



第3章 Python语言入门与Web信息解析

3.1 Python语言入门

3.2 Python语言进阶

3.3 HTTP解析与Python实现

3.4 HTML解析与Python实现

3.5 Web信息解析

《Python 爬虫开发与项目实战》 范传辉编著 机械工业出版社
2017年11月第一版



3.3 HTTP解析与Python实现

3.3.1 HTTP协议

3.3.2 Python-Requests解析HTTP方法

3.3.3 实例

- 1.HTTP概述
- 2.请求报文
- 3.响应报文
- 4.Cookies处理

3.3.1 HTTP协议-1.概述

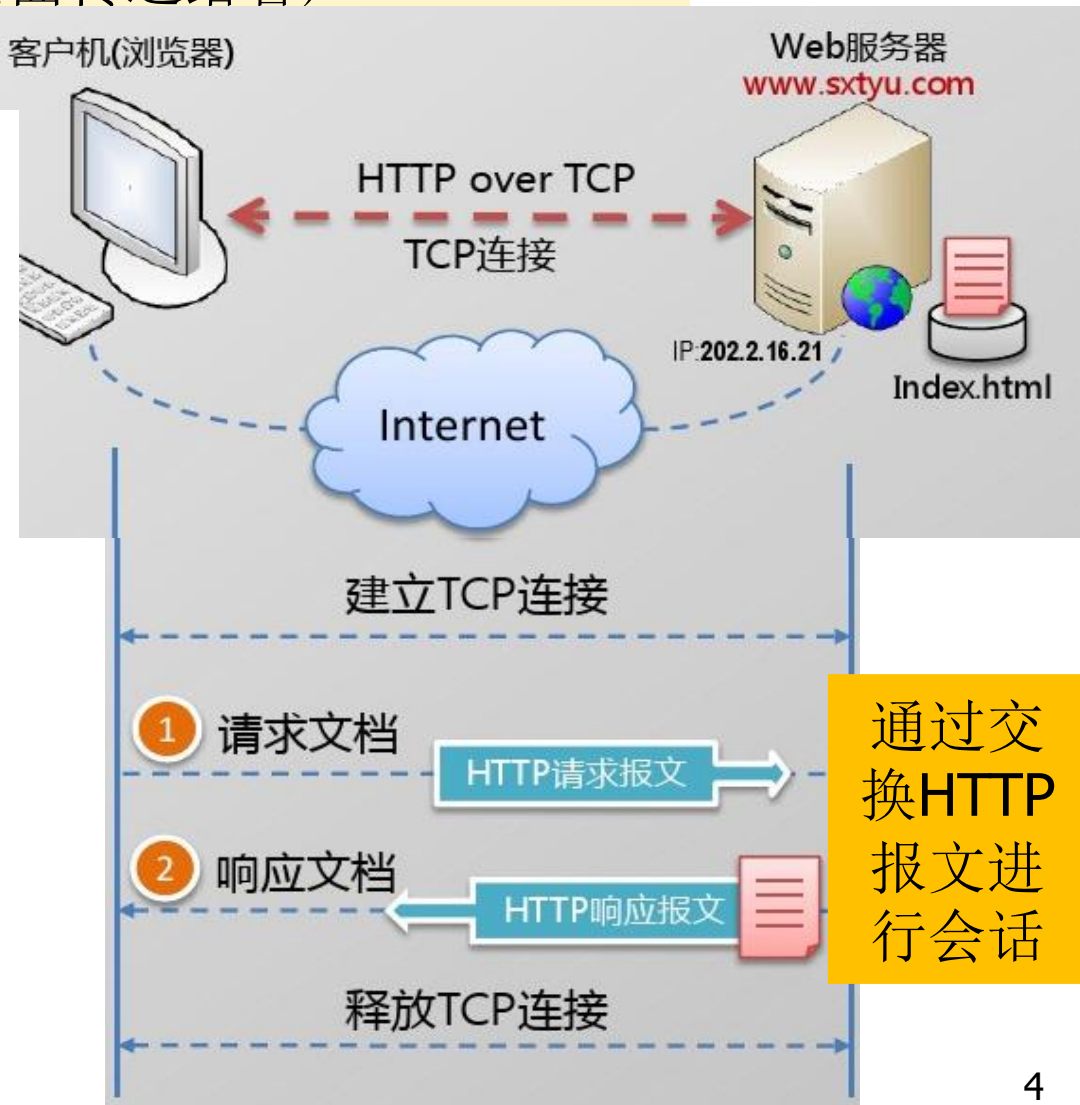
- Web网页提取，首先要和HTTP协议打交道
- HTTP协议（HyperText Transfer Protocol:超文本传输协议）
 - TCP/IP协议栈应用层协议
 - 用于从WWW服务器传输超文本到本地浏览器的传输协议
 - HTTP协议采取的是请求/响应模型，客户端发起请求，服务器回送响应



