[鼠绘漫画网 1](#_Toc26596)

[一拳超人 1](#_Toc23560)

[方案1，公众号抓包 1](#_Toc28116)

[方案2，解决网页图片懒加载的问题 1](#_Toc24572)

[方案3，解析资源库的URL规则 2](#_Toc31825)

[总结 3](#_Toc21146)

[海贼王 3](#_Toc723)

[百度贴吧 6](#_Toc31253)

[黑咪，监狱学园 6](#_Toc22779)

[飞丸网 6](#_Toc6391)

[东京食尸鬼 6](#_Toc742)

# 鼠绘漫画网

爬取鼠绘漫画网的资源。

## 一拳超人

### 方案1，公众号抓包

即便成功抓包也涉及到脚本的问题，因为我不可能一话一话的点过去。

而我不会python，我也不会shell。

pass。

### **方案2，解决网页图片懒加载的问题**

可以先用crawler4j试试。

2018-03-03 10:51:23

试过了，这吊东西不靠谱，获取的网页源代码还是什么都没有，就算是因为我没研究出来，那这吊东西还是不靠谱，一个好的工具就应该简便易用。搞那么多配置，或入口藏的深的要死，老子哪有那功夫去研究。

鼠标滑动过后，图片才会显示，看起来像是vue.js导致的。

那么现在有两种做法，找到那段js，然后…… ……然后个毛，我js那么差，pass。

第二种做法是继续找工具，可以直接获取js加载后的网页源代码。

2018-03-03 11:22:23

找到了，叫Selenium。

crawler4j还有jsoup这种工具都是直接访问url，那么服务器返回什么，他们就接收什么，静态资源还好说，但这种js加载出来的动态html代码，这些工具就获取不到了，而selenium，这个工具会调用浏览器。因为是靠浏览器去访问的，所以碰到js后，浏览器会自动执行，等浏览器执行完js，渲染出整个页面后，selenium在从浏览器那获取html。

（启动程序后，看着浏览器自动弹开有那么点诡异，跟中毒似的）

我用的是Chrome，所以要下载Chrome的驱动提供给selenium。

<dependency>

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

<artifactId>selenium-java</artifactId>

<version>3.10.0</version>

</dependency>

System.*getProperties*().setProperty("webdriver.chrome.driver","D:\\Workspace\\tools\\chromeDriver\\chromedriver.exe");

WebDriver webDriver = **new** ChromeDriver();

webDriver.get("http://hanhuazu.cc/cartoon/post?id=2292");

String responseBody = webDriver.getPageSource();

System.***out***.println(responseBody);

webDriver.close();

### **方案3，解析资源库的URL规则**

第1话，第1页

http://pic02.ishuhui.com/cartoon/book-7/53/1-6775/0001.jpg?0c850560

第128话，第1页

<http://pic03.ishuhui.com/cartoon/book-7/53/128-10463/0001.jpg?ef791431>

* ？后跟的是随机数，防缓存的，直接过
* 从pic02，变到pic03中间有没有pic02呢

~~我猜鼠绘分域名是因为资源量过大。~~

这你麻痹的就尴尬了，第128话，第1页，我用02，或是01都能正常访问。

[http://pic02.ishuhui.com/cartoon/book-7/53/128-10463/0001.jpg?ef791431](http://pic03.ishuhui.com/cartoon/book-7/53/128-10463/0001.jpg?ef791431)

