

目录

前言	1.1
Python爬虫简介	1.2
裸写Python爬虫代码	1.3
用Python库写爬虫代码	1.4
用Python框架写爬虫代码	1.5
举例	1.6
抓取百度热榜	1.6.1
用Chrome分析逻辑	1.6.1.1
三种实现方式	1.6.1.2
用纯内置库裸写	1.6.1.2.1
用第三方库	1.6.1.2.2
用爬虫框架	1.6.1.2.3
附录	1.7
参考资料	1.7.1

如何用Python写爬虫

- 最新版本: v1.5
- 更新时间: 20200731

简介

总结如何用Python去写爬虫，包括如何裸写爬虫代码，如何用Python库去写爬虫，如何用Python爬虫框架写爬虫，并给出实例详细解释具体的操作过程，比如抓取百度首页中百度热榜标题列表，以及三种实现方式的完整代码和效果截图。

源码+浏览+下载

本书的各种源码、在线浏览地址、多种格式文件下载如下：

Gitbook源码

- [crifan/use_python_write_spider: 如何用Python写爬虫](#)

如何使用此Gitbook源码去生成发布为电子书

详见：[crifan/gitbook_template: demo how to use crifan gitbook template and demo](#)

在线浏览

- [如何用Python写爬虫 book.crifan.com](#)
- [如何用Python写爬虫 crifan.github.io](#)

离线下载阅读

- [如何用Python写爬虫 PDF](#)
- [如何用Python写爬虫 ePUB](#)
- [如何用Python写爬虫 Mobi](#)

版权说明

此电子书教程的全部内容，如无特别说明，均为本人原创和整理。其中部分内容参考自网络，均已备注了出处。如有发现侵犯您版权，请通过邮箱联系我 `admin 艾特 crifan.com`，我会尽快删除。谢谢合作。

鸣谢

感谢我的老婆陈雪的包容理解和悉心照料，才使得我 crifan 有更多精力去专注技术专研和整理归纳出这些电子书和技术教程，特此鸣谢。

更多其他电子书

本人 crifan 还写了其他 100+ 本电子书教程，感兴趣可移步至：

[crifan/crifan_ebook_readme: Crifan的电子书的使用说明](#)

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新：2021-01-17 12:15:09

Python爬虫简介

爬取你要的数据：爬虫技术中已经解释了爬虫的核心步骤了和相关涉及内容，也提到了很多语言都可以实现爬虫，都能爬取到你要的数据。

不过不同语言有自己的侧重点，而其中爬虫领域，最方便的要数Python。Python在爬虫领域，有很多的现成的库和框架可供使用，便于快速高效的实现爬虫的功能。

用Python写爬虫的不同方式

正如爬取你要的数据：爬虫技术中所整理的，用Python去写爬虫，也有三种方式：

- 裸写Python爬虫代码
 - 下载
 - python的内置http网络库
 - [urllib](#)
 - [crifanLibPython](#)中的[getUrlRespHtml](#)
 - [re](#)模块
 - [Python中的正则表达式：re模块详解](#)
 - 提取
 - [txt](#)
 - [csv / excel](#)
 - [Python心得：操作CSV和Excel](#)
 - 保存
 - [txt](#)
 - [csv / excel](#)
- 用各种Python库组合去写爬虫代码
 - 下载
 - 选择第三方的、更强大的、更好用的网络库
 - [Python心得：http网络库](#)
 - Requests
 - aiohttp
 - 提取
 - [BeautifulSoup](#)
 - [Python专题教程：BeautifulSoup详解](#)
 - v3 -> Python2
 - v4 -> Python3
 - [PyQuery](#)
 - [Python心得：HTML解析库PyQuery](#)
 - [lxml](#)
 - [【记录】Python中尝试用lxml去解析html – 在路上](#)
 - 等等
 - 保存
 - [csv / excel](#)
 - [PyMySQL](#)

- 主流关系数据库: MySQL
- PyMongo
 - 主流文档型数据库: MongoDB
- 等等
- 用爬虫框架去写爬虫代码
 - 常见Python爬虫框架
 - PySpider
 - Python爬虫框架: PySpider
 - Scrapy
 - 主流Python爬虫框架: Scrapy
 - 其他相关
 - 【整理】pyspider vs scrapy

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新: 2020-08-09 10:17:56

裸写Python爬虫代码

TODO:

用python内置urllib去裸写代码，去下载，再用re正则去提取，汽车之家车型车系数据。

旧教程

之前已写过一些相关教程，供参考：

- [详解抓取网站，模拟登陆，抓取动态网页的原理和实现（Python, C#等）](#)
 - [【教程】模拟登陆网站 之 Python版（内含两种版本的完整的可运行的代码） – 在路上](#)
- [Python专题教程：抓取网站，模拟登陆，抓取动态网页](#)

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新：2020-08-09 10:17:56

用Python库写爬虫代码

TODO:

用python的第三方http库，比如requests，去下载，再去用BeautifulSoup去提取，汽车之家车型车系数据。

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新：2020-08-09 10:17:56

用Python框架写爬虫代码

用Python爬虫框架PySpider去爬取汽车之家的车型车系数据

此处举例说明，用PySpider这个Python爬虫框架去爬取汽车之家的车型车系数据

详细过程参见：

[【已解决】写Python爬虫爬取汽车之家品牌车系车型数据 – 在路上](#)

期间包括：

- [【记录】Mac中安装和运行pyspider](#)
- [【已解决】pyspider中如何写规则去提取网页内容](#)
- [【已解决】pyspider中如何加载汽车之家页面中的更多内容](#)
- [【已解决】PySpider如何把json结果数据保存到csv或excel文件中](#)
- [【已解决】PySpider中如何清空之前运行的数据和正在运行的任务](#)

TODO：

把 `autohomeCarData` 代码上传到GitHub，并在此贴出地址

而关于PySpider更多的介绍，详见：

[Python爬虫框架：PySpider](#)

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新：2020-08-09 10:17:56

举例

下面通过举例例子来说明，去实现同一个爬虫目标，三种不同方式抓包是什么样的。

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新：2020-08-09 10:17:56

抓取百度热榜

具体详见：

- 【记录】演示如何实现简单爬虫：用Python提取百度首页中百度热榜内容列表
- 【已解决】用Python代码获取到百度首页源码并提取保存百度热榜内容列表
- 【已解决】Mac中用Chrome开发者工具分析百度首页的百度热榜内容加载逻辑
- 【已解决】用Python爬虫框架PySpider实现爬虫爬取百度热榜内容列表

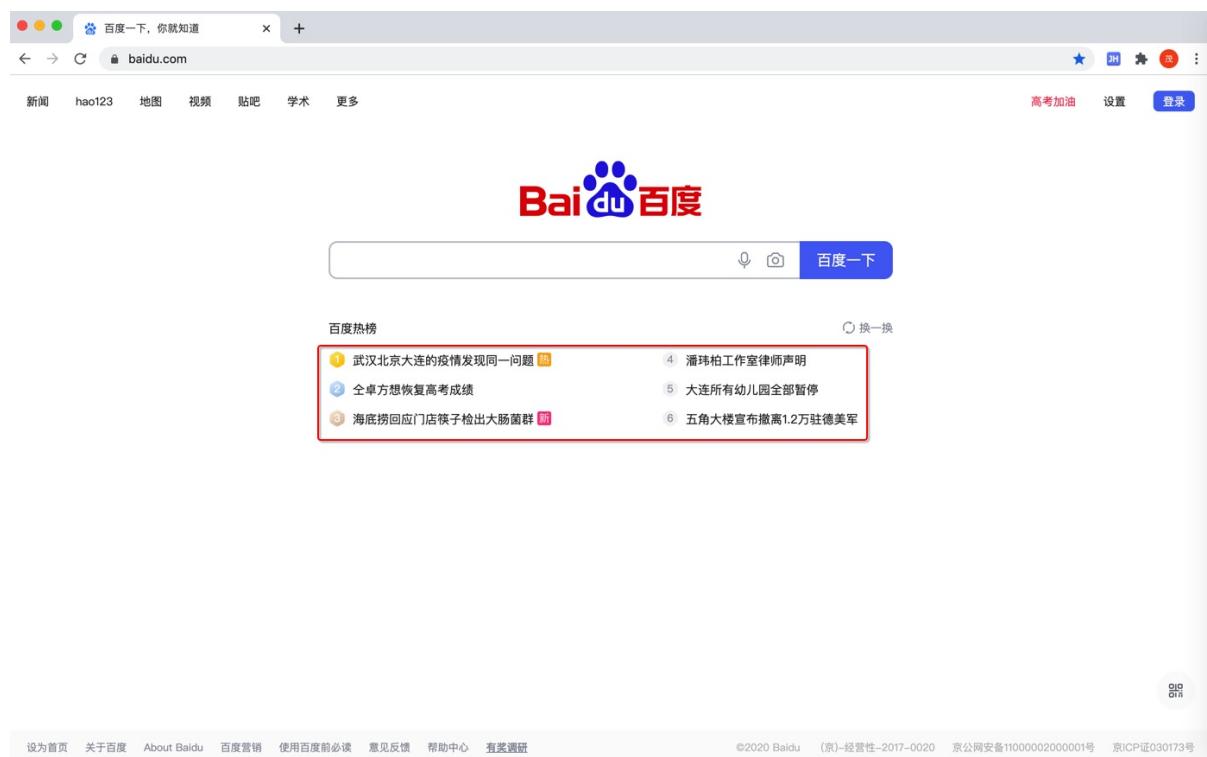
目标

爬取百度首页

百度一下，你就知道

<https://www.baidu.com/>

中的 百度热榜的内容的标题的列表：



希望输出的内容：

一个字符串列表：

- 武汉北京大连的疫情发现同一问题
- ...
- 五角大楼宣布撤离1.2万驻德美军

保存格式，暂定为csv文件。

先了解基础逻辑

入手之前，先要了解清楚：

- 写爬虫的思路
 - 先去（用工具）分析流程
 - 此处：用Chrome中 开发者工具 去分析
 - 用Chrome的开发者工具分析百度首页的内容加载的流程
 - 再去用代码实现逻辑
 - 此处：用Python代码实现
 - 要做的事情可以分成3个步骤
 - Download=下载：html网页源码
 - 期间可能涉及
 - 多次利用Chrome的开发者工具去调试页面内容加载逻辑
 - Parse=分析：分析html中源码中我们要的内容的提取规则是什么
 - 需要事先
 - 分析要抓取的内容，所对应的规则
 - 然后用代码实现规则，提取内容
 - Save=保存：把抓取到的内容保存出来

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新：2020-08-09 10:17:56

用Chrome分析逻辑

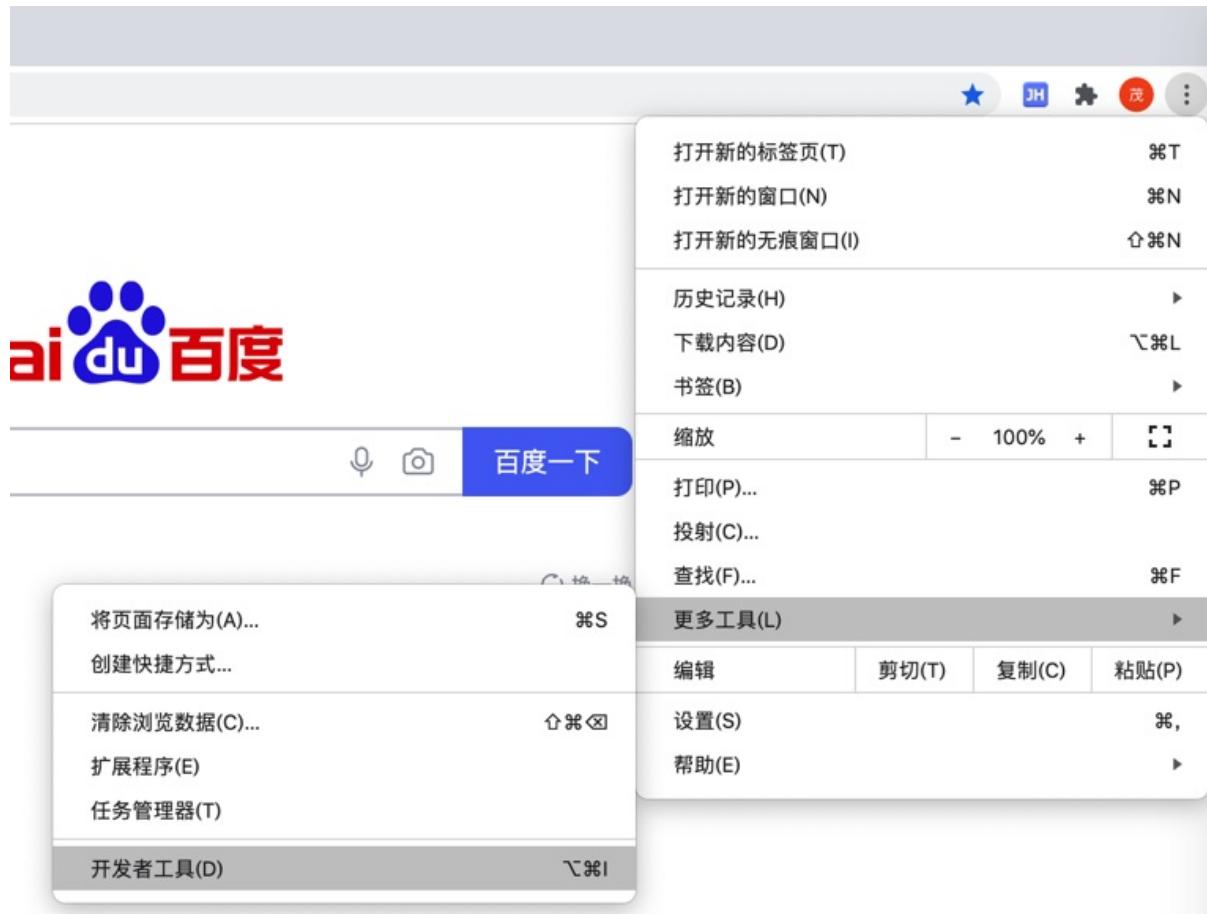
此处去用 Chrome 的 开发者工具 去分析百度首页中加载出百度热榜中的内容的基本逻辑。