- HTTP协议定义了：
 - Web客户如何向Web服务器请求Web页面
 - 服务器如何将Web页面传送给客户
 - 交互报文的格式

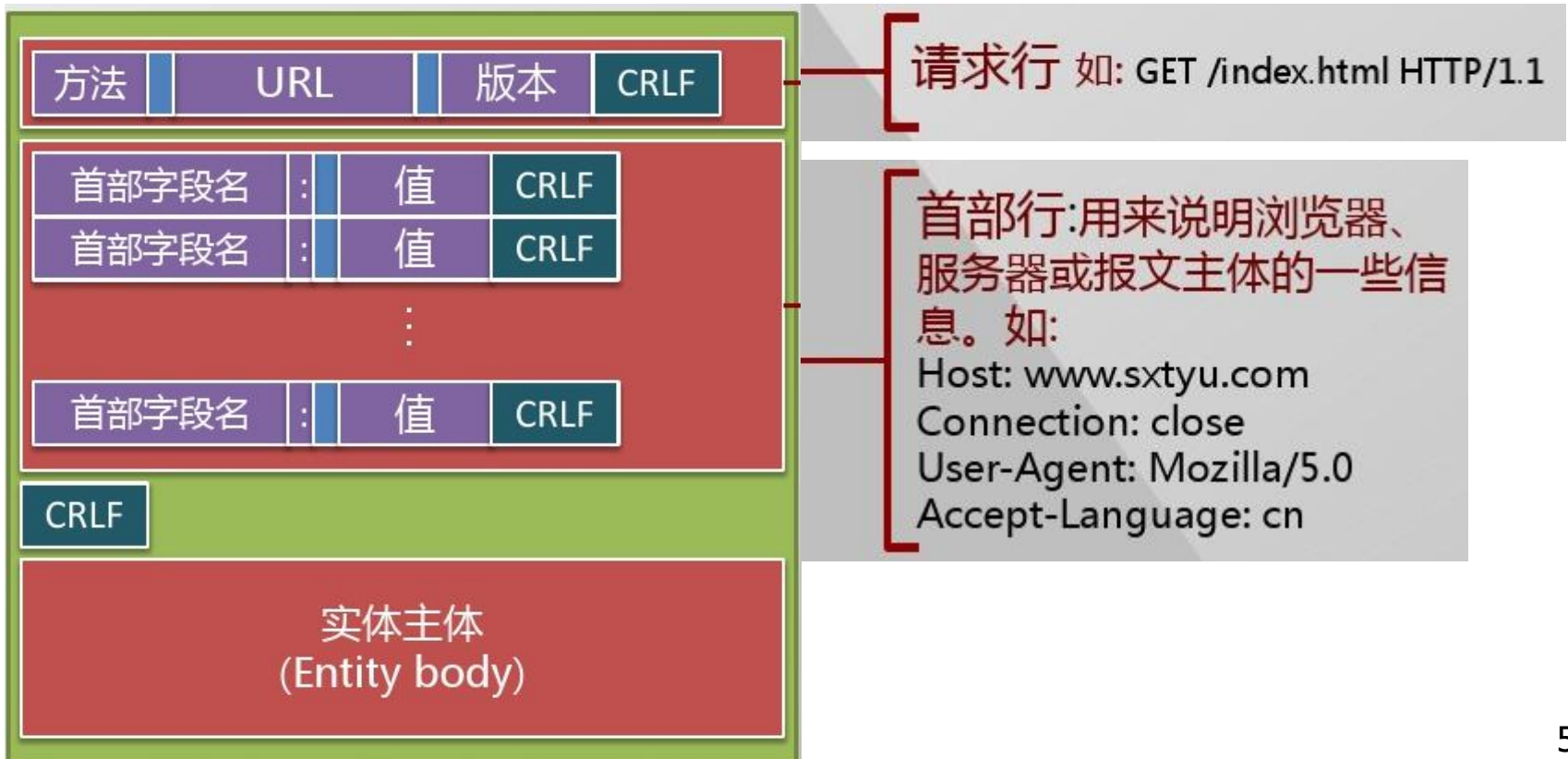
■ HTTP请求过程

- 首先客户端与服务器需要建立TCP连接
- 建立连接后，客户端发送一个请求给服务器
- 服务器接到请求后，给予相应的响应信息
- 客户端接收服务器所返回的信息，通过浏览器将信息显示在用户的显示屏上，然后客户端与服务器断开连接。