也就是说，这些域名最后解析的都是同一个ip，那么这个问题我就算解决了。

* cartoon，book-7

cartoon是固定的。

book-7有可能变成book-1，其它的还没见过。这东西好比一个大文件夹。

* /53/1-6775/、/53/128-10463/

这一段的话，53是一拳超人的这部漫画的代号，1表示第一话，128表示第128话。关键是6775以及10463，这些代表某一话的文件资源位置。

这个迭代太麻烦了，资源位置不对会报404，超出这一话的长度也会报404，book-7或是book-1不对还是报404。

我现在不知道怎么去判断这个404是在哪个环节抛出来的。

另外一个问题是访问的太频繁可能被鼠绘服务器禁用IP，限制访问间隔时间的话，像海贼王那种几百话的总计上千页的漫画，又不知道什么时候能爬完。

但这个问题又有点多余，因为我规则还没解析出来。。。。

### **总结**

那就方案二了。

## 海贼王

只有一个方案，解析资源库的url并迭代。。。。。。。。

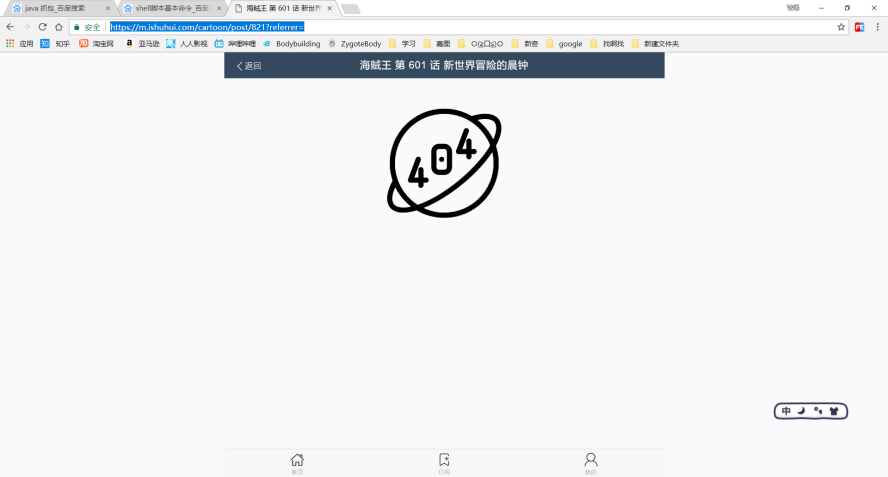
2018.3.19

公众号就是H5，换言之，这里头的内容是可以在浏览器里打开的。鼠绘公众号里的海贼王漫画也能在浏览器中打开，但图片文件却被屏蔽了。

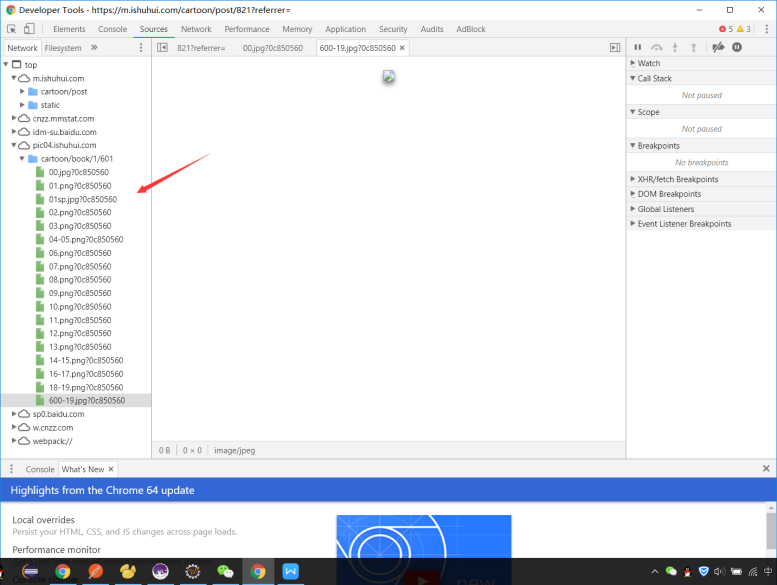
<https://m.ishuhui.com/cartoon/post/821>

以这个连接为例（这个连接是通过分享某一话的内容获得的，如果直接点击微信上“复制链接地址”的按钮，那么只能得到这个http://m.ishuhui.com/cartoon/book/1）