先去 Mac 中打开 Chrome 中的 开发者工具 = Developer Tools :

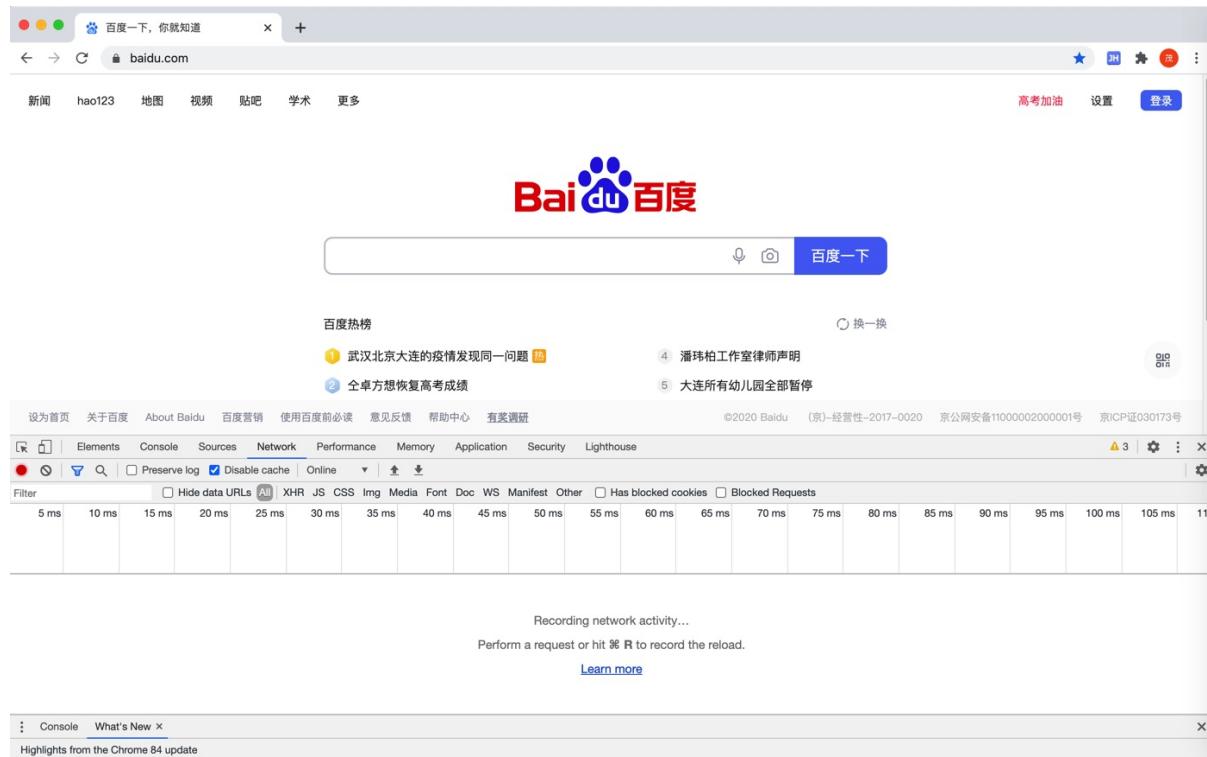
直接用快捷键 (Mac中是) : Option+Command+I

或:

更多 -> 更多工具 -> 开发者工具



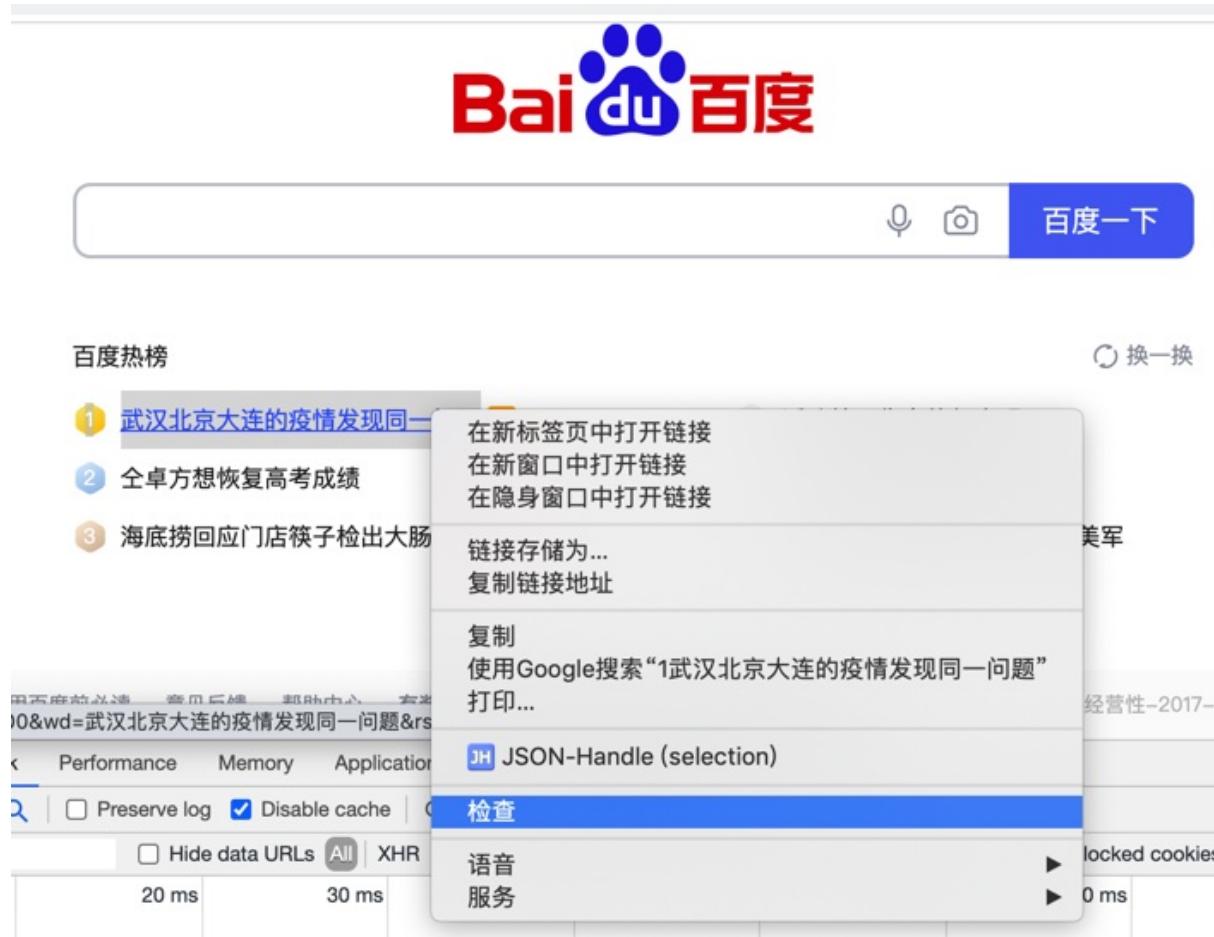
打开后效果：



先了解一下简单但常用的功能

比如最常见的 查看元素：

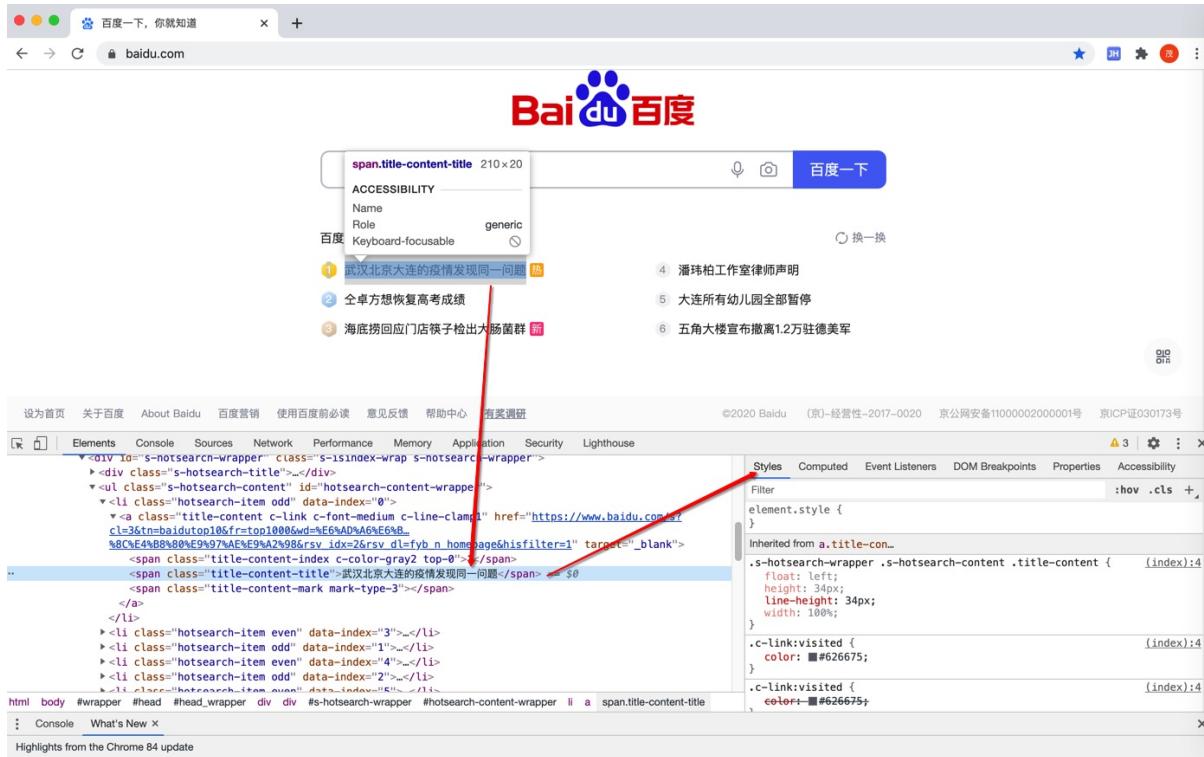
右键 -> 检查



即可看到：

坐标的Elements，表示 显示html网页源码

以及 右边是 css的Styles部分



此处，我们目的是：分析百度热榜中的内容列表是如何加载出来的

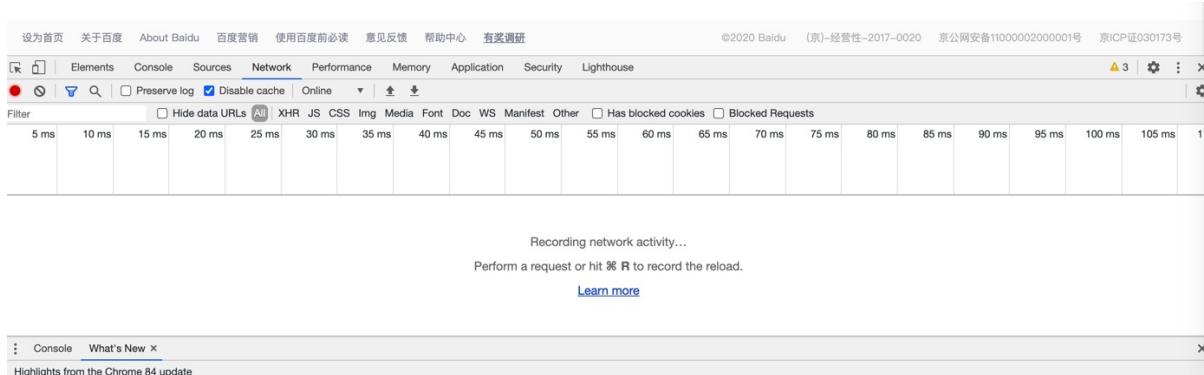
所以先去根据列表中第一个元素的内容：

武汉北京大连的疫情发现同一问题

去尝试搜索Network部分中的内容（请求或返回的响应中）能否搜索到

步骤：

切换到 Network 一栏：



可见，此处是空的

原因是，在打开开发者工具之前我们就已经加载完毕页面，所以开发者工具没有记录到内容。

先去使得工具能看到当前网页加载所有内容的过程和请求的列表，则思路就是：重新加载网页

其中也可以看到也有对应提示：

Recording network activity ...