3.3.1 HTTP协议-请求报文

- 请求报文
- 三个部分组成：请求行、首部行、请求（实体）主体



3.3.1 HTTP协议-请求报文

- 请求行：方法、URL、HTTP版本

方法

URL

版本

CRLF

请求行 如: GET /index.html HTTP/1.1

方法(Method)是对所请求对象所进行的操作,也就是一些命令。

方法(操作)	含义	方法(操作)	含义
GET	请求读取一个Web页面	HEAD	请求读取一个Web页面的首部
POST	附加一个命名资源(如Web页面)	PUT	请求存储一个Web页面
DELETE	删除Web页面	TRACE	用于测试, 要求服务器送回收到的请求
CONNECT	用于代理服务器	OPTION	查询特定选项



■ 最常用的两种方法：

- **GET方式**：向目的服务器请求URL所指定资源
- **POST方式**：向目的服务器发出请求，要求它接受被附在请求后的实体，并把它当作请求队列中请求URL所指定资源的附加新子项。

■ GET与POST方法区别：

- 在客户端，**Get**方式通过URL提交数据，数据在URL中可以看到；**POST**方式，数据放置在实体区内提交。
- **GET**方式提交的数据最多只能有**1024**字节，而**POST**则没有此限制。
- 安全性问题。使用**Get**的时候，参数会显示在地址栏上，而**Post**不会。所以，如果这些数据是非敏感数据，那么使用**Get**；如果用户输入的数据包含敏感数据，那么还是使用**Post**为好。

- 在爬虫开发中基本处理的也是GET和POST请求
- GET请求在访问网页时很常见，POST请求则是常用在登录框、提交框的位置使用
- 例.一个完整的POST请求
 - 登录知乎社区时捕获的请求

```
POST /login/phone_num HTTP/1.1
Host: www.zhihu.com
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:49.0) Gecko/20100101 Firefox/49.0
Accept: */*
Accept-Language: zh-CN,zh;q=0.8,en-US;q=0.5,en;q=0.3
Accept-Encoding: gzip, deflate, br
X-Xsrftoken: ade0896dc13cc3b2204a8f7742ad7f48
Content-Type: application/x-www-form-urlencoded; charset=UTF-8
X-Requested-With: XMLHttpRequest
Referer: https:// www.zhihu.com/
Content-Length: 117
Cookie: q_c1=7bc53a12dd7942d3b64776441ab69983|1477975324000|1465870098000; d_c0="ACAAa1M-EwqPTgdv2RIP3IIzHwN2ZhYzgyZmEx0TE=|1477975348|735a805117328df9e557f0126eb348e7712e310c"
Connection: keep-alive
```

```
_xsrf=ade0896dc13cc3b2204a8f7742ad7f48&password=xxxxxxx&captcha_type=cn&remember_me=true&phone_num=xxxxx:
```


■ 请求首部行字段：

User- Agent	关于浏览器和它平台的信息，如Mozilla5.0
Accept	客户能处理的页面的类型，如text/html
Accept-Charset	客户可以接受的字符集，如Unicode-1-1
Accept-Encoding	客户能处理的页面编码方法，如gzip
Accept-Language	客户能处理的自然语言，如en(英语)，zh-cn(简体中文)
Host	服务器的DNS名称。从URL中提取出来，必需。
Authorization	客户的信息凭据列表
Cookie	将以前设置的Cookie送回服务器，可用来作为会话信息
Date	消息被发送时的日期和时间

```
Host: www.cnblogs.com
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:49.0) Gecko/20100101 Firefox/49.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: zh-CN,zh;q=0.8,en-US;q=0.5,en;q=0.3
Accept-Encoding: gzip, deflate
Connection: keep-alive
If-Modified-Since: Sun, 30 Oct 2016 10:13:18 GMT
```