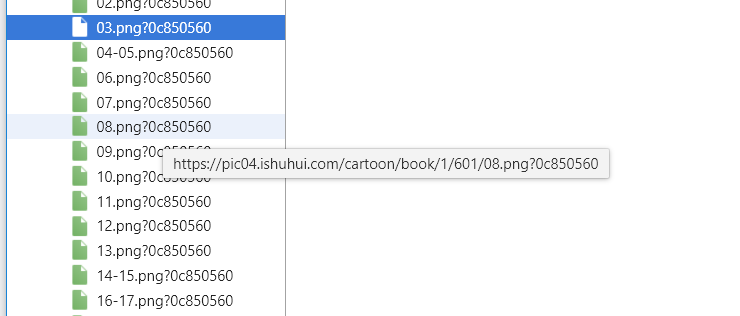
用浏览器访问这个连接的话，显示如下内容。



不过呼出控制台后却看到了有意思的东西



把鼠标移上去

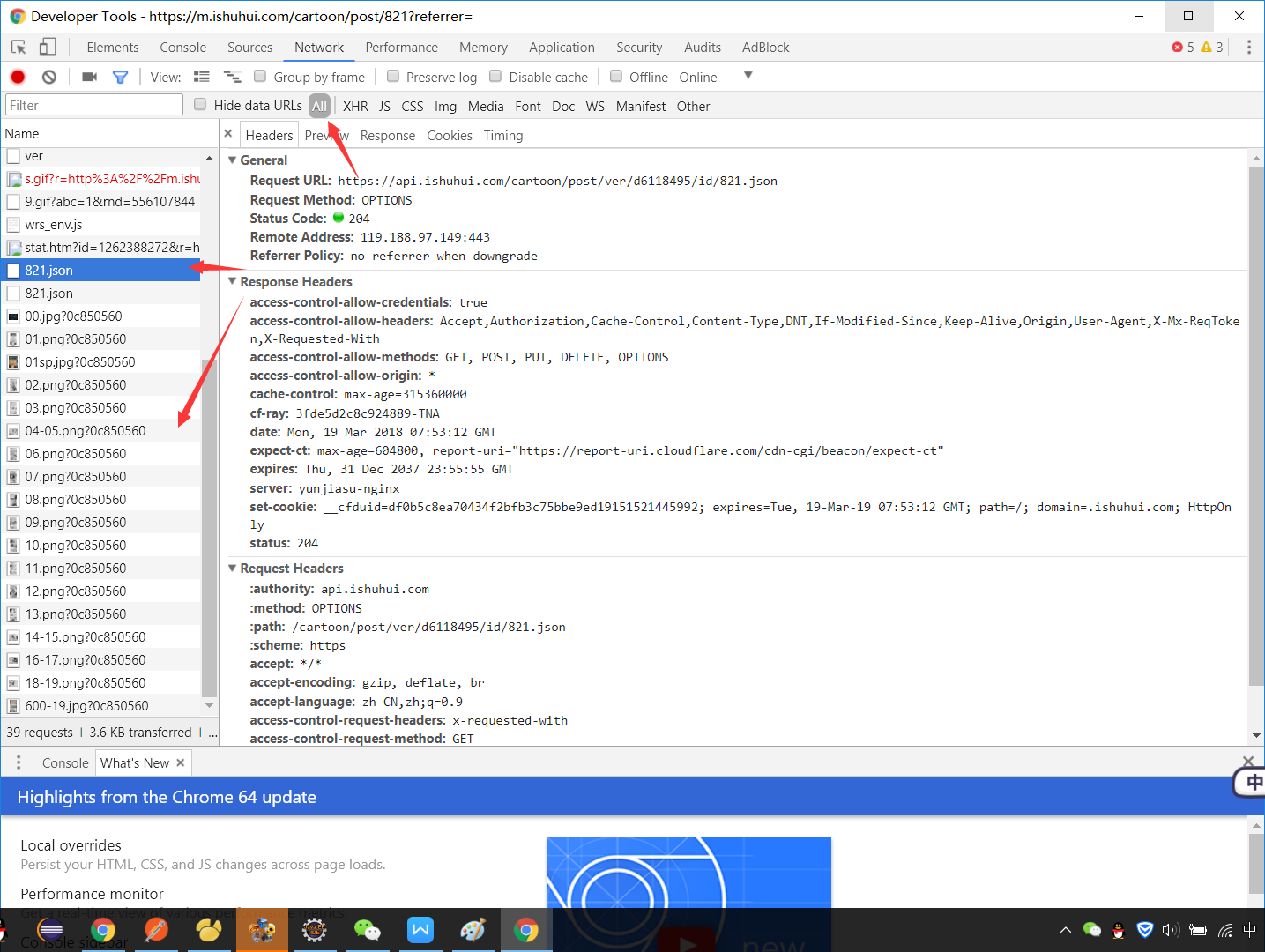


右键新标签打开



excetied !