Perform a request or hit Command+R to record the reload

所以去 重新刷新页面

快捷键：Command+R

The screenshot shows the Chrome DevTools Network tab for a request to `www.baidu.com`. The timeline at the top shows the duration of each resource load. Below the timeline is a table of requests:

Name	Status	Type	Initiator	Size	Time	Waterfall
<code>www.baidu.com</code>	200	document	Other	72.1 kB	72 ms	[Waterfall icon]
<code>baiduyun@2x-e0be79e69e.png</code>	200	png	(index)	5.0 kB	24 ms	[Waterfall icon]
<code>zhidao@2x-e9b427ec04.png</code>	200	png	(index)	3.0 kB	27 ms	[Waterfall icon]
<code>baike@2x-1fe3db7fa6.png</code>	200	png	(index)	3.4 kB	33 ms	[Waterfall icon]
<code>tupian@2x-482fc011fc.png</code>	200	png	(index)	2.4 kB	35 ms	[Waterfall icon]
<code>baobaozhidaօ@2x-a409f9dbe.png</code>	200	png	(index)	7.0 kB	45 ms	[Waterfall icon]
<code>wenku@2x-f3aba893c1.png</code>	200	png	(index)	3.2 kB	33 ms	[Waterfall icon]
<code>linnvn@2x-a53ea48cbh.nm</code>	200	nm	(index)	4.1 kB	33 ms	[Waterfall icon]

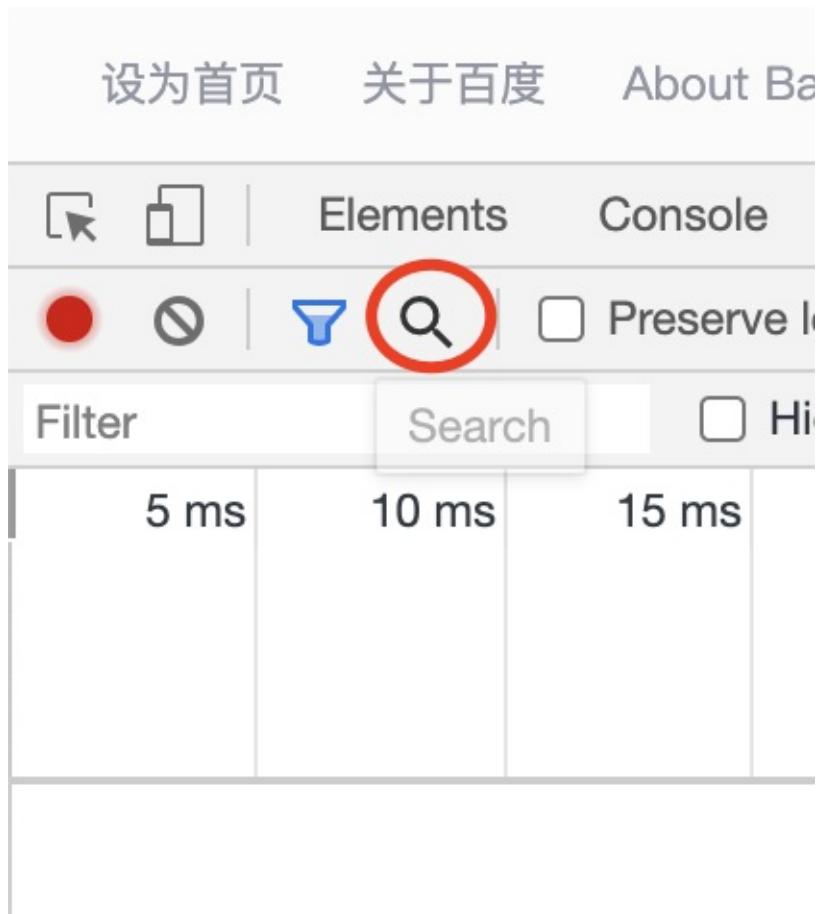
At the bottom of the Network tab, there are links for 'Console', 'What's New', and 'Highlights from the Chrome 84 update'.

就可以看到网页内容加载的细节和具体每个请求和详情了。

接下来再去找我们的要的内容。

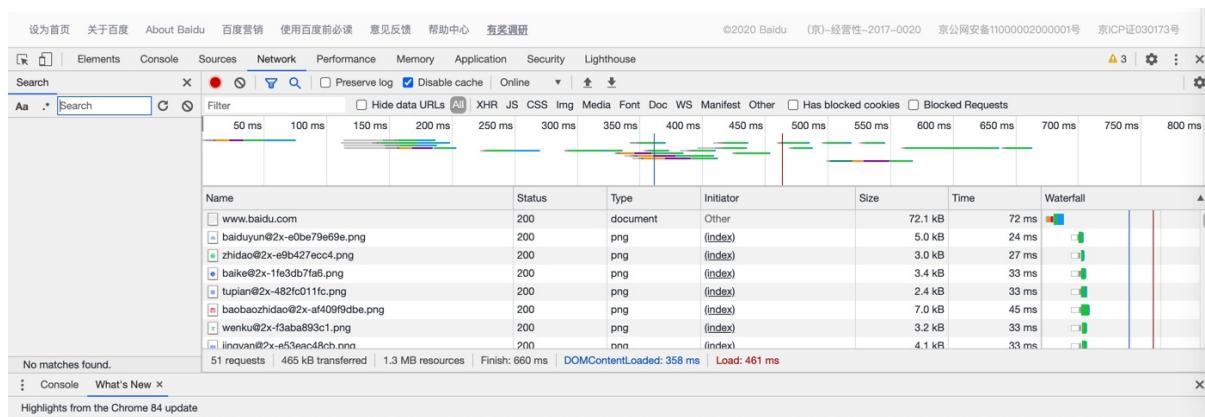
先打开搜索：

点击 搜索 按钮：



或快捷键： Command+F

即可打开搜索界面：



输入要搜的内容： 武汉北京大连的疫情发现同一问题， 并回车触发搜索

此处可以搜到一条记录， 点击会跳转过去：

The screenshot shows the Chrome DevTools Network tab. A request for `www.baidu.com` is selected. The 'Response' tab is open, displaying the HTML source code of the search results page. A red arrow points from the 'Name' column to the 'Response' tab.

```

50 .s-isindex-wrap(display:none)
51 #nv(display:none!important)
52 #head .head_wrapper(display:block;padding-top:0!important)
53 .s-bottom-ctner(display:none!important)
54 #head .s-upfunc-menus(display:none)
55 #s_skin_upload(display:none)</style><textarea><div id="wrapper" class="wrapper_new"><script>if(window.bd:
56 <div id="s_wrap" class="s-isindex-wrap"><div id="s_main" class="main clearfix "></div></div>
57 <div id="bottom_layer" class="s-bottom-layer s-isindex-wrap"><div class="s-bottom-layer-left"><p clas:
58 <div class="s_tab" id="s_tab">
59 <div class="s_tab_inner">
60 </div>
61

```

此处可以看出是：

`www.baidu.com`

的这条记录中的 Response 部分返回的 html 源码中包含了我们要搜索的内容

双击选中并复制该行：

The screenshot shows the same Network tab with the response body selected. A large red rectangular highlight covers the entire content of the 'Response' tab, indicating the selected text.

粘贴出来，放到编辑器中去研究内容，比如放到VSCode中：

```
#s_skin_upload{display:none}</style></te * Untitled-1 — gitbook_template
[[ SUMMARY.md ]]
```

1 #s_skin_upload{display:none}</textarea><div id="wrapper_new"><script>if(window.bds&&bds.util.setContainerWidth){bds.util.setContainerWidth();}</script><div id="head"><div id="s_top_wrap" class="s-top-wrap s-isindex-wrap"><div class="s-top-nav"></div><div class="s-center-box"></div></div><div id="u">百度首页设置<i class="c-icon c-icon-triangle-down"></i>登录<div class="bdpmenu"></div></div><div id="s-top-left" class="s-top-left s-isindex-wrap">新闻hao123地图视频贴吧学术<div class="mnav s-top-more-btn">更多<div class="s-top-more" id="s-top-more"><div class="s-top-more-content row-1 clearfix"><div class="s-top-more-title c-font-normal c-color-t">网盘</div><div class="s-top-more-title c-font-normal c-color-t">知道</div><div class="s-top-more-title c-font-normal c-color-t">百科</div><div class="s-top-more-title c-font-normal c-color-t">图片</div><div class="s-top-more-title c-font-normal c-color-t">宝宝知道</div><div class="s-top-more-title c-font-normal c-color-t">文库</div><div class="s-top-more-title c-font-normal c-color-t">经验</div><div class="s-top-more-title c-font-normal c-color-t">音乐</div>查看全部百度产品</div></div><div id="u1" class="s-top-right s-isindex-wrap">高考加油设置登录<div id="s-user-setting-menu" class="s-top-userset-menu c-floating-box c-font-normal"><div class="s-user-setting-pfmenu"></div>关闭热榜开启热榜</div></div><div id="head_wrapper" class="head_wrapper s-isindex-wrap nologin "><div class="s_form s_form_nologin"><div class="s_form_wrapper"><style>.index-logo-srcnew {display: none;}@media (-webkit-min-device-pixel-ratio: 2),

行 2, 列 1 空格: 4 UTF-8 LF 纯文本 ⌂

找到对应的内容所在位置：

搜 武汉北京大连的疫情发现同一问题，找到2处：

```
#s_skin_upload{display:none}</style></te> Untitled-1 — gitbook_template
SUMMARY.md □ #s_skin_upload{display:none}</style></te> Untitled-1 ●
class="hot-title" href="http://top.baidu.com/?fr=mhd_ca" > 北京大连的疫情发现同一问题 Aa Ab A* 2 中的 1 ↑ ↓ = ×
c-color-t">百度热榜</div></a><a id="hotsearch-refresh-btn" class="c-icon">&#xe619;</i><span class="hot-refresh-text">换一换</span></a></div><ul class="s-hotsearch-content" id="hotsearch-content-wrapper"><li class="hotsearch-item odd" data-index="0"><a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%AD%A6%E6%B1%89%E5%8C%97%E4%BA%C5%A4%A7%E8%BF%9E%7%9A%84%E7%96%AB%E6%83%85%E5%8F%91%E7%8E%B0%E5%90%8C%E4%8B%80%E9%97%AE%2&#8226; rsv_idx=2& rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank" ><span class="title-content-index c-color-gray2 top-0">1</span><span class="title-content-title">武汉北京大连的疫情发现同一问题</span><span class="title-content-mark mark-type-3"></span></a></li><li class="hotsearch-item even" data-index="3"><a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%BD%98%7%8E%AE%6%9F%8F%5B%7A%5E%4%BD%9C%5E%AE%A4%5E%BE%8B%5B%88%5E%3%A3%8E%6%98%8E& rsv_idx=2& rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank" ><span class="title-content-index c-color-gray2 top-3">4</span><span class="title-content-title">潘玮柏工作室律师声明</span><span class="title-content-mark mark-type-0"></span></a></li><li class="hotsearch-item odd" data-index="1"><a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E4%BB%9D%5E%8D%93%6%96%89%6%83%83%E6%81%2A%5E%44%8D%9B%98%8E%80%83%6%88%90%7%BB%A9& rsv_idx=2& rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank" ><span class="title-content-index c-color-gray2 top-1">2</span><span class="title-content-title">全卓方想恢复高考成绩</span><span class="title-content-mark mark-type-0"></span></a></li><li class="hotsearch-item even" data-index="4"><a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E5%A4%A7%8E%8F%9E%6%89%80%6%9C%89%5E%5B%9C%5E%84%BF%5%9B%AD%5E%85%8A%9%83%8A%6%9A%82%5E%81%9C& rsv_idx=2& rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank" ><span class="title-content-index c-color-gray2 top-4">5</span><span class="title-content-title">大连所有幼儿园全部暂停</span><span class="title-content-mark mark-type-0"></span></a></li><li class="hotsearch-item odd" data-index="2"><a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%85%87%5E%BA%95%6%8D%9E%5%9B%9E%5%BA%94%6%97%8A%5E%BA%97%7%AD%7%5E%AD%90%6%3%80%5E%87%BA%5E%4%7%8E%82%80%8E%8C%7%BE%A4& rsv_idx=2& rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank" ><span class="title-content-index c-color-gray2 top-2">3</span><span class="title-content-title">海底捞回应门店筷子检出大肠菌群</span><span class="title-content-mark mark-type-1"></span></a></li><li class="hotsearch-item even" data-index="5"><a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E4%BA%94%8E%7%9E%5%4%7%8E%8A%5E%84%BF%5%9B%AD%5E%85%8A%9%83%8A%6%9A%82%5E%81%9C&2%4%88%87%9A%9B%5E%8E%87%7%8E%8E%5%86%9B& rsv_idx=2& rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank" ><span class="title-content-index c-color-gray2 top-5">6</span><span class="title-content-title">五角大楼宣布撤离1.2万驻德美军</span><span class="title-content-mark mark-type-0"></span></a></li></div><textarea id="hotsearch_data" style="display:none;">{"hotsearch": [{"pure_title": "武汉北京大连的疫情发现同一问题", "linkurl": "https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%3Fcl%3D3%26tn%3Dbaidutop10%26fr%3Dtop1000%26wd%3D%25E6%25AD%25A6%25E6%25B1%2589%25E5%258C%2597%25E4%25BA%25AC%25E5%25A4%25A7%25E8%259E%259A%2584%25E7%2596%25AB%25E6%2583%2585%25E5%258F%2591%25E7%258E%25B0%25E5%2590%2558%25E4%25B8%2580%25E9%2597%25E9%25A2%255A%252598%25rsv_idx%25D2%25rsv_d%25fyb_n_homepage%26hisfilter%25D1", "views": "", "isViewed": "", "isNew": "", "heat_score": "4912516", "hotTags": "3"}, {"pure_title": "全卓方想恢复高考成绩", "linkurl": "https://www.baidu.com/s?cl=3&tn=baidutop10%26fr%3Dtop1000%26wd%3D%25E4%25BB%259D%25E5%258D%2593%25E6%2596%25B9%25E6%2583%25B3%2525", "views": "", "isViewed": "", "isNew": "", "heat_score": "4912516", "hotTags": "3"}]}行 1, 列 8335 (已选择 15) 空格: 4 UTF-8 LF 纯文本 ↻ 🔍
```