■ 请求首部行字段：

Host: www.cnblogs.com

User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:49.0) Gecko/20100101 Firefox/49.0

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8

Accept-Language: zh-CN,zh;q=0.8,en-US;q=0.5,en;q=0.3

Accept-Encoding: gzip, deflate

Connection: keep-alive

If-Modified-Since: Sun, 30 Oct 2016 10:13:18 GMT

• **Connection:** 允许发送用于指定连接的选项。例如指定连接的状态是连续，或者指定“**close**”选项，通知服务器，在响应完成后，关闭连接。

• **If-Modified-Since:** 用于在发送HTTP请求时，把浏览器端缓存页面的最后修改时间一起发到服务器去，服务器会把这个时间与服务器上实际文件的最后修改时间进行比较。如果时间一致，那么返回HTTP状态码**304**（不返回文件内容），客户端收到之后，就直接把本地缓存文件显示到浏览器中。如果时间不一致，就返回HTTP状态码**200**和新的文件内容，客户端收到之后，会丢弃旧文件，把新文件缓存起来，并显示到浏览器中。

3.3.1 HTTP协议

- 请求报文：请求(实体)主体
 - 若方法字段是GET，则此项为空，没有数据
 - 若方法字段是POST,则通常来说此处放置的就是要提交的数据



比如：

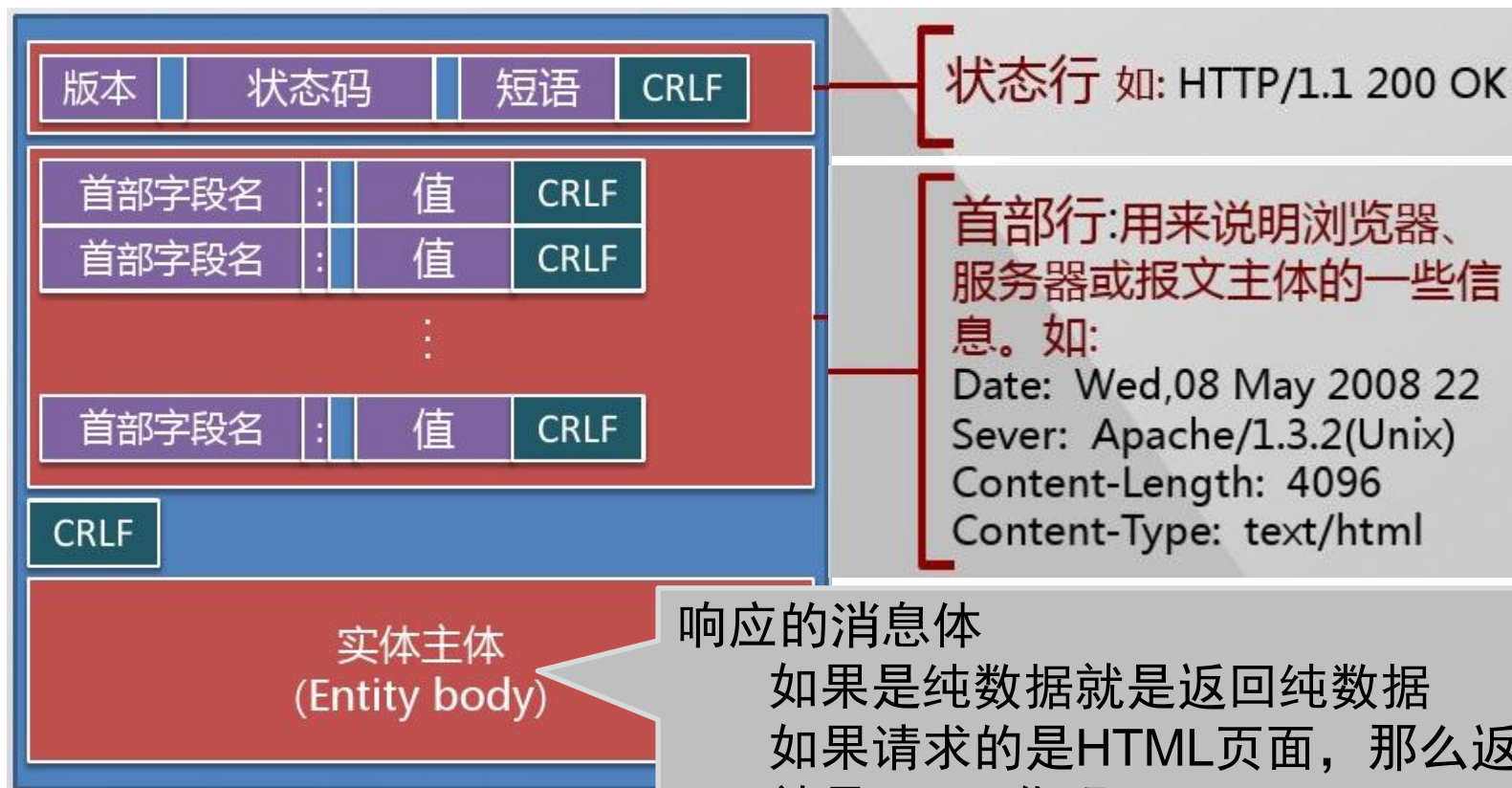
要使用POST方法提交一个表单，其中有user字段中数据为“admin”，password字段为123456，那么这里的请求数据就是

