

[博客](#)[学院](#)[下载](#)[GitChat](#)[论坛](#)[问答](#)[商城](#)[写博客](#)[发Chat](#)[登录](#)[注册](#)

panda-star的博客

每天淘汰昨天的自己，形成一种习惯

[目录视图](#)[摘要视图](#)[RSS 订阅](#)

网页抓取方式（四）--phantomjs

2017-06-11 12:22

634人阅读

[评论\(0\)](#)

[收藏](#)

[举报](#)

分类：[爬虫（6）](#)

版权声明：本文为博主原创文章，未经博主允许不得转载。 <http://blog.csdn.net/chinabestchina/article/details/73017086>

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

phantomjs操作方式有两种：

- 1、原生的phantomjs操作；
- 2、借助页面测试中的selenium操作phantomjs；

三、安装phantomjs

其实就是一个可执行文件，各位同学可以根据自己的操作系统到<http://phantomjs.org/>这上面进行下载

四、原生的phantomjs实例

- 1、代码实例

```
var page = require('webpage').create();

page.settings.resourceTimeout=5000;

page.onConsoleMessage = function (msg) {

    console.log(msg);

}

phantom.outputEncoding = "gbk";

var url = "http://www.ifeng.com/";

page.open(url,function (status) {

    if (status !== 'success') {

        console.log('Unable to access network');

    } else {

        page.includeJs("http://libs.baidu.com/jquery/1.11.1/jquery.min.js",function () {

            var result = page.evaluate(function(){

                var result = $("#headLineDefault h1 a").html();

                console.log("ifeng headline is : "+result);

                return result;

            })

            console.log("result is :"+result);

        })

    }

    setTimeout(function () {

        phantom.exit();

    },5000);

})
```

运行结果：

ifeng headline is : 习近平哈萨克斯坦之行纪实

result is :习近平哈萨克斯坦之行纪实

五、借助selenium的phantomjs操作

1、添加maven依赖

```
<dependency>

    <groupId>org.seleniumhq.selenium</groupId>

    <artifactId>selenium-java</artifactId>

    <version>3.1.0</version>

</dependency>
```

2、代码实例

```
public class PhantomjsCrawlerMain {

    public static void main(String[] args) throws Exception {

        phantomjsCrawler();

    }

    static void phantomjsCrawler() throws Exception {

        String url = "http://www.ifeng.com/";

        String phantomjsPath = "/home/mong/bin/phantomjs.exe";

        DesiredCapabilities caps = new DesiredCapabilities();

        caps.setJavascriptEnabled(true);

        caps.setCapability(PhantomJSDriverService.PHANTOMJS_EXECUTABLE_PATH_PROPERTY, phantomjsPath);

        caps.setCapability(PhantomJSDriverService.PHANTOMJS_PAGE_SETTINGS_PREFIX + "loadImages", "false");

        final WebDriver driver = new PhantomJSDriver(caps);

        WebDriverWait webDriverWait = new WebDriverWait(driver, 2);

        final By headlineBy = By.cssSelector("#headLineDefault > h1 > a");

        driver.get(url);

        webDriverWait.until((ExpectedCondition<Boolean>) webDriver -> webDriver.findElement(headlineBy) != null);

        WebElement ele = driver.findElement(headlineBy);
```

```
        if(Objects.nonNull(ele)){

            String result = ele.getText().trim();

            System.out.println("ifeng headline is : " + result);

        }

    }

}
```

运行结果：

ifeng headline is : 习近平哈萨克斯坦之行纪实

- [上一篇](#) 网页抓取方式（三）--HtmlUnit
- [下一篇](#) mybatis配置

顶
0

踩
0

您还没有登录,请[\[登录\]](#)或[\[注册\]](#)

用Canvas实现截取网页内容保存为图片



cx067261

2018-01-13 17:06

51

最近有个客户提出需求，要把报表导出成电子版，什么格式都行。当时我想啊，转成word、Excel不可取，转成pdf可以完整保留页面样式，所以就找各种html转pdf的插件，也试了pechkin、itextsharp等等，然后就发现一个致命的问题。我的数据是js绑定的，页面里点和线的高度也是根据...

phantomjs 抓取网页



tengdazhang770960436

2014-08-13 16:34

8676

phantomjs：我的理解就是它是一个无显示的浏览器，也就是说除了不能显示页面内容以外，浏览器能干的活儿它基本上都能干。so,最近由于实验需要，要从某电商爬一点图片，但是它又是AJAX生成的，单纯的爬取HTML的方法是行不通的，o(′□′)o，于是在经过一些求助后，；了解到了PHANTO MJS，鉴...