从经验来看，对于是要显示的html内容来说，第一条更像是我们要的

且再仔细看源码发现，前面有 百度热榜 的字样以及前面是 ul 的列表：

baidu_hotlist.html — gitbook_template

SUMMARY.md baidu_hotlist.html B

n_nologin > div.s_form_wrapper > div#s-hotsearch-wrapper.s-isindex-wrap.s-hotsearch-wrapper.hide. > ul#hotsearch-content-wrapper

value=""><input type=hidden name=tu value="baidu"><input type="hidden" name="s_ipt_wr" value="北京大连的疫情发现同一问题" />

<input id="kw" name="wd" class="s_ipt" value="北京大连的疫情发现同一问题" />

<input type="submit" id="su" value="百度一下" class="bg_s_btn"/><div id="mCon">输入法</div><ul id="mMenu">手写关闭<input type="hidden" name="rn" value="" /><input type="hidden" name="fenli" value="256" /><input type="hidden" name="oq" value="" /><input type="hidden" name="rsv_pq" value="95c0b7ba0004clf" /><input type="hidden" name="rsv_t" value="be8brSls2TpLbkPHBSpQvrdCTEhvTngTpJzs478j31Ju6wxtseS1C6jyoELM" /><input type="hidden" name="rlang" value="cn" /></form><div id="m" class="under-tips s_lm_hide" ><div id="lm-new" /></div></div><div id="s-hotsearch-wrapper" class="s-isindex-wrap s-hotsearch-wrapper hide" ><div class="s-hotsearch-title"><div class="title-text c-font-medium c-color-t" >百度热榜</div><i class="c-icon"></i>换一换</div><ul class="s-hotsearch-content" id="hotsearch-content-wrapper" ><li class="hotsearch-item odd" data-index="0" >1武汉北京大连的疫情发现同一问题<li class="hotsearch-item even" data-index="3" >4潘玮柏工作室律师声明<li class="hotsearch-item odd" data-index="1" >2

更验证了之前的推断。

把ul这部分html源码拷贝出来：

- [武汉北京大连的疫情发现同一问题1](https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%AD%A6%E6%B1%89%E5%8C%97%E4%BA%AC%E5%A4%A7%E8%BF%9E%E7%9A%84%E7%96%AB%E6%83%85%E5%8F%91%E7%8E%B0%E5%90%8C%E4%B8%80%E9%97%AE%E9%A2%98&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1)
- [潘玮柏工作室律师声明2](https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%BD%98%E7%8E%AE%E6%9F%8F%E5%87%A5%E4%BD%9C%E5%AE%A4%E5%BE%8B%E5%88%E5%A3%B0%E6%98%8E&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1)
- [全卓方想恢复高考成绩3](https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E4%BB%9D%E5%8D%93%E6%96%B9%E6%83%B3%E6%81%A2%E5%A4%8D%E9%AB%98%E8%80%83%E6%88%90%E7%BB%A9&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1)
- [大连所有幼儿园全部暂停4](https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E5%A4%A7%E8%BF%9E%E6%89%80%E6%9C%89%E5%BC%E5%84%BF%E5%9B%AD%E5%85%A8%E9%83%A8%E6%9A%82%E5%81%9C&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1)
- [大连所有幼儿园全部暂停5](https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E5%A4%A7%E8%BF%9E%E6%89%80%E6%9C%89%E5%BC%E5%84%BF%E5%9B%AD%E5%85%A8%E9%83%A8%E6%9A%82%E5%81%9C&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1)

```

="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%B5%B7%E5%BA%95%E6%8D%9E%E5%9B%9E%E5%BA%94%E9%97%A8%E5%BA%97%E7%AD%B7%E5%AD%90%E6%A3%80%E5%87%BA%E5%A4%A7%E8%82%A0%E8%8F%8C%E7%BE%A4&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank">><span class="title-content-index c-color-gray2 top-2">3</span><span class="title-content-mark mark-type-1"></span></a></li><li class="hotsearch-item even" data-index="5"><a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E4%BA%94%E8%A7%92%E5%A4%A7%E6%A5%BC%E5%AE%A3%E5%88%83%E6%92%A4%E7%A6%BB1.2%E4%BB%87%E9%A9%BB%E5%BE%87%E7%BE%8E%E5%86%9B&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank">><span class="title-content-index c-color-gray2 top-5">6</span><span class="title-content-title">五角大楼宣布撤离1.2万驻德美军</span><span class="title-content-mark mark-type-0"></span></a></li></ul>

```

且为了便于研究，再单独存到另外一个文件，且设置为HTML格式，使得语法高亮，便于阅读：

```

hotlist_ul_li.html — gitbook_template
SUMMARY.md  baidu_hotlist.html  hotlist_ul_li.html
Users > liamao > dev > tmp > demo_spider > hotlist_ul_li.html > ul#hotsearch-content-wrapper.s-hotsearch-content > li.hotsearch-item.odd > a.title-cont
1  <ul class="hotsearch-content" id="hotsearch-content-w> 北京大连的疫情发现同一问题 Aa * 1 中的 1 ↑ ↓ ⌂ × ; 
    <li class="hotsearch-item even" data-index="1"> <a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%AD%A6%E6%B1%89%E5%8C%97%E4%BA%AC%E5%A4%A7%E8%BF%9E%E7%9A%84%96%E6%83%85%E5%8F%91%E7%8E%80%50%8C%E4%8B%80%E9%7%AE%9A%98&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank">><span class="title-content-index c-color-gray2 top-0">1</span><span class="title-content-title">武汉北京大连的疫情发现同一问题</span><span class="title-content-mark mark-type-1"></span></a></li><li class="hotsearch-item even" data-index="3"> <a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%8D%98%E7%8E%AE%E6%9F%8F%E5%B7%A5%E4%BD%9C%E5%AE%A4%E5%BE%8B%E5%88%88%E5%A3%80%6%98%8E&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank">><span class="title-content-index c-color-gray2 top-0">4</span><span class="title-content-title">潘玮柏工作室律师声明</span><span class="title-content-mark mark-type-0"></span></a></li><li class="hotsearch-item odd" data-index="1"> <a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E4%BB%9D%58%93%E6%96%99%E6%83%83%6%81%2E5%4%80%9A%98%8E%80%83%6%88%90%7%BB%9A&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank">><span class="title-content-index c-color-gray2 top-1">2</span><span class="title-content-title">仝卓方想恢复高考成绩</span><span class="title-content-mark mark-type-0"></span></a></li><li class="hotsearch-item even" data-index="4"> <a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E5%44%A7%E8%BF%9E%6%89%80%6%9C%89%5%9B%C5%84%BF%5%9B%AD%5%85%8%9%83%AB%6%9A%82%5%81%9C&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank">><span class="title-content-index c-color-gray2 top-4">5</span><span class="title-content-title">大连所有幼儿园全部暂停</span><span class="title-content-mark mark-type-0"></span></a></li><li class="hotsearch-item odd" data-index="2"> <a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%85%87%5%9B%95%E6%8D%9E%5%9B%9E%5%BA%94%E9%97%8A%5%BA%97%E7%AD%7%5%AD%90%6%3%80%5%87%BA%5%A4%A7%8%82%A0%8%8F%8C%7%BE%A4&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank">><span class="title-content-index c-color-gray2 top-2">3</span><span class="title-content-title">海底捞回应门店筷子检出大肠菌群</span><span class="title-content-mark mark-type-1"></span></a></li><li class="hotsearch-item even" data-index="5"> <a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E4%BA%94%E8%A7%92%5%4%5%BC%5%AE%A3%5%88%83%6%92%A4%7%7%BB1.2%&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank">><span class="title-content-index c-color-gray2 top-5">6</span><span class="title-content-title">五角大楼宣布撤离1.2万驻德美军</span><span class="title-content-mark mark-type-0"></span></a></li></ul>

```

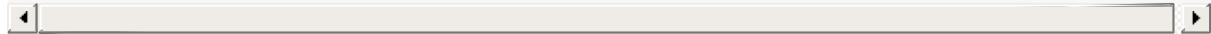
然后去研究代码，稍微懂点html的，即可了解其基本逻辑：

此处第一个li的元素：武汉北京大连的疫情发现同一问题

对应源码就是：

```
<li class="hotsearch-item odd" data-index="0"><a class="title-content c-link c-font-medium
```

```
c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%AD%A6%E6%B1%89%E5%8C%97%E4%BA%AC%E5%A4%A7%E8%BF%9E%E7%9A%84%E7%96%AB%E6%83%85%E5%8F%91%E7%8E%B0%E5%90%8C%E4%B8%80%E9%97%AE%E9%A2%98&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank" ><span class="title-content-index c-color-gray2 top-0">1</span><span class="title-content-title">武汉北京大连的疫情发现同一问题</span><span class="title-content-mark mark-type-3"></span></a></li>
```



其后的其他几个热榜节点的格式也是类似的，只不过是

- hotsearch-item 是 even
- data-index=1

等等细节不同：

```
<li class="hotsearch-item even" data-index="3"><a class="title-content c-link c-font-medium c-line-clamp1" href="https://www.baidu.com/s?cl=3&tn=baidutop10&fr=top1000&wd=%E6%BD%98%E7%8E%AE%E6%9F%8F%E5%87%A5%E4%BD%9C%E5%AE%A4%E5%BE%8B%E5%B8%88%E5%A3%B0%E6%98%8E&rsv_idx=2&rsv_dl=fyb_n_homepage&hisfilter=1" target="_blank" ><span class="title-content-index c-color-gray2 top-3">4</span><span class="title-content-title">潘玮柏工作室律师声明</span><span class="title-content-mark mark-type-0"></span></a></li>
```

所以，对于上述内容，此处研究出来的逻辑是：

如果能正常获取到

<https://www.baidu.com/>

的html源码，且内容已加载完毕的情况下

则直接去使用此简化规则去匹配内容：

```
<span class="title-content-title">xxx</span>
```

其中xxx是中文字符串，是此处希望找到的内容的标题。

说明：自己要确保此规则不会和导致误判，多了或少了，即匹配出其他的额外的不想要的内容，或漏了某些想要的内容。

当误判时，就需要加上其他限定条件，比如此处的：

父级节点是：li中 class="hotsearch-item even" 或 "hotsearch-item odd"

对于上述简化规则，再去用代码实现，提取要的内容：

(1) Python中re正则

```
contentTitleP = '<span\s+class="title-content-title">(?P<contentTitle>[^>]+)</span>'
```

(2) Python的用于解析html的第三方库BeautifulSoup

```
allTitleSoupList = soup.find_all("span", attrs={"class": "title-content-title"})
```

