracle 2017年04月25日 10:46 🕮 5082

Jsoup是一款 Java 的 HTML 解析器,可直接解析某个 URL 地址、HTML 文本内容。它提供了一套非常省力的 API,可通过 DO M, CSS 以及类似于 jQuery 的操作方法来取出和...

jsoup+httpclient

2014年03月07日 14:17 1.03MB 下载



httpclient和jsoup



wrb492960423 2014年08月21日 13:24 🚇 296

加入CSDN,享受更精准的内容推荐,与500万程序员共同成长!

登录

注册

JAVA 利用Jsoup 在网络获取数据

利用Jsoup可以更容易更快的获取数据下面我演示几个事例:

🎁 PeicongHe 2017年02月10日 00:51 🕮 1844

【网络爬虫】HttpClient抓取+解析+存储数据

🌏 tsj11514oo 2017年04月30日 17:25 🖺 791

前面使用了HttpClient抓取数据(http://blog.csdn.net/tsj11514oo/article/details/71023314),现在我们就要进行对数据的 分 析和存储。实现整一…

□ 叫中心系统

地门呼叫中心系统大全

日度广告



爬虫系列(一)——网页请求HttpClient

Daybreak1209 2017年02月07日 22:41 🚨 2596

爬虫系列博客将从以下几个方面介绍相对编写网页爬虫核心过程。 ——网页解析Jsoup 爬虫系列(三)——多线程爬… 爬虫系列 (—) ——网页请求HttpClient 爬虫系列 (二)

java httpclient 抓取网页 POST GET

w tanzuozhev 2016年03月01日 23:18 🕮 5649

httpclient post方法 以TTD数据库为例 // //Licensed to the Apache Software Foundation (ASF) under one * or mor...

Java HttpClient 实现自动登录与获取网页信息

● llwwlql 2016年10月11日 20:39 🖺 7902

用HttpGet获取网页上的信息: public void testGet(String url) throws ClientProtocolException, IOException { //...

Java使用HttpClient的HttpGet获取网页内容

🌘 testcs_dn 2015年03月02日 17:56 🕮 11118

项目添加HttpClient jar包引用 引用: import org.apache.http.HttpEntity; import org.apache.http.HttpResponse; im...

【基于Jsoup】Android通过Jsoup抓取网页信息详解(一)

1.关于Jsoup Jsoup是在Java中应用较为广泛的一种对HTML做解析的解析器,直接解析某个URL或本地的HTML文档内容,它提供了一套非常省力的API,可通过DOM,CSS以及类...

● u011669081 2015年11月09日 19:41 🚇 1207



联系我们



请扫描二维码联系客服

webmaster@csdn.net

2 400-660-0108

▲ QQ客服 ● 客服论坛

关于 招聘 广告服务 📸 百度 ©1999-2018 CSDN版权所有 京ICP证09002463号

经营性网站备案信息 网络110报警服务 中国互联网举报中心 北京互联网违法和不良信息举报中心