使用phantomjs抓取JS动态生成的页面



u010814849

2016-10-18 10:52

3614

关于phantomjsphantomjs实现了一个无界面的webkit浏览器。虽然没有界面，但dom渲染、js运行、网络访问等API都很完整。可以利用phantomjs来下载js生成的页面。下载phantomjs (<http://phantomjs.org/download.html>)。解压到任意目录...

java使用phantomJs抓取动态页面



hacker_Lees

2017-08-09 11:11

📖 2653

phantomjs的镜像网站：<http://npm.taobao.org/dist/phantomjs/> 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...

程序员不会英语怎么行？！

免费报名网易公开课，一个公式教你秒懂天下英语



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——phantomjs抓取ajax动态加载网页



EQ__

2016-10-02 01:34

📖 7149

（说好的第二期终于来了 >_ 1、phantomjs介绍 phantomjs实现了一个无界面的webkit浏览器。虽然没有界面，但dom渲染、js运行、网络访问、canvas/svg绘制等功能都很完备，在页面抓取、页面输出、自动化测试等方面有广泛的应用。 官网: <http://phantomj...>

phantomjs 抓取网页



hblfyla

2016-09-28 16:01

📖 972

phantomjs：我的理解就是它是一个无显示的浏览器，也就是说除了不能显示页面内容以外，浏览器能干的活儿它基本上都能干。so,最近由于实验需要，要从某电商爬一点图片，但是它又是AJAX生成的，单纯的爬取HTML的方法是行不通的，o(′□′)o，于是在经过一些求助后，；了解到了PHANTOMJS，鉴...

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

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。...

Python网页信息采集：使用PhantomJS采集淘宝天猫商品内容



fullerhua

2016-07-06 11:48

📖 4790

最近一直在看Scrapy 爬虫框架，并尝试使用Scrapy框架写一个可以实现网页信息采集的简单的小程序。尝试过程中遇到了很多小问题，希望大家多多指教。

phantomjs使用说明

 mecho 2015-05-21 10:36  13243

phantomjs实现了一个无界面的webkit浏览器。虽然没有界面，但dom渲染、js运行、网络访问、canvas/svg绘制等功能都很完备，在页面抓取、页面输出、自动化测试等方面有广泛的应用。 安装 下载phantomjs（ 官方下载，下载失败请访问另一个下载点 ）。解压到任意目录，并将包含ph...

为什么casperjs比phantomjs好

 shaojwa 2015-11-20 17:30  3039

casperjs

PhantomJS快速入门

 libsyc 2017-10-11 09:15  4138

PhantomJS快速入门 本文简要介绍了PhantomJS的相关基础知识点，主要包括PhantomJS的介绍、下载与安装、HelloWorld程序、核心模块介绍等。由于鄙人才疏学浅，难免有疏漏之处，欢迎指正交流。 1、PhantomJS是什么？ PhantomJS...

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...

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 ...

PhantomJS简介

 violetgo 2015-08-30 15:56  4186

PhantomJS是一个可编程的无头浏览器.无头浏览器：一个完整的浏览器内核,包括js解析引擎,渲染引擎,请求处理等,但是不包括显示和用户交互页面的浏览器。可以使用Phantomejs做一些页面渲染的工作；如获取js的页面内容、截图等；

恭喜：一个公式教你秒懂天下英语

老司机教你一个数学公式秒懂天下英语



PhantomJs 快速入门github_367041582017-04-28 15:131005

PhantomJS 是一个基于 WebKit 的服务器端 JavaScript API。它全面支持web而不需浏览器支持，其快速，原生支持各种Web标准：DOM 处理, CSS 选择器, JSON, Canvas, 和 SVG。 PhantomJS 可以用于 页面自动化 ， 网络监测 ， 网页截屏...

自从有了Phantomjs和Casperjs，后台网页抓取和交互变得异常的简单

Casperjs是基于Phantomjs的，而Phantom JS是一个服务器端的 JavaScript API 的 WebKit。 这跟我一直想找个自带浏览器内核的后台东西的想法“暗合”。所以，在我发现这东西的时候就已经开始不由自主的兴奋起...

alexdream2013-08-30 09:5416369

PhantomJSfree_to_fly2016-04-17 18:21776

PhantomJS PhantomJS 是一个基于WebKit的服务器端 JavaScript API。它全面支持web而不需浏览器支持，其快速，原生支持各种Web标准：DOM 处理, CSS 选择器, JSON, Canvas, 和 SVG。 PhantomJS可以用于页面自动化，网络监测，网页截...

[Python爬虫] 在Windows下安装PIP+Phantomjs+SeleniumEastmount2015-08-19 20:0424251

最近准备深入学习Python相关的爬虫知识了，如果说在使用Python爬取相对正规的网页使用"urllib2 + BeautifulSoup + 正则表达式"就能搞定的话；那么动态生成的信息页面，如Ajax、JavaScript等就需要通过"Phantomjs + Ca...