user=admin&password=123456

使用&来连接各个字段。

3.3.1 HTTP协议-响应报文

- 响应报文
- 三个部分组成：状态行、响应头、响应(实体)主体



响应的消息体

如果是纯数据就是返回纯数据
如果请求的是HTML页面，那么返回的就是HTML代码
如果是js就是JS代码，如此之类

3.3.1 HTTP协议-响应报文

- 响应报文状态行：版本、状态码、状态信息




状态码(Status-Code)是响应报文状态行中包含的一个3位数字，指明特定的请求是否被满足，如果没有满足，原因是什么。状态码分为以下五类：

状态码	含义	例子
1xx	通知信息	100=服务器正在处理客户请求
2xx	成功	200=请求成功(OK)
3xx	重定向	301=页面改变了位置
4xx	客户错误	403=禁止的页面；404=页面未找到
5xx	服务器错误	500=服务器内部错误；503=以后再试

■ 响应首部行字段：

Server	关于服务器的信息，如Microsoft-IIS/6.0
Content-Encoding	内容是如何被编码的（如gzip）
Content-Language	页面所使用的自然语言
Content-Length	以字节计算的页面长度
Content-Type	页面的MIME类型
Last-Modified	页面最后被修改的时间和日期，在页面缓存机制中意义重大
Location	指示客户将请求发送给别处，即重定向到另一个URL
Set-Cookie	服务器希望客户保存一个Cookie
Date	消息被发送时的日期和时间

```
Date: Sun, 30 Oct 2016 10:13:50 GMT
Content-Type: text/html; charset=utf-8
Transfer-Encoding: chunked
Connection: keep-alive
Vary: Accept-Encoding
Cache-Control: public, max-age=3
Expires: Sun, 30 Oct 2016 10:13:54 GMT
Last-Modified: Sun, 30 Oct 2016 10:13:24 GMT
Content-Encoding: gzip
```



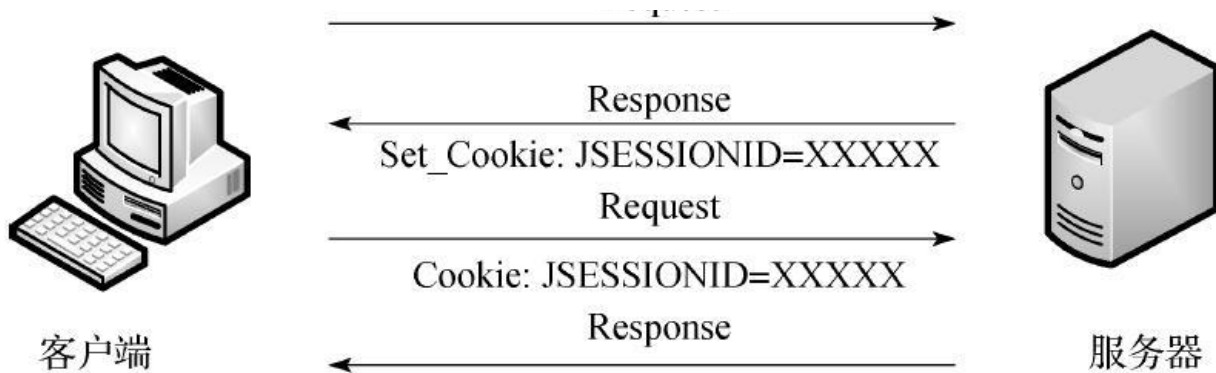
```
Date: Sun, 30 Oct 2016 10:13:50 GMT
Content-Type: text/html; charset=utf-8
Transfer-Encoding: chunked
Connection: keep-alive
Vary: Accept-Encoding
Cache-Control: public, max-age=3
Expires: Sun, 30 Oct 2016 10:13:54 GMT
Last-Modified: Sun, 30 Oct 2016 10:13:24 GMT
Content-Encoding: gzip
```