至此，要抓取的内容的提取规则，已分析完毕。

接下来，就是回头再去确保，可以正常获取到

<https://www.baidu.com/>

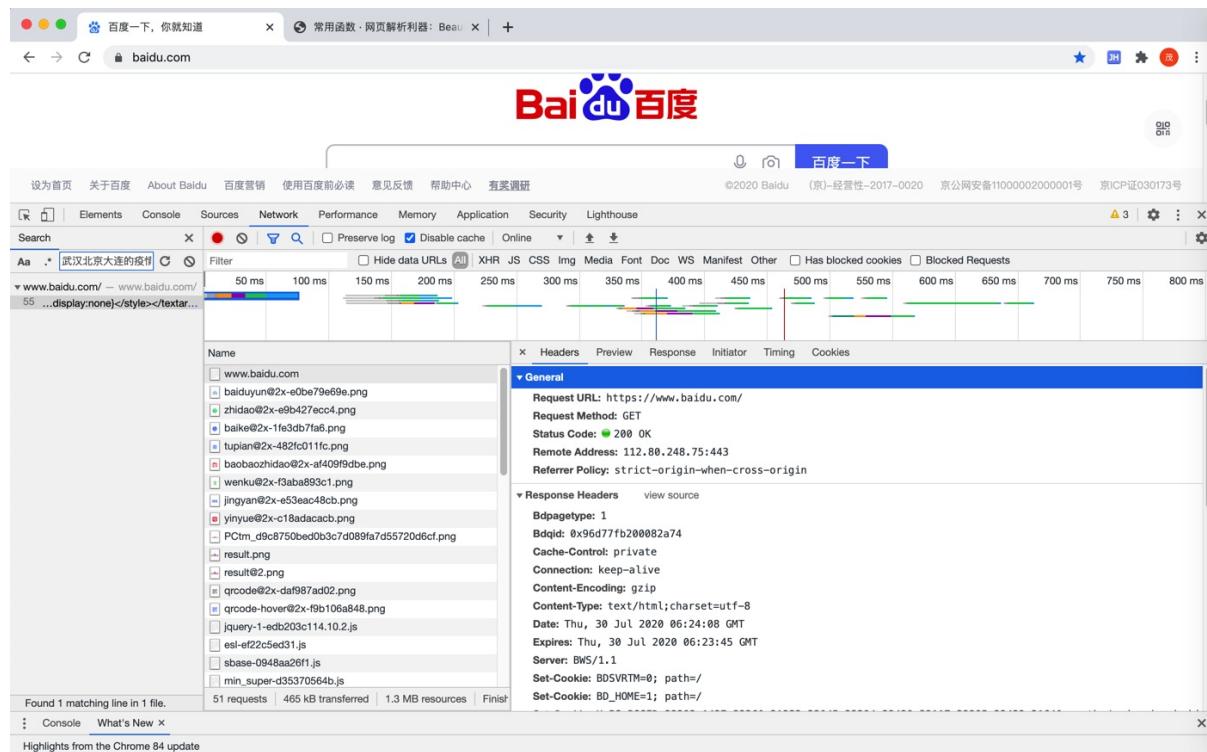
的源码，即可。

对此，往往不太容易一次性就很轻松的获取各个网站的网页源码。

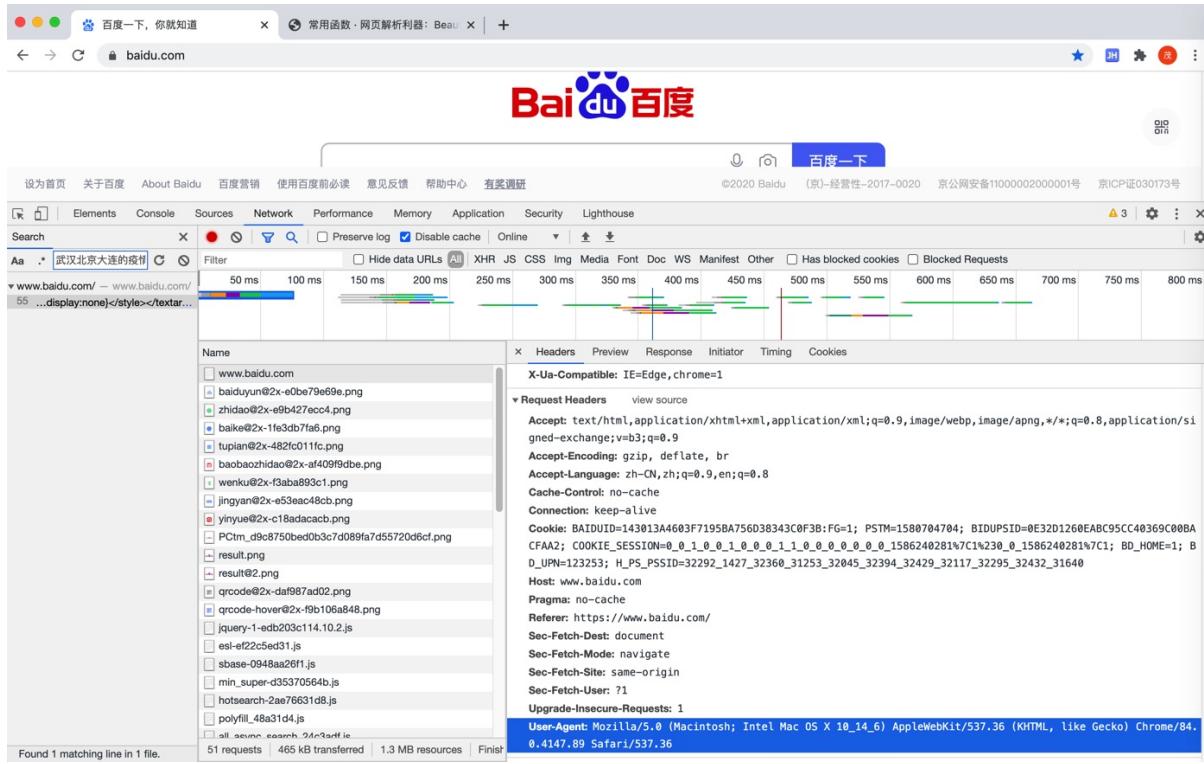
所以一般的逻辑是：直接去写代码，然后出现问题，变调试，变优化代码，直到最终获取到源码

期间继续调试逻辑：

切换到Headers界面：



找到 Request Headers 中的 User-Agent 部分：



拷贝出来是：

```
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/84.0.4147.89 Safari/537.36
```

用于后续代码中使用。

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新：2020-08-09 10:17:56

用Python实现爬虫逻辑

通过前面用Chrome的开发者工具分析完逻辑后，再去用Python代码实现爬取的全部逻辑。

此处如之前所述，主要可以分3种实现方式：

- 裸写代码，纯内置库，不用第三方库
- [用第三方库](#)
 - 记录了从无到有写出完整代码的详细过程
- [用爬虫框架](#)

下面分别介绍具体实现方式。

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新：2020-08-09 10:17:56

用纯内置库裸写

此处使用纯Python的库，主要是：

- 下载=HTTP网络库下载HTML源码
 - urllib
 - 解析=解析提取所需内容
 - 正则
 - re

核心代码：

用内置网络库：urllib：

```
import urllib.request

baiduUrl = "https://www.baidu.com/"

# Method 1 (pure python built-in lib, no third-party lib): urllib
baiduResp = urllib.request.urlopen(baiduUrl)
baiduHtmlBytes = baiduResp.read()
baiduHtml1 = baiduHtmlBytes.decode()
```

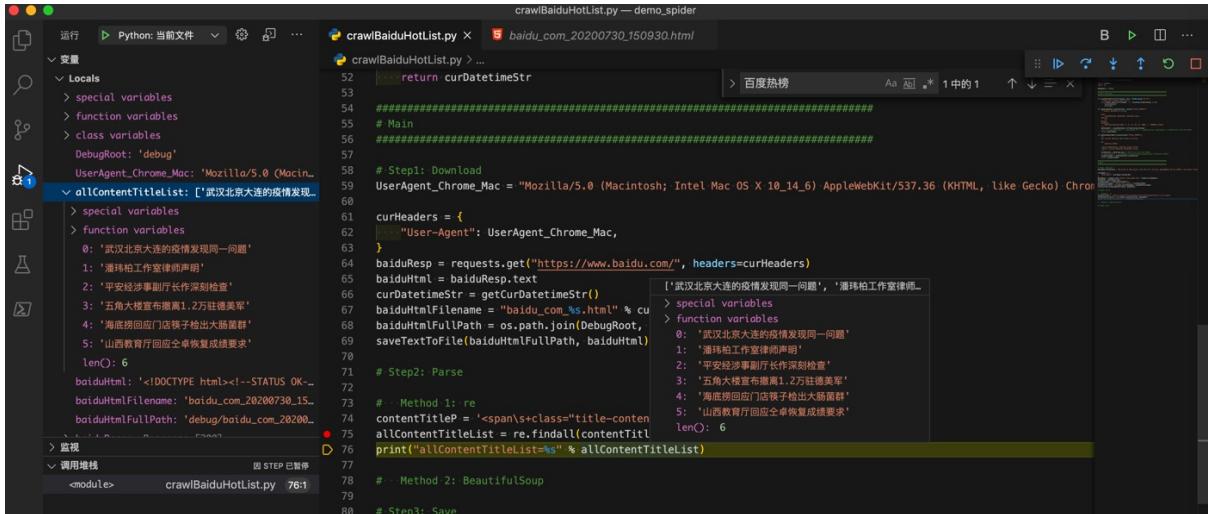
即可获取到HTML源码：

```
baiduUrl = "https://www.baidu.com"  
> codecs: <module 'codecs' from '/Users/limou/_  
> csv: <module 'csv' from '/Users/limou/.pyenv/_  
'<!DOCTYPE html><!-STATUS OK-->\n<html><head><meta http-equiv="Content-Type" content="text/html;charset=utf-8"><meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1"><meta conte  
> curHeaders: {User-Agent: Mozilla/5.0 (Mac_ OS_X/10.12.6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36  
> os: <module 'os' from '/Users/limou/.pyenv/_  
> re: <module 're' from '/Users/limou/.pyenv/_  
> urllib: <module 'urllib' from '/Users/limou/_
```

或许用内置库：正则 re :

```
# Step2: Parse
#   Method 1: re
contentTitleP = '<span\s+class="title-content-title">(?P<contentTitle>[^<>]+)</span>'
allContentTitleList = re.findall(contentTitleP, baiduHtml)
print("allContentTitleList=%s" % allContentTitleList)
```

即可匹配出要的热榜标题列表：



完整代码

```

# Function: Demo how use Python crawl baidu.com 百度热榜
# Author: Crifan
# Update: 20200731

import os
import codecs
from datetime import datetime, timedelta

import urllib.request
import re

# import requests
# from bs4 import BeautifulSoup

import csv

DebugRoot = "debug"
OutputRoot = "output"

#####
# Utils Functions
#####

def createFolder(folderFullPath):
    """
        create folder, even if already existed
        Note: for Python 3.2+
    """
    os.makedirs(folderFullPath, exist_ok=True)

def saveTextToFile(fullFilename, text, fileEncoding="utf-8"):
    """save text content into file"""
    with codecs.open(fullFilename, 'w', encoding=fileEncoding) as fp:
        fp.write(text)

```

```

fp.close()

def datetimeToStr(inputDatetime, format="%Y%m%d_%H%M%S"):
    """Convert datetime to string

Args:
    inputDatetime (datetime): datetime value
Returns:
    str
Raises:
Examples:
    datetime.datetime(2020, 4, 21, 15, 44, 13) -> '20200421_154413'
"""

datetimeStr = inputDatetime.strftime(format)
# print("inputDatetime=%s -> datetimeStr=%s" % (inputDatetime, datetimeStr)) # 2020-04
-21 15:08:59.787623
return datetimeStr

def getCurDatetimeStr(outputFormat="%Y%m%d_%H%M%S"):
    """
    get current datetime then format to string

eg:
    20171111_220722

:param outputFormat: datetime output format
:return: current datetime formatted string
"""

curDatetime = datetime.now() # 2017-11-11 22:07:22.705101
# curDatetimeStr = curDatetime.strftime(format=outputFormat) #'20171111_220722'
curDatetimeStr = datetimeToStr(curDatetime)
return curDatetimeStr

def saveToCsvByDictList(csvDictList, outputPath):
    # generate csv headers from dict list
    firstItemDict = csvDictList[0]
    csvHeaders = list(firstItemDict.keys())
    with codecs.open(outputPath, "w", "UTF-8") as outCsvFp:
        csvDictWriter = csv.DictWriter(outCsvFp, fieldnames=csvHeaders)

        # write header by inner function from fieldnames
        csvDictWriter.writeheader()

        for eachRowDict in csvDictList:
            csvDictWriter.writerow(eachRowDict)

def saveToCsvByHeaderAndList(csvHeaderList, csvRowListList, outputPath):
    with codecs.open(outputPath, "w", "UTF-8") as outCsvFp:
        csvWriter = csv.writer(outCsvFp)

        # write header from list
        csvWriter.writerow(csvHeaderList)

