hacker_Lees的博客

java使用phantomJs抓取动态页面

标签： java phantomjs

2017-08-09 11:11 2653人阅读 评论(0) 收藏 举报

分类： phantomjs (9)

1. phantomjs的镜像网站：http://npm.taobao.org/dist/phantomjs/
2. phantomjs内置webkit内核，也就是chrome的内核。可以无界面加载页面，指的是和浏览器上面的页面一致，也就是解析完js的页面。所以需要爬取或者获得动态页面的，这算是利器。

3.之前自己也试了HttpUnit，不行的。网上找到的例子自己运行不了。报错太多。但是有没有文档，因为HttpUnit是2008年出的。官网上面啥也没有。所以我也没有资料参考，就放弃了。
4. 开始使用phantomjs，发现phantomjs算是动态爬取网页的主流。当然，所谓动态爬取从来不是问题，问题是速度。直接使用webkit等浏览器内核还是比较麻烦，而且速度不理想。

1. 自己使用的java + phantomjs在window上面开发。放到ubuntu上面。
   首先是安装，其实window版下载解压即可。但是如果你想要直接在cmd可以使用phantomjs的命令，请把bin下面的phantomjs.exe文件路径添加到path里面。此处程序不要依赖path路径。也就是直接使用绝对路径。当然绝对路径里面使用了项目的相对路径。这样是为了更好的迁移。phantomJS的使用过程就是java程序调用phantomJS调用js文件来获取指定页面，然后传回相应的内容。
   先给出代码：java端

```
1  import jdk.internal.org.xml.sax.SAXException;
2
3  import java.io.*;
4
5  public class JSUtil {
6
7      // 如果要更换运行环境，请注意exePath最后的phantom.exe需要更改。因为这个只能在window版本上运行。前面的路径名
8      // 也需要和exePath里面的保持一致。否则无法调用
```

```
    private static String projectPath = System.getProperty("user.dir");
    private static String jsPath = projectPath + File.separator + "test" + File.separator
            + "hello2.js";
    private static String exePath = projectPath + File.separator + "phantomjs-2.1.1" + File.separator + "bin"
            + File.separator + "phantomjs.exe";

    public static void main(String[] args) throws IOException, SAXException {

        // 测试调用。传入url即可
//        String html = getParseredHtml("https://b2b.10086.cn/b2b/main/showBiao!showZhaobiaoResult.html");
        String html = getParseredHtml2("http://huisheng99.b2b.hc360.com/");
        System.out.println("html: " + html);
    }

    public static String getParseredHtml(String url) throws IOException {
        Runtime rt = Runtime.getRuntime();

//        url传入并没用，建议url想办法传入.js文件中
//        Process p = rt.exec(exePath + " " + jsPath + " " + url);

        Process p = rt.exec(exePath + " " + jsPath);
        InputStream is = p.getInputStream();
        BufferedReader br = new BufferedReader(new InputStreamReader(is));
        StringBuffer sbf = new StringBuffer();
        String tmp = "";
        while ((tmp = br.readLine()) != null) {
            sbf.append(tmp);
        }

//        对数据进行处理，摘取自己需要的数据
//        String[] result = sbf.toString().split("companyServiceMod");
//        String result2 = "";
//        if(result.length >= 2)
//        {
//            result2 = result[1];
//            if(result2.length() > 200)
//            {
//                result2 = result2.substring(0, 200);
//            }
//        }
//        //System.out.println("resut2: "+result2);
//        return result2;

        return sbf.toString();
    }
```

```
54
55      public static String getParseredHtml2(String url) throws IOException {
56          Runtime rt = Runtime.getRuntime();
57
58          Process p = rt.exec(exePath + " " + jsPath + " " + url);
59
60          InputStream is = p.getInputStream();
61          BufferedReader br = new BufferedReader(new InputStreamReader(is));
62          StringBuffer sbf = new StringBuffer();
63          String tmp = "";
64          while ((tmp = br.readLine()) != null) {
65              sbf.append(tmp);
66          }
67
68  //          对数据进行处理，摘取自己需要的数据
69          String[] result = sbf.toString().split("companyServiceMod");
70          String result2 = "";
71          if (result.length >= 2) {
72              result2 = result[1];
73              if (result2.length() > 200) {
74                  result2 = result2.substring(0, 200);
75              }
76          }
77          return result2;
78
79  //          return sbf.toString();
80      }
81
82  }
```