- Transfer-Encoding: chunked表示输出的内容长度不能确定。
- Connection: 允许发送用于指定连接的选项。例如指定连接的状态是连续，或者指定“close”选项，通知服务器，在响应完成后，关闭连接
- Vary: 指定了一些请求头域，这些请求头域用来决定当缓存中存在一个响应，并且该缓存没有过期失效时，是否被允许利用此响应去回复后续请求而不需要重复验证。
- Cache-Control: 用于指定缓存指令，缓存指令是单向的，且是独立的
- Expires: 给出响应过期的日期和时间。为了让代理服务器或浏览器在一段时间以后更新缓存中（再次访问曾访问过的页面时，直接从缓存中加载，缩短响应时间和降低服务器负载）的页面，我们可以使用Expires实体报头域指定页面过期的时间。

3.3.1 HTTP协议-Cookie处理

■ Cookie状态管理

- Cookie和Session都用来保存状态信息，都是保存客户端状态的机制，是为了解决HTTP无状态的问题
- Cookie的工作方式：
 - 服务器给每个Session分配一个唯一的JSESSIONID，并通过Cookie发送给客户端
 - 当客户端发起新的请求的时候，将在Cookie头中携带这个JSESSIONID





3.3.1 HTTP协议-Cookie处理

- Cookie是服务器在本地机器上存储的小段文本并随每一个请求发送至同一个服务器
- **cookie数据格式(属性):**
 - 域名"Domain"
 - 路径"Path"
 - 内容"Content"
 - 过期时间"Expires"
 - 安全域"Secure"