他妈的，点开network后惊喜更多



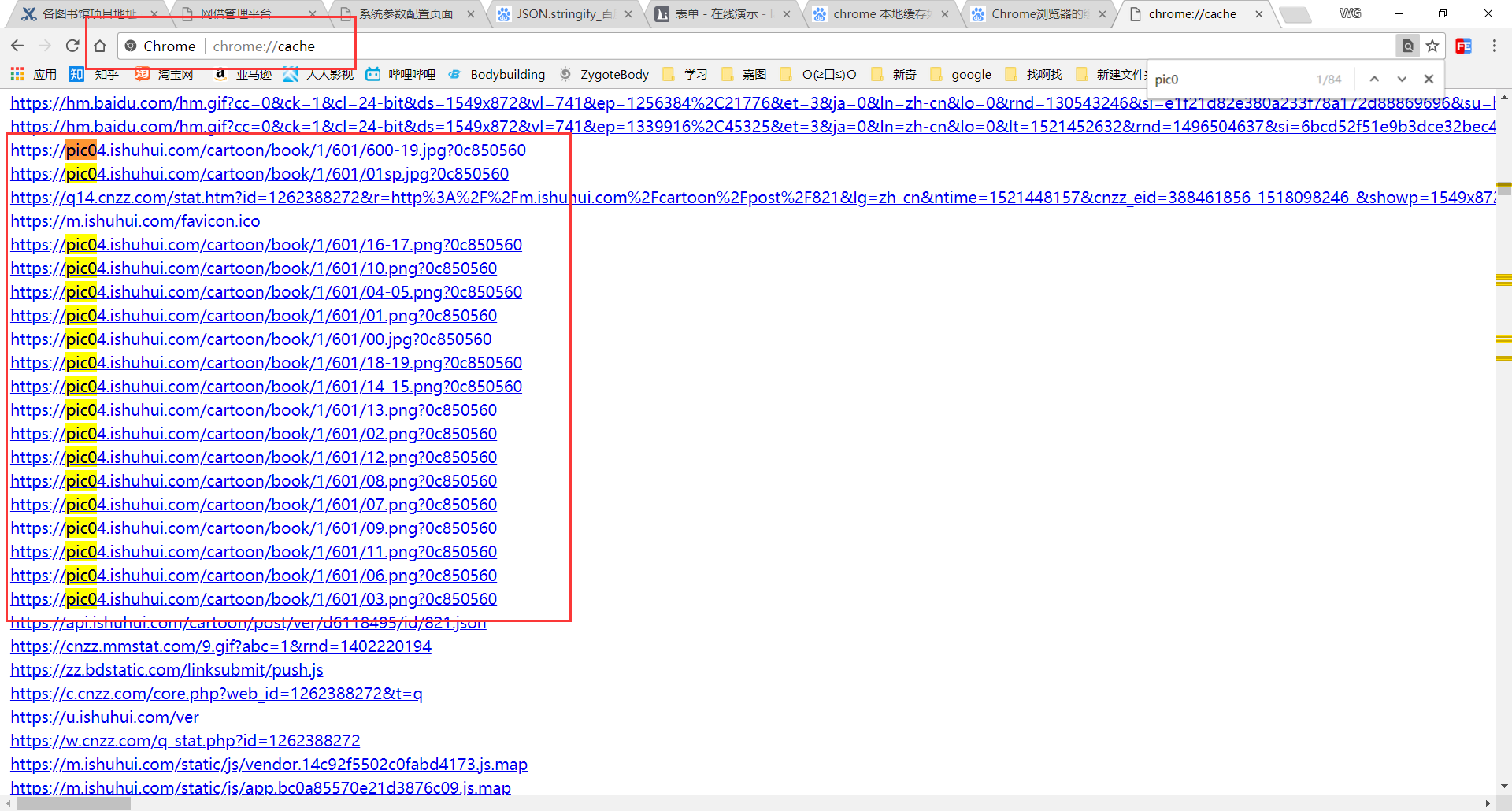
821.json的内容



拷了一部分

content\_img:"{"00.jpg":"/upload/cartoon/book/1/601/00.jpg","01.png":"/upload/cartoon/book/1/601/01.png","01sp.jpg":"/upload/cartoon/book/1/601/01sp.jpg","02.png":"/upload/cartoon/book/1/601/02.png","03.png":"/upload/cartoon/book/1/601/03.png","04-05.png":"/upload/cartoon/book/1/601/04-05.png","06.png":"/upload/cartoon/book/1/601/06.png","07.png":"/upload/cartoon/book/1/601/07.png","08.png":"/upload/cartoon/book/1/601/08.png","09.png":"/upload/cartoon/book/1/601/09.png","10.png":"/upload/cartoon/book/1/601/10.png","11.png":"/upload/cartoon/book/1/601/11.png","12.png":"/upload/cartoon/book/1/601/12.png","13.png":"/upload/cartoon/book/1/601/13.png","14-15.png":"/upload/cartoon/book/1/601/14-15.png","16-17.png":"/upload/cartoon/book/1/601/16-17.png","18-19.png":"/upload/cartoon/book/1/601/18-19.png","600-19.jpg":"/upload/cartoon/book/1/601/600-19.jpg"}"

到这里的话，又出现了一种可取的方式，因为控制台里的那些图片，肯定是存在Chrome缓存里的。那么把这些东西都提取出来，也实现了获取图片url的目的。



# 百度贴吧

## 黑咪，监狱学园

# 飞丸网

## 东京食尸鬼