```

```

# type1: write each row
# for eachRowList in csvRowListList:
#     csvWriter.writerow(eachRowList)

# type2: write all rows
csvWriter.writerows(csvRowListList)

#####
# Main
#####

createFolder(DebugRoot)
createFolder(OutputRoot)

curDatetimeStr = getCurDatetimeStr()

# Step1: Download
UserAgent_Chrome_Mac = "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/84.0.4147.89 Safari/537.36"

curHeaders = {
    "User-Agent": UserAgent_Chrome_Mac,
}

baiduUrl = "https://www.baidu.com/"

# Method 1 (pure python built-in lib, no third-party lib): urllib
baiduResp = urllib.request.urlopen(baiduUrl)
baiduHtmlBytes = baiduResp.read()
baiduHtml = baiduHtmlBytes.decode()

# # Method 2 (use third-party lib): requests
# baiduResp = requests.get(baiduUrl, headers=curHeaders)
# baiduHtml = baiduResp.text

# for debug
baiduHtmlFilename = "baidu_com_%s.html" % curDatetimeStr
baiduHtmlFullPath = os.path.join(DebugRoot, baiduHtmlFilename)
saveTextToFile(baiduHtmlFullPath, baiduHtml)

# Step2: Parse=Extract

# Method 1 (pure python built-in lib, no third-party lib): re
contentTitleP = '<span\s+class="title-content-title">(?P<contentTitle>[^<>]+)</span>'
allContentTitleList = re.findall(contentTitleP, baiduHtml)

# # Method 2 (use third-party lib): BeautifulSoup
# soup = BeautifulSoup(baiduHtml, 'html.parser')
# allTitleSoupList = soup.find_all("span", attrs={"class": "title-content-title"})
# print("allTitleSoupList=%s" % allTitleSoupList)
# allContentTitleList = []
# for eachTitleSoup in allTitleSoupList:
#     titleStr = eachTitleSoup.string

```

```
#     allContentTitleList.append(titleStr)

print("allContentTitleList=%s" % allContentTitleList)

# Step3: Save

# save to csv
OutputCsvHeader = ["序号", "百度热榜标题"]
OutputCsvFilename = "BaiduHotTitleList_%s.csv" % curDatetimeStr
OutputCsvFullPath = os.path.join(OutputRoot, OutputCsvFilename)

outputCsvDictList = []
for curIdx, eachTitle in enumerate(allContentTitleList):
    curNum = curIdx + 1
    csvDict = {
        "序号": curNum,
        "百度热榜标题": eachTitle
    }
    outputCsvDictList.append(csvDict)

saveToCsvByDictList(outputCsvDictList, OutputCsvFullPath)
print("Completed save data to %s" % OutputCsvFullPath)
```

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新: 2020-08-09 10:17:56

用第三方Python库

此处记录从无到有的核心过程：

先写出核心代码：

```
import requests

baiduResp = requests.get("https://www.baidu.com/")
baiduHtml = baiduResp.text
curDatetimeStr = getCurDatetimeStr()
baiduHtmlFilename = "baidu_com_%s.html" % curDatetimeStr
baiduHtmlFullPath = os.path.join(DebugRoot, baiduHtmlFilename)
saveTextToFile(baiduHtmlFullPath, baiduHtml)
```

调试返回的html源码是：

```
baidu_com_20200730_150323.html — demo_spider
debug > baidu_com_20200730_150323.html > ...
1  <!DOCTYPE html>
2  <!--STATUS OK--><html><head><meta http-equiv=content-type content=text/html;charset=utf-8><meta
http-equiv=X-UA-Compatible content=IE=Edge><meta content=always name=referrer><link rel=stylesheet type=text/css
href=https://ss1.bdstatic.com/5eNbjq8AAUyM2zgoY3K/r/www/cache/bdorz/baidu.min.css><title>ç¾ºä¹ä¸å, å¾ºä¹äº¤ç
¥é</title></head><body link=#0000c><div id=wrapper><div id=head><div class=head_wrapper><div class=s_form>
<div class=s_form_wrapper><div id=lg><img hiderefocus=true src=/www.baidu.com/img/bd_logo1.png width=270
height=129></div><form id=form name=f action=/www.baidu.com/s class=f><input type=hidden name=bdorz_come
value=1><input type=hidden name=ie value=utf-8><input type=hidden name=f value=8><input type=hidden name=rsv_bp
value=1><input type=hidden name=rsv_idx value=1><input type=hidden name=t value=baidu><span class=bg
s_ipt_wr><input id=kw name=wd class=s_ipt value maxlength=255 autocomplete=off autofocus=autofocus></span><span
class=bg_s_btn_wr><input type=submit id=su value=ç¾ºä¹ä¸å, å, class=bg s_btn" autofocus=</span></form></div>
</div><div id=u1><a href=http://news.baidu.com name=tj_trnews class=mnav>æ°é»</a><a href=https://www.hao123.
com name=tj_trhao123 class=mnav>hao123</a><a href=http://map.baidu.com name=tj_trmap class=mnav>å¾ºä¹ä¸å</a>
<a href=http://v.baidu.com name=tj_trvideo class=mnav>è§é¢</a><a href=http://tieba.baidu.com name=tj_trtieba
class=mnav>è‘‘å §</a><noscript><a href=http://www.baidu.com/bdorz/login.gif?login&tpl=mn&u=
u=http%3A%2F%2Fwww.baidu.com%2F3fbdf02_com%3d1 name=tj_login class=lb>ç »å½“</a></noscript><script>document.
write('<a href="http://www.baidu.com/bdorz/login.gif?login&tpl=mn&u='+ encodeURIComponent(window.location.href+
(window.location.search === "" ? "?" : "&") + "bdorz_come=1")+'"' name="tj_login" class="lb">ç »å½“</a>');
<script><a href="http://www.baidu.com/more/" name=tj_briicon class=bri style="display: block;
">æ å¾ºä¹ä¸å</a></div></div><div id=ftCon><div id=ftConv><p id=lh><a href=http://
home.baidu.com/å¾ºä¹ä¸å</a><a href=http://ir.baidu.com>About Baidu</a></p><p id=cp>&copy;
2017&nbsp;Baidu&nbsp;<a href=http://www.baidu.com/duty>å¼ç ç å¾ºä¹ä¸åè®“</a>&nbsp;<a
href=http://jianyu.baidu.com/ class=cp-feedback>è§å é</a>&nbsp;&nbsp;ICPè° 030173å·&nbsp;
<img src=/www.baidu.com/img/gs.gif></p></div></div></body></html>
```

很明显：没有包含我们希望的 百度热榜 的内容，且连其他的中文，比如 百度一下 之类的字眼都看不到

那么根据经验，需要加其他参数，甚至额外逻辑，才可能获取完整的html代码

而最先要去加上的，就是 User-Agent

先去回去用 Chrome的开发者工具，看看当前的User-Agent是啥，找到值。

再去把User-Agent部分，加到requests中：

```
UserAgent_Chrome_Mac = "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/84.0.4147.89 Safari/537.36"

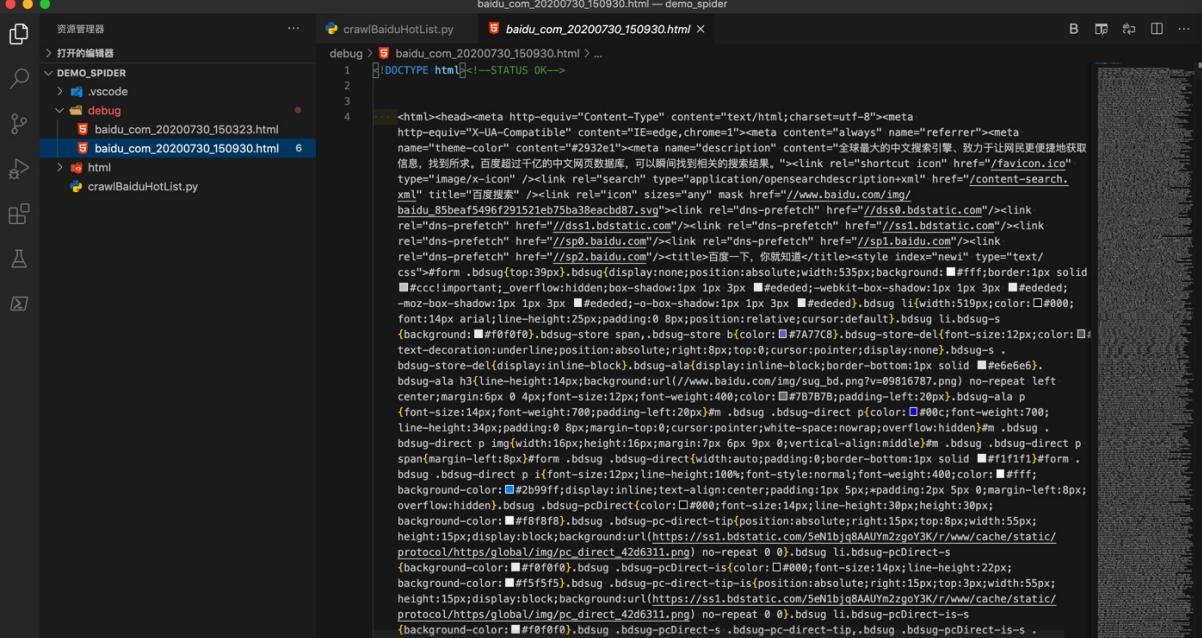
curHeaders = {
    "User-Agent": UserAgent_Chrome_Mac,
```

```

}
baiduResp = requests.get("https://www.baidu.com/", headers curHeaders)

```

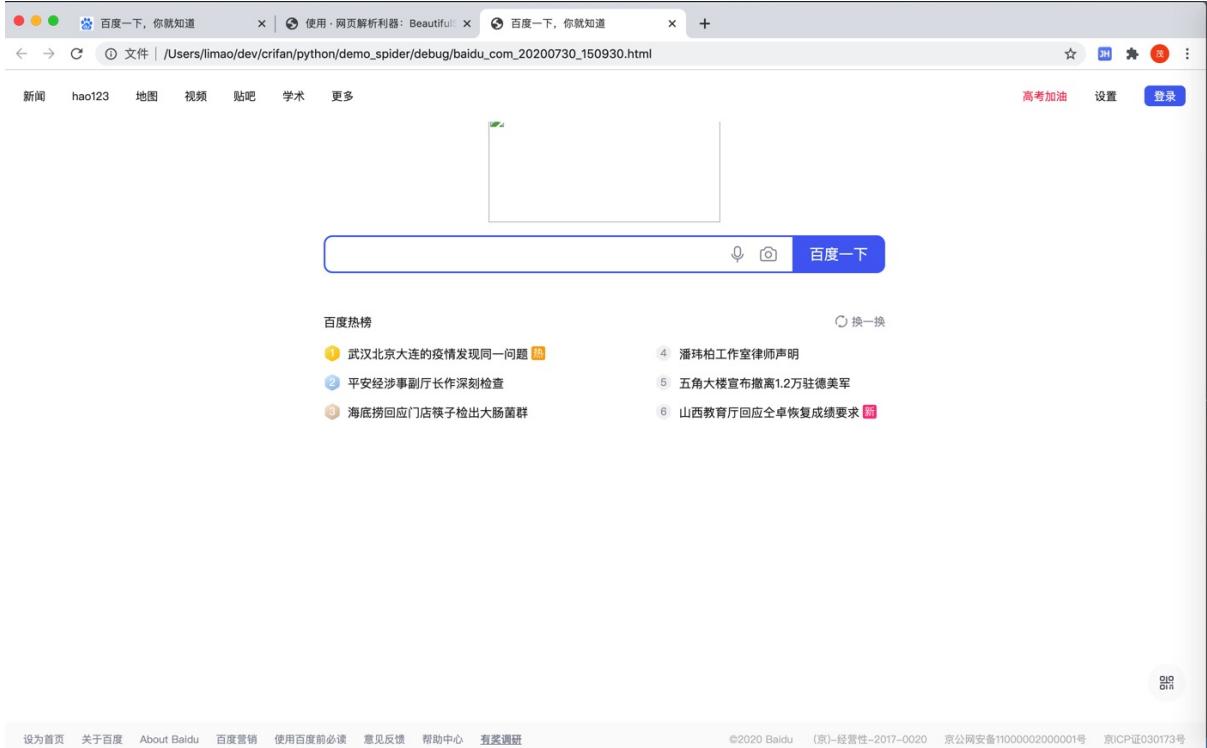
再去试试，此处我们很幸运，立刻就可以返回，大量的内容：



The screenshot shows a code editor with two tabs open: 'crawlBaiduHotList.py' and 'baidu_com_20200730_150930.html'. The 'baidu_com_20200730_150930.html' tab displays a large amount of HTML code, which is the captured content of the Baidu homepage. The code includes various meta tags, links, and the main content area.