内容
Cart=1-21652465; 2-3468654

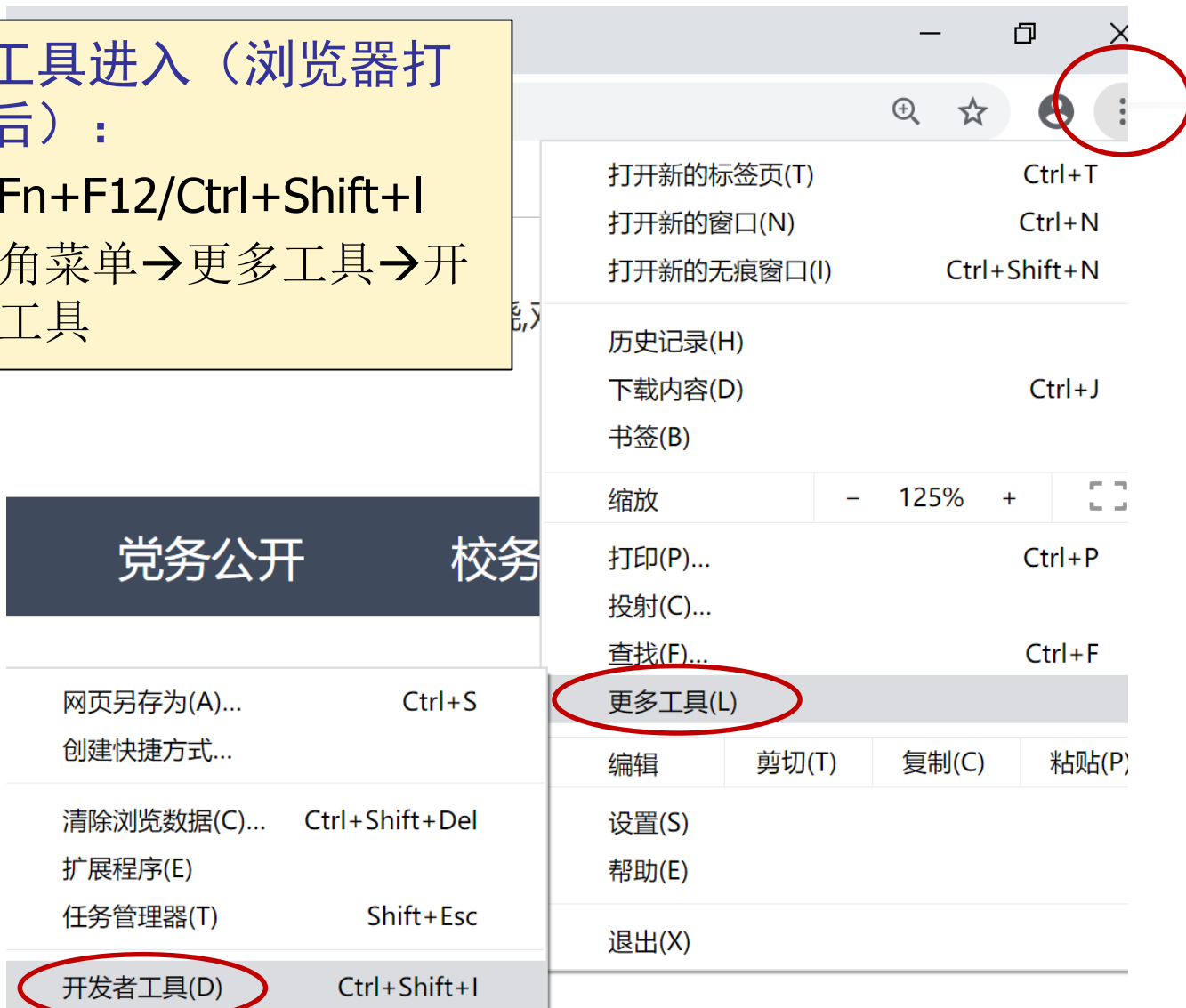
开发者工具

https://blog.csdn.net/weixin_41819731/article/details/80472232 (2020)

<https://www.jianshu.com/p/2641ed6b48a9> (2021)

■ 开发者工具进入（浏览器打开页面后）：

- F12/Fn+F12/Ctrl+Shift+I
- 右上角菜单→更多工具→开发者工具



北邮信息 邮箱 科研网站 教学 娱乐 译 http://fanyi.baidu....

南

学生党

4月20

织学习

参观学

Perform a request

北邮信息 邮箱 科研网站 教学 娱乐 译 http://fanyi.baidu....

Network Performance Memory Application Security Audits

Filter Hide data URLs All XHR JS CSS Img Media Font Doc WS Manifest Other

10 ms 20 ms 30 ms 40 ms

500 ms 1000 ms 1500 ms 2000 ms 2500 ms

Name Headers Preview Response Cookies Timing

t_index.jsp?urltype=tree.TreeTempUrl&wb...

General

Request URL: https://webvpn.bupt.edu.cn/http/7772647670e1e7b0c9ce29b5b/t_index.jsp?urltype=tree.TreeTempUrl&wb...

Request Method: GET

Status Code: 200 OK

Remote Address: 211.68.69.250:443

Referrer Policy: no-referrer-when-downgrade

Response Headers view source

Connection: keep-alive

Content-Encoding: gzip

Content-Language: zh-CN

Content-Type: text/html; charset=UTF-8

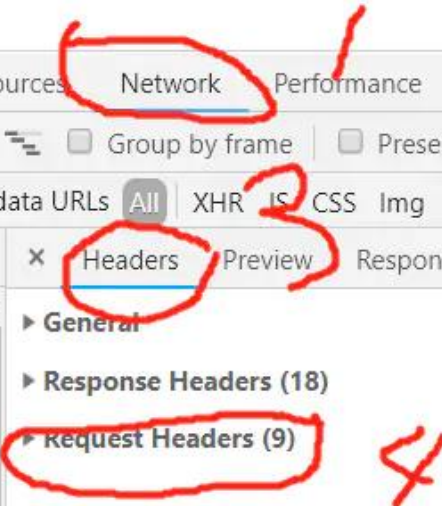
Date: Thu, 22 Apr 2021 09:11:09 GMT

Server: none

Transfer-Encoding: chunked

Vary: Accept-Encoding

X-Frame-Options: SAMEORIGIN



■ 开发者工具查看HTTP协议

× Headers Preview Response Cookies Timing

▼ General

Request URL: https://webvpn.bupt.edu.cn/http/77726476706e69737468656265737421fdee0f9e32207c1e7b0c9ce29b5b/t_index.jsp?urltype=tree.TreeTempUrl&wbtreeid=1545