然后是js文件，hello.js

```
1  /**
2   * Created by lee on 2017/8/9.
3   */
4  // a phantomjs example
5  var page = require('webpage').create();
6  phantom.outputEncoding="utf-8";
7  // page.open("https://b2b.10086.cn/b2b/main/showBiao!showZhaobiaoResult.html", function(status) {
8  page.open("http://www.baidu.com", function(status) {
9      if ( status === "success" ) {
10         console.log(page.title);
11         console.log(page);
12
```

```
        //将page转存为图片
        // page.render("D:/mavenProject/wnadmin/test/front-Thinking.jpg");

        //将page转存为pdf
        // page.paperSize = { format: 'A4',
        //     orientation: 'portrait',
        //     border: '1cm' };
        // page.render("front-Thinking.pdf");

        // 下面为无效代码，测试用的
        // var content = page.evaluate(function () {
        //     var element = document.getElementsByTagName('table');
        //     return element.textContent;
        // });
        // console.log(content);

    } else {
        console.log("Page failed to load.");
    }
    phantom.exit(0);
});
```

js文件2,hello2.js：

```
var page = require('webpage').create(),
  system = require('system'),
  t, address;
//写入文件，用来测试。正式版本可以注释掉用来提高速度。
var fs = require("fs");
//读取命令行参数，也就是js文件路径。
if (system.args.length === 1) {
  console.log('Usage: loadspeed.js <some URL>');
//这行代码很重要。凡是结束必须调用。否则phantomjs不会停止
  phantom.exit();
}
page.settings.loadImages = false;  //为了提升加载速度，不加载图片
page.settings.resourceTimeout = 10000;//超过10秒放弃加载
//此处是用来设置截图的参数。不截图没啥用
page.viewportSize = {
  width: 1280,
  height: 800
};
block_urls = ['baidu.com'];//为了提升速度，屏蔽一些需要时间长的。比如百度广告
page.onResourceRequested = function(requestData, request){
```

```
    for(url in block_urls) {
        if(requestData.url.indexOf(block_urls[url]) !== -1) {
            request.abort();
            //console.log(requestData.url + " aborted");
            return;
        }
    }
}
t = Date.now();//看看加载需要多久。
address = system.args[1];
page.open(address, function(status) {
  if (status !== 'success') {
    console.log('FAIL to load the address');
  } else {
    t = Date.now() - t;
//此处原来是为了提取相应的元素。只要可以用document的，还是看可以用。但是自己的无法用document，只能在用字符分割在java里。
    //  var ua = page.evaluate(function() {
    //   return document.getElementById('companyServiceMod').innerHTML;

    // });
    // fs.write("qq.html", ua, 'w');
   // console.log("测试qq: "+ua);
//console.log就是传输回去的内容。
    console.log('Loading time ' + t + ' msec');
    console.log(page.content);
    setTimeout(function(){ phantom.exit(); }, 6000);
  }
  phantom.exit();
});
```

请把js文件放到java的程序里面指定的路径。二者要一直。建议就是项目的根目录下面。
此处我是放在了项目的根目录下面。文件名是huicong.js

1. 有一个巨大的问题，就是速度。官网解释如下：

   stackoverflow给出的，如果截图，10秒算是正常。可以体会一下其速度。
   然后自己查了一下stackoverflow，找到了一个很好的回答。感谢大神们的无私分享：
   http://blog.csdn.net/kaka0930/article/details/68941932
   http://stackoverflow.com/questions/42703760/phantomjs-open-too-slow
   具体就是三点：
   6.1. 换个好点的电脑。