看起来就是正确的，估计包含我们要找到的 百度热榜 的内容了。

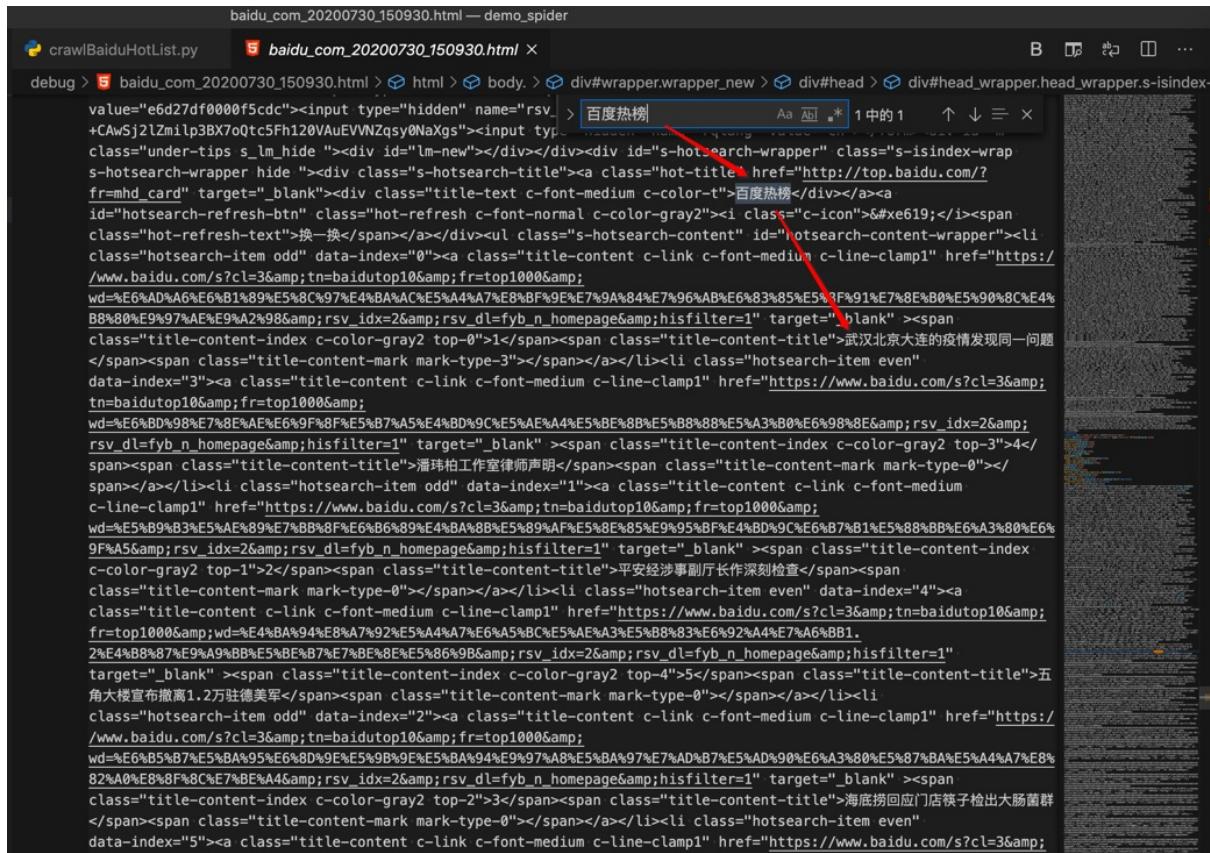
另外顺带，直接用浏览器打开此处抓取到的本地的离线的html，看看效果是什么样的：



可见（由于本身页面简单不复杂），除了首页logo外，页面效果和浏览器打开的基本一致。

也验证了前面的推测，确认就是完整的源码了。

去搜索 百度热榜



的确可以找到我们要的内容。

更多情况下获取完整的全部源码要难很多

此处只加了 User-Agent 就可以返回所需全部的完整的页面源码，是很幸运的。

因为随着web技术发展，反扒技术进步，稍微有点点技术含量的公司所做的web页面，尤其是页面逻辑复杂的，涉及到多个页面的

想要获取完整页面源码，往往都需要加上其他更多参数，才（可）能获取到期望的返回结果。

而关于 其他更多参数，常见的一些有：

- 简单的
 - Accept
 - Accept-Encoding
 - Accept-Language
 - Host
 - Referer
 - 复杂的
 - Cookie
 - 很多值，很难获取到（搞懂生成的逻辑）

所以可以接着，去用之前分析出的规则，去解析内容了。

此处用第三方Python的HTML解析库 BeautifulSoup :

```
# Method 2: BeautifulSoup
soup = BeautifulSoup(baiduHtml, 'html.parser')
allTitleSoupList = soup.find_all("span", attrs {"class": "title-content-title"})
print("allTitleSoupList=%s" % allTitleSoupList)
```

可以解析到所需内容：

```
crawlBaiduHotList.py — demo_spider
crawlBaiduHotList.py x baidu_com_20200730_150930.html
crawlBaiduHotList.py > ...
60 UserAgent_Chrome_Mac = "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/85.0.4183.122 Safari/537.36"
61 curHeaders = {
62     "User-Agent": UserAgent_Chrome_Mac,
63 }
64 baiduResp = requests.get("https://www.baidu.com/", headers=curHeaders)
65 baiduHtml = baiduResp.text
66 curDatetimeStr = getCurDatetimeStr()
67 baiduHtmlFilename = "baidu_com_%s.html" % curDatetimeStr
68 baiduHtmlFullPath = os.path.join(DebugRoot, baiduHtmlFilename)
69 saveTextToFile(baiduHtmlFullPath, baiduHtml)
70 # Step2: Parse
71 # Method 1: re
72 # contentTitleP = '<span>s+class="title-content-title"</span>'.re.findall
73 # allContentTitleList = re.findall(contentTitleP, baiduHtml)
74 # print("allContentTitleList=%s" % allContentTitleList)
75 # Method 2: BeautifulSoup
76 soup = BeautifulSoup(baiduHtml, 'html.parser')
77 allTitleSoupList = soup.find_all("span", attrs {"class": "title-content-title"})
78 print("allTitleSoupList=%s" % allTitleSoupList)
79 # Step3: Save
80
81
82
83
84 # Step4: Save
```

再去加上代码，把 soup 的 string 保存出来：

```
allContentTitleList = []
for eachTitleSoup in allTitleSoupList:
    titleStr = eachTitleSoup.string
    allContentTitleList.append(titleStr)
print("allContentTitleList=%s" % allContentTitleList)
```

就是我们要的列表了：

```
allContentTitleList=['武汉北京大连的疫情发现同一问题', '潘玮柏工作室律师声明', '平安经涉事副厅长作深刻检查', '五角大楼宣布撤离1.2万驻德美军', '海底捞回应门店筷子检出大肠菌群', '山西教育厅回应全卓恢复成绩要求']
```

至此下载和提取都完成了

接着去保存内容，如前面假设，比如保存到csv文件中

```
def saveToCsvByDictList(csvDictList, outputPath):
    # generate csv headers from dict list
    firstItemDict = csvDictList[0]
    csvHeaders = list(firstItemDict.keys())
    with codecs.open(outputPath, "w", "UTF-8") as outCsvFp:
        csvDictWriter = csv.DictWriter(outCsvFp, fieldnames=csvHeaders)

        # write header by inner function from fieldnames
        csvDictWriter.writeheader()
```

```

        for eachRowDict in csvDictList:
            csvDictWriter.writerow(eachRowDict)

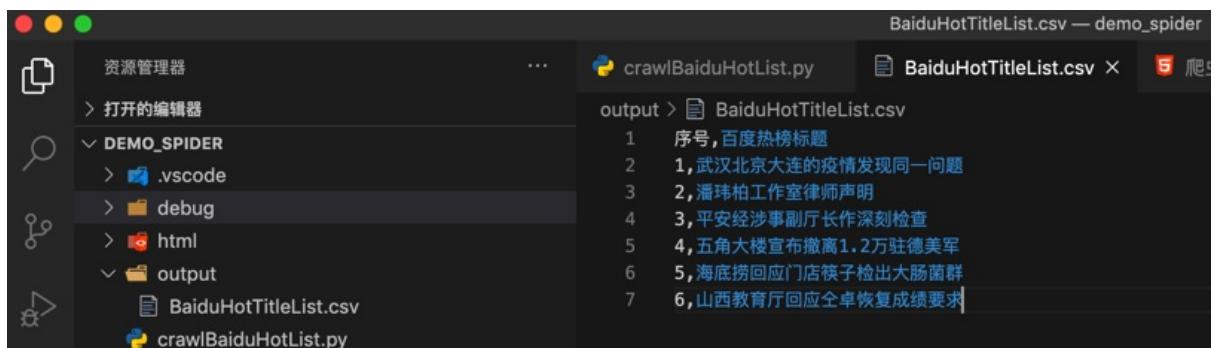
# save to csv
OutputCsvHeader = ["序号", "百度热榜标题"]
OutputCsvFilename = "BaiduHotTitleList.csv"
OutputCsvFullPath = os.path.join(OutputRoot, OutputCsvFilename)

outputCsvDictList = []
for curIdx, eachTitle in enumerate(allContentTitleList):
    curNum = curIdx + 1
    csvDict = {
        "序号": curNum,
        "百度热榜标题": eachTitle
    }
    outputCsvDictList.append(csvDict)

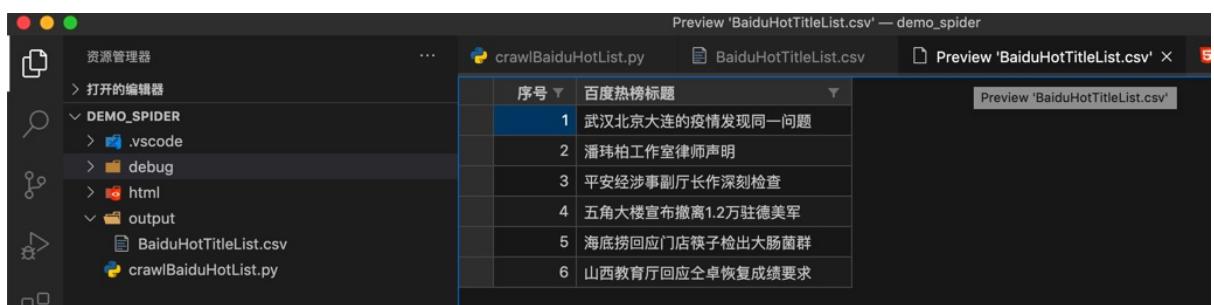
saveToCsvByDictList(outputCsvDictList, OutputCsvFullPath)

```

即可保存出我们要的csv文件：



以及，用VSCode中csv插件去以列表方式查看的效果：



和Mac中的预览效果：



至此，实现完整的爬虫功能：

- 下载百度首页源码
- 提取所需的百度热榜的标题内容
- 保存内容为csv格式

完整代码

```

# Function: Demo how use Python crawl baidu.com 百度热榜
# Author: Crifan
# Update: 20200731

import os
import codecs
from datetime import datetime, timedelta

# import urllib.request
# import re

import requests
from bs4 import BeautifulSoup

import csv

DebugRoot = "debug"
OutputRoot = "output"

#####
# Utils Functions
#####

def createFolder(folderFullPath):
    """
    """

```

```

        create folder, even if already existed
        Note: for Python 3.2+
"""

os.makedirs(folderFullPath, exist_ok=True)

def saveTextToFile(fullFilename, text, fileEncoding="utf-8"):
    """save text content into file"""
    with codecs.open(fullFilename, 'w', encoding=fileEncoding) as fp:
        fp.write(text)
        fp.close()

def datetimeToStr(inputDatetime, format="%Y%m%d_%H%M%S"):
    """Convert datetime to string

Args:
    inputDatetime (datetime): datetime value
Returns:
    str
Raises:
Examples:
    datetime.datetime(2020, 4, 21, 15, 44, 13, 2000) -> '20200421_154413'
"""

datetimeStr = inputDatetime.strftime(format=format)
# print("inputDatetime=%s -> datetimeStr=%s" % (inputDatetime, datetimeStr)) # 2020-04
-21 15:08:59.787623
return datetimeStr

def getCurDatetimeStr(outputFormat="%Y%m%d_%H%M%S"):
    """
get current datetime then format to string

eg:
    20171111_220722

:param outputFormat: datetime output format
:return: current datetime formatted string
"""

curDatetime = datetime.now() # 2017-11-11 22:07:22.705101
# curDatetimeStr = curDatetime.strftime(format=outputFormat) #'20171111_220722'
curDatetimeStr = datetimeToStr(curDatetime)
return curDatetimeStr

def saveToCsvByDictList(csvDictList, outputPath):
    # generate csv headers from dict list
    firstItemDict = csvDictList[0]
    csvHeaders = list(firstItemDict.keys())
    with codecs.open(outputPath, "w", "UTF-8") as outCsvFp:
        csvDictWriter = csv.DictWriter(outCsvFp, fieldnames=csvHeaders)

        # write header by inner function from fieldnames
        csvDictWriter.writeheader()

        for eachRowDict in csvDictList:

```

```

        csvDictWriter.writerow(eachRowDict)

def saveToCsvByHeaderAndList(csvHeaderList, csvRowListList, outputPath):
    with codecs.open(outputPath, "w", "UTF-8") as outCsvFp:
        csvWriter = csv.writer(outCsvFp)

        # write header from list
        csvWriter.writerow(csvHeaderList)

        # type1: write each row
        # for eachRowList in csvRowListList:
        #     csvWriter.writerow(eachRowList)

        # type2: write all rows
        csvWriter.writerows(csvRowListList)

#####
# Main
#####

createFolder(DebugRoot)
createFolder(OutputRoot)

curDatetimeStr = getCurDatetimeStr()

# Step1: Download
UserAgent_Chrome_Mac = "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/84.0.4147.89 Safari/537.36"

curHeaders = {
    "User-Agent": UserAgent_Chrome_Mac,
}

baiduUrl = "https://www.baidu.com/"

# # Method 1 (pure python built-in lib, no third-party lib): urllib
# baiduResp = urllib.request.urlopen(baiduUrl)
# baiduHtmlBytes = baiduResp.read()
# baiduHtml = baiduHtmlBytes.decode()

# Method 2 (use third-party lib): requests
baiduResp = requests.get(baiduUrl, headers curHeaders)
baiduHtml = baiduResp.text

# for debug
baiduHtmlFilename = "baidu_com_%s.html" % curDatetimeStr
baiduHtmlFullPath = os.path.join(DebugRoot, baiduHtmlFilename)
saveTextToFile(baiduHtmlFullPath, baiduHtml)

# Step2: Parse=Extract

# # Method 1 (pure python built-in lib, no third-party lib): re
# contentTitleP = '<span\s+class="title-content-title">(?P<contentTitle>[^<>]+)</span>'
```

```
# allContentTitleList = re.findall(contentTitleP, baiduHtml)

# Method 2 (use third-party lib): BeautifulSoup
soup = BeautifulSoup(baiduHtml, 'html.parser')
allTitleSoupList = soup.find_all("span", attrs {"class": "title-content-title"})
print("allTitleSoupList=%s" % allTitleSoupList)
allContentTitleList = []
for eachTitleSoup in allTitleSoupList:
    titleStr = eachTitleSoup.string
    allContentTitleList.append(titleStr)

print("allContentTitleList=%s" % allContentTitleList)

# Step3: Save

# save to csv
OutputCsvHeader = ["序号", "百度热榜标题"]
OutputCsvFilename = "BaiduHotTitleList_%s.csv" % curDatetimeStr
OutputCsvFullPath = os.path.join(OutputRoot, OutputCsvFilename)

outputCsvDictList = []
for curIdx, eachTitle in enumerate(allContentTitleList):
    curNum = curIdx + 1
    csvDict = {
        "序号": curNum,
        "百度热榜标题": eachTitle
    }
    outputCsvDictList.append(csvDict)

saveToCsvByDictList(outputCsvDictList, OutputCsvFullPath)
print("Completed save data to %s" % OutputCsvFullPath)
```

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新: 2020-08-09 10:17:56

用爬虫框架

此处再去把同样爬虫功能，换成第三方的爬虫框架PySpider去实现。

- 准备工作

- 安装： pip install pyspider
- 启动： pyspider

经过调试，效果是：

调试时能看到输出多个message的结果：

```

不安全 — 0.0.0:5000/debug/crawlBaiduHotList_PySpider_1501
pyspider > crawlBaiduHotList_PySpider_1501
run Documentation WebDAV Mode save
{
  "fetch": {
    "headers": {
      "User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/84.0.4147.89 Safari/537.36"
    }
  },
  "process": {
    "callback": "baiduHome"
  },
  "project": "crawlBaiduHotList_PySpider_1501",
  "schedule": {},
  "taskid": "e8lc1f5749545c5f7d247b3a100ff662",
  "url": "https://www.baidu.com/"
}

titleItemList=[(<span>title-content-title</span>), (<span>title-content-title</span>), (<span>title-content-title</span>)]
[0] eachItem<span class="title-content-title">北斗正式开通</span>
itemTitleStr=北斗正式开通
curlUrl=https://www.baidu.com/#0
[1] eachItem<span class="title-content-title">阿云嘎粉丝后援会闭站</span>
itemTitleStr=阿云嘎粉丝后援会闭站
curlUrl=https://www.baidu.com/#1
[2] eachItem<span class="title-content-title">赵正永一审被判死缓</span>
itemTitleStr=赵正永一审被判死缓
curlUrl=https://www.baidu.com/#2
[3] eachItem<span class="title-content-title">19岁贫困男孩高考后查出白血病</span>
itemTitleStr=19岁贫困男孩高考后查出白血病
curlUrl=https://www.baidu.com/#3
[4] eachItem<span class="title-content-title">特朗普改口称不想推迟选举</span>
itemTitleStr=特朗普改口称不想推迟选举
curlUrl=https://www.baidu.com/#4
[5] eachItem<span class="title-content-title">2020ChinaJoy逛展指南</span>
itemTitleStr=2020ChinaJoy逛展指南
curlUrl=https://www.baidu.com/#5

```

显示菜单 enable css selector helper web html follows messages 6

去 Run 运行项目：

scheduler	0	fetcher	0	processor	0	result_worker
			0 + 0			

Recent Active Tasks

group	project name	status	rate/burst	avg time	progress	actions
(group)	crawlBaiduHotList_PySpider_1501	RUNNING	1/3	51.9+4.53	<div style="width: 33.33%;">5m: 6 1h: 6 1d: 6 all: 3</div>	Run Active Tasks Results

all of 3 tasks:
pending(0.0%): 0
success(100.0%): 3
retry(0.0%): 0
failed(0.0%): 0

运行完毕后，点击 Results，进入结果页面：

三种实现方式

url	百度热榜标题	...
https://www.baidu.com/#5	"2020ChinaJoy逛展指南"	0
https://www.baidu.com/#4	"特朗普改口称不想推迟选举"	0
https://www.baidu.com/#3	"19岁贫困男孩高考后查出白血病"	0
https://www.baidu.com/#2	"赵正永一审被判死缓"	0
https://www.baidu.com/#1	"阿云嘎粉丝后援会闭站"	0
https://www.baidu.com/#0	"北斗正式开通"	0

点击 csv 显示 (也可以保存下载) 结果:

```
url,百度热榜标题,…
https://www.baidu.com/#5,2020ChinaJoy逛展指南,{}
https://www.baidu.com/#4,特朗普改口称不想推迟选举,{}
https://www.baidu.com/#3,19岁贫困男孩高考后查出白血病,{}
https://www.baidu.com/#2,赵正永一审被判死缓,{}
https://www.baidu.com/#1,阿云嘎粉丝后援会闭站,{}
https://www.baidu.com/#0,北斗正式开通,{}
```

下载或拷贝出来，放到VSCode中，预览效果为：

即可实现最终要的结果。

完整代码

```
#!/usr/bin/env python
# -*- encoding: utf-8 -*-
# Created on 2020-07-31 15:01:00
# Project: crawlBaiduHotList_PySpider_1501

from pyspider.libs.base_handler import *
from pyspider.database import connect_database

class Handler(BaseHandler):
    crawl_config = {
    }

    # @every(minutes=24 * 60)
    def on_start(self):
        UserAgent_Chrome_Mac = "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/84.0.4147.89 Safari/537.36"
        curHeaderDict = {
            "User-Agent": UserAgent_Chrome_Mac,
        }
```

```
        self.crawl('https://www.baidu.com/', callback self.baiduHome, headers curHeaderDict
    )

# @config(age=10 * 24 * 60 * 60)
def baiduHome(self, response):
    # for eachItem in response.doc('span[class="title-content-title"]').items():

        titleItemGenerator = response.doc('span[class="title-content-title"]').items()
        titleItemList = list(titleItemGenerator)
        print("titleItemList=%s" % titleItemList)
        # for eachItem in titleItemList:
        for curIdx, eachItem in enumerate(titleItemList):
            print("[%d] eachItem=%s" % (curIdx, eachItem))
            itemTitleStr = eachItem.text()
            print("itemTitleStr=%s" % itemTitleStr)
            curUrl = "%s%d" % (response.url, curIdx)
            print("curUrl=%s" % curUrl)
            curResult = {
                # "url": response.url,
                # "url": curUrl,
                "百度热榜标题": itemTitleStr,
            }
            # return curResult
            # self.send_message(self.project_name, curResult, url=response.url)
            self.send_message(self.project_name, curResult, url=curUrl)

    def on_message(self, project, msg):
        print("on_message: msg=%s", msg)
        return msg
```

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新: 2020-08-09 10:17:56

附录

下面列出相关参考资料。

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新: 2019-03-29 21:30:12

参考资料

- 【记录】演示如何实现简单爬虫：用Python提取百度首页中百度热榜内容列表
- 【已解决】用Python爬虫框架PySpider实现爬虫爬取百度热榜内容列表
- 【已解决】PySpider中如何在单个页面返回多个结果保存到自带的Results页面中的列表中
- 【已解决】PySpider抓包百度热榜标题列表结果
- 【已解决】Mac中安装phantomjs
- 【已解决】Mac中启动PySpider
- 【已解决】Mac中pip安装pycurl报错：fatal error openssl/ssl.h file not found
- 【已解决】Mac中给Python3安装PySpider
- 【已解决】用Python纯内置库无第三方库实现爬虫爬取百度热榜内容列表
- 【已解决】用Python3的urllib下载百度首页源码
- 【已解决】Mac中用Chrome开发者工具分析百度首页的百度热榜内容加载逻辑
- 【已解决】用Python代码获取到百度首页源码并提取保存百度热榜内容列表
-
- 爬取你要的数据：[爬虫技术](#)
- [crifanLibPython](#)
- [getUrlRespHtml](#)
- Python中的正则表达式：[re模块详解](#)
- Python心得：[操作CSV和Excel](#)
- Python心得：[http网络库](#)
- Python专题教程：[BeautifulSoup详解](#)
- Python心得：[HTML解析库PyQuery](#)
- 【记录】Python中尝试用lxml去解析html – 在路上 [在路上](#)
- 主流关系数据库：[MySQL](#)
- 主流文档型数据库：[MongoDB](#)
- Python爬虫框架：[PySpider](#)
- 主流Python爬虫框架：[Scrapy](#)
- 【整理】[pyspider vs scrapy](#)
- 【教程】模拟登陆网站 之 Python版（内含两种版本的完整的可运行的代码） – 在路上 [在路上](#)
- Python专题教程：[抓取网站，模拟登陆，抓取动态网页](#)
- 【整理】各种浏览器中的开发人员工具Developer Tools：[IE9的F12, Chrome的Ctrl+Shift+J, Firefox的Firebug](#)
- 【总结】浏览器中的开发人员工具（IE9的F12和Chrome的Ctrl+Shift+I） -网页分析的利器
- 【教程】如何利用IE9的F12去分析网站登陆过程中的复杂的（参数， cookie等）值（的来源）
- 【教程】手把手教你如何利用工具(IE9的F12)去分析模拟登陆网站(百度首页)的内部逻辑过程
- app抓包利器：[Charles](#)
- 【已解决】写Python爬虫爬取汽车之家品牌车系车型数据 – 在路上 [在路上](#)
- 【记录】Mac中安装和运行pyspider
- 【已解决】pyspider中如何写规则去提取网页内容
- 【已解决】pyspider中如何加载汽车之家页面中的更多内容
- 【已解决】PySpider如何把json结果数据保存到csv或excel文件中

- [【已解决】PySpider中如何清空之前运行的数据和正在运行的任务](#)
- [【已解决】Python中实现带Cookie的Http的Post请求 – 在路上](#)
- [【已解决】Python中如何获得访问网页所返回的cookie – 在路上](#)
-
- [Requests](#)
- [re](#)
- [aiohttp](#)
- [PyMySQL](#)
- [PyMongo](#)
- [urllib](#)
- [BeautifulSoup](#)
- [PyQuery](#)
- [lxml](#)
- [PySpider](#)
- [Scrapy](#)
- [Chrome 开发者工具 | Tools for Web Developers](#)
- [rmax/scrapy-redis: Redis-based components for Scrapy.](#)
- [grangier/python-goose: Html Content / Article Extractor, web scrapping lib in Python](#)
- [Bloom Filters by Example](#)
- [Bloom Filters by Example 中文](#)
- [Scrapy入门教程 — Scrapy 0.24.6 文档](#)
- [Scrapy爬虫框架教程（一）-- Scrapy入门](#)
-

crifan.com, 使用[署名4.0国际\(CC BY 4.0\)协议](#)发布 all right reserved, powered by Gitbook最后更新: 2020-08-09 10:17:56