6.2. 不加载图片。参考上面的js文件。
6.3. 屏蔽相关广告等。参考上面的js文件。自己用了，成功吧时间压缩到2s。

7.自己是为了提取一个div里面的qq链接。但是没有找到怎么用dom来做。所以就直接传回整个page，然后手动用字符串解析。这里也许可以用各种selector。但是自己没有研究。

- 上一篇 phantomjs快速入门和使用说明
- 下一篇 java.lang.String中的trim()方法的详细说明

java使用phantomJs抓取动态页面

kaka0930 2017-04-01 10:45 6521

1. phantomjs的镜像网站：http://npm.taobao.org/dist/phantomjs/ 2. phantomjs内置webkit内核，也就是chrome的内核。可以无界面加载页面，指的是和浏览器上面的页面一致，也就是解析完js的页面。所以需要爬取或者获得动态页面的，这算是利...

Java爬虫进阶-Selenium+PhantomJs的运用

Smile_Miracle 2017-04-26 18:17 9730

selenium Selenium是一个用于Web应用程序测试的工具。Selenium测试直接运行在浏览器中，就像真正的用户在操作一样。支持的浏览器包括IE、Mozilla Firefox、Mozilla Suite等。这个工具的主要功能包括：测试与浏览器的兼容性——测试你的应用程序看是否能够很好...

java 调用 phantomjs

tengdazhang770960436 2014-11-21 13:55 16745

日前有采集需求，当我把所有的对应页面的链接都拿到手，准备开始根据链接去采集（写爬虫爬取）对应的终端页的时候，发觉用程序获取到的数据根本没有对应的内容，可是我的浏览器看到的内容明明是有的，于是浏览器查看源代码也发觉没有，此时想起该网页应该是ajax加载的。不知道ajax的小朋友可以去学下web开发啦。...

基于phantomJs的Java后台网页截图技术

ontologyFhcj 2017-02-14 14:24 4304

无

Java网络爬虫（十三）--PhantomJs的使用及性能优化

 championhengyi 2017-10-10 22:29 1011

先说点题外话吧，在我刚开始学习爬虫的时候，有一次一个学长给了我一个需求，让我把京东图书的相关信息抓取下来。恩，因为真的是刚开始学习爬虫，并且是用豆瓣练得手，抓取了大概500篇左右的影评吧，然后存放到了mysql中，当时觉得自己厉害的不行，于是轻松的接下了这个需求。。。 然后信心满满的开始干活。。首...

Java之网络爬虫WebCollector+selenium+phantomjs(一)

 jiangsanfeng1111 2016-08-27 14:07 3547

http://blog.csdn.net/osaymissyou0/article/details/49386637 最近研究了一下爬虫技术，与大家分享一下。 由于目前有很多成熟的框架(奉劝不要自己花时间再写爬虫框架了，真心没必要)，俺也就从中选一个适合我目前需求或者说相对简单的框架...

java调用phantomjs采集ajax加载生成的网页

 linsongze2011 2014-04-23 09:04 9452

日前有采集需求，当我把所有的对应页面的链接都拿到手，

java 与 phantomjs 的运用 (一)

 u013025479 2014-07-29 23:06 6376

最近在搞爬虫，一直以来用的都是

Java之网络爬虫WebCollector+selenium+phantomjs(三)

 jiangsanfeng1111 2016-08-27 14:14 1455

经过前面两篇的学习Java之网络爬虫WebCollector+selenium+phantomjs(一)与Java之网络爬虫WebCollector+selenium+phantomjs(二)的学习后，我们来做一个小例子。我们所要做的东西为:爬取到京东列表页面，在页面上抽取出商品信息(名称、价格、评...

java调用phantomjs采集ajax加载生成的网页

 hong0220 2014-12-03 23:24 3736

日前有采集需求，当我把所有的对应页面的链接都拿到手，准备开始根据链接去采集（写爬虫爬取）对应的终端页的时候，发觉用程序获取到的数据根本没有对应的内容，可是我的浏览器看到的内容明明是有的，于是浏览器查看源代码也发觉没有，此时想起该网页应该是ajax加载的。不知道ajax的小朋友可以去学下web开发啦。...

使用PhantomJS实现模拟登陆（Java爬虫）

 Nightmare_Zero 2017-09-18 21:17 558

记录了利用PhantomJS+Java进行模拟登陆的过程，同时介绍了PhantomJS，并记录了开发过程中遇到的问题

Java之网络爬虫WebCollector+selenium+phantomjs(一)

oSayMissyou0 2015-10-24 17:27 9010

java 爬虫学习 webcollector+selenium+phantomjs

selenium+phantomjs+java

2017-06-10 15:12 38.43MB 下载

java调用phantomjs

qq_33382373 2016-06-29 16:36 1244

目录 目录 前言 代码前言 下载完phantomjs之后直接解压就可以使用，然后在path目录加入phantomjs的路径（以便直接在命令行就可以执行phantomjs命令）。 接下来要完成个代码，一个是用phantomjs去获取页面（采用js编写行为），一个是采用java去调用phantomj...

phantomjs的使用+Java代码+依赖js（兼容Linux和windows版本）

1、 在使用phantomjs的时候需要下载phantomjs，网上有window版本和Linux版本。将phantomjs放在Linux上的指定位置之后（如下面的/home/tpl/phantomjs-1.9.7-linux-x86_64/处），2、 按照类似如下命令进行测...

toto1297488504 2016-05-28 15:53 8674

phantomjs的使用+Java代码+依赖js（兼容Linux和windows版本）

1、 在使用phantomjs的时候需要下载phantomjs，网上有window版本和Linux版本。将phantomjs放在linux上的指定位置之后（如下面的/home/tpl/phantomjs-1.9.7-linux-x86_64/处）， 2、 按照类似如下命令进行测试，是否可以生成...

hacker_Lees 2017-08-18 17:17 491

安装phantomjs，使用java代码进行截图

Zzhou1990 2016-07-22 10:11 3007

如何安装： 下载phantomjs安装文件，直接解压到相关目录，解包：tar xvf FileName.tar 创建软连接方便调用：(如果报错使用 ln -sf 强制执行) ln –s /root/satanbox/phantomjs/phantomjs-1.9....

phantomjs 简介


tengdazhang770960436 2014-11-20 20:35 53933

在爬虫、自然语言处理群320349384中的交流中，偶然接触到phantomjs、casper
等相对于httpclient较新的框架及采集解决方案，微查之后发现方案可行，故尽清明三日之力，将其二次开发应用于百度元搜索信息采集项目中，达到预期效果，下一步将重点应用到腾讯微博采集和抢票抢手机项目中。下面...

Java之网络爬虫WebCollector+selenium+phantomjs(二)


jiangsanfeng1111 2016-08-27 14:10 1890

上一篇做小例子的时候，在获取页面上价格的时候发现，获取不到，查了下说是webcollector需要结合selenium与phantomjs来获取js生成的动态。下面就做个例子来学习。 准备材料在上一篇已经准备完毕，我是在windows系统上进行的测试，所以phantomjs运行环境下载phan...

网页抓取方式（四）--phantomjs


chinabestchina 2017-06-11 12:22 635

一、phantomjs简介 phantomjs是基于webkit内核的无界面浏览器，因此我们可以借此进行网页抓取。 它的优点是： 1、本身就是在浏览器上操作，对js、css支持良好； 2、不容易被查封； 3、支持jquery操作； 缺点： 1、速度慢。 二、操作方式 ...