Request Method: GET

Status Code:  200 OK

Remote Address: 211.68.69.250:443

Referrer Policy: no-referrer-when-downgrade

▼ Response Headers [view source](#)

Connection: keep-alive

Content-Encoding: gzip

Content-Language: zh-CN

Content-Type: text/html; charset=UTF-8

Date: Thu, 22 Apr 2021 09:11:09 GMT

Server: none

Transfer-Encoding: chunked

Vary: Accept-Encoding

X-Frame-Options: SAMEORIGIN

X-Ua-Compatible: IE=edge,chrome=1

▼ Request Headers [view source](#)

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8

Accept-Encoding: gzip, deflate, br

Accept-Language: zh-CN,zh;q=0.9

Cache-Control: max-age=0

Connection: keep-alive

Cookie: wengine_vpn_ticketwebvpn_bupt_edu_cn=315d2c619d65e

Host: webvpn.bupt.edu.cn

Referer: https://webvpn.bupt.edu.cn/http/77726476706e69737421fdee0f9e32207c1e7b0c9ce29b5b/t_index.jsp?urltype=tree.TreeTempUrl&wbtreeid=1545

Sec-Fetch-Mode: navigate

Sec-Fetch-Site: same-origin

Sec-Fetch-User: ?1

Upgrade-Insecure-Requests: 1

User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/90.0.4430.21 Safari/537.36

■ 开发者工具查看HTTP协议

▼ Request Headers

[view source](#)

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8

-exchange;v=b3

Accept-Encoding: gzip, deflate, br

Accept-Language: zh-CN,zh;q=0.9

Cache-Control: max-age=0

Connection: keep-alive

Cookie: wengine_vpn_ticketwebvpn_bupt_edu_cn=315d2c619d65eeaa; show_vpn=0; refresh=1

Host: webvpn.bupt.edu.cn

Referer: https://webvpn.bupt.edu.cn/http/77726476706e69737468656265737421fdee0f9e32207c1e7b0c9ce29b5b/t_index.jsp?url=eeid=1545 HTTP/1.1

Upgrade-Insecure-Requests: 1

User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.18 Safari/537.36

Sec-Fetch-Mode: navigate

Sec-Fetch-Site: same-origin

Sec-Fetch-User: ?1

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8

-exchange;v=b3

Accept-Encoding: gzip, deflate, br

Accept-Language: zh-CN,zh;q=0.9

Cookie: wengine_vpn_ticketwebvpn_bupt_edu_cn=315d2c619d65eeaa; show_vpn=0; refresh=1

▼ Request Headers

[view parsed](#)

GET /http/77726476706e69737468656265737421fdee0f9e32207c1e7b0c9ce29b5b/t_index.jsp?url=eeid=1545 HTTP/1.1

Host: webvpn.bupt.edu.cn

Connection: keep-alive

Cache-Control: max-age=0

Upgrade-Insecure-Requests: 1

User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.18 Safari/537.36

Sec-Fetch-User: ?1

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8

-exchange;v=b3

Accept-Encoding: gzip, deflate, br

Accept-Language: zh-CN,zh;q=0.9

Cookie: wengine_vpn_ticketwebvpn_bupt_edu_cn=315d2c619d65eeaa; show_vpn=0; refresh=1

Referer: https://webvpn.bupt.edu.cn/http/77726476706e69737468656265737421fdee0f9e32207c1e7b0c9ce29b5b/t_index.jsp?url=eeid=1545

Upgrade-Insecure-Requests: 1

Cache-Control: max-age=0

Connection: keep-alive

Host: webvpn.bupt.edu.cn

User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.18 Safari/537.36

■ 开发者工具查看HTTP协议

▼ Response Headers

[view source](#)

Connection: keep-alive

Content-Encoding: gzip

Content-Language: zh-CN

Content-Type: text/html; charset=UTF-8

Date: Thu, 22 Apr 2021 09:11:09 GMT

Server: none

Transfer-Encoding: chunked

Vary: Accept-Encoding

X-Frame-Options: SAMEORIGIN

X-Ua-Compatible: IE=edge,chrome=1

▼ Response Headers

[view parsed](#)

HTTP/1.1 200 OK

Server: none

Date: Thu, 22 Apr 2021 09:11:09 GMT

Content-Type: text/html; charset=UTF-8

Transfer-Encoding: chunked

Connection: keep-alive

Content-Language: zh-CN

Vary: Accept-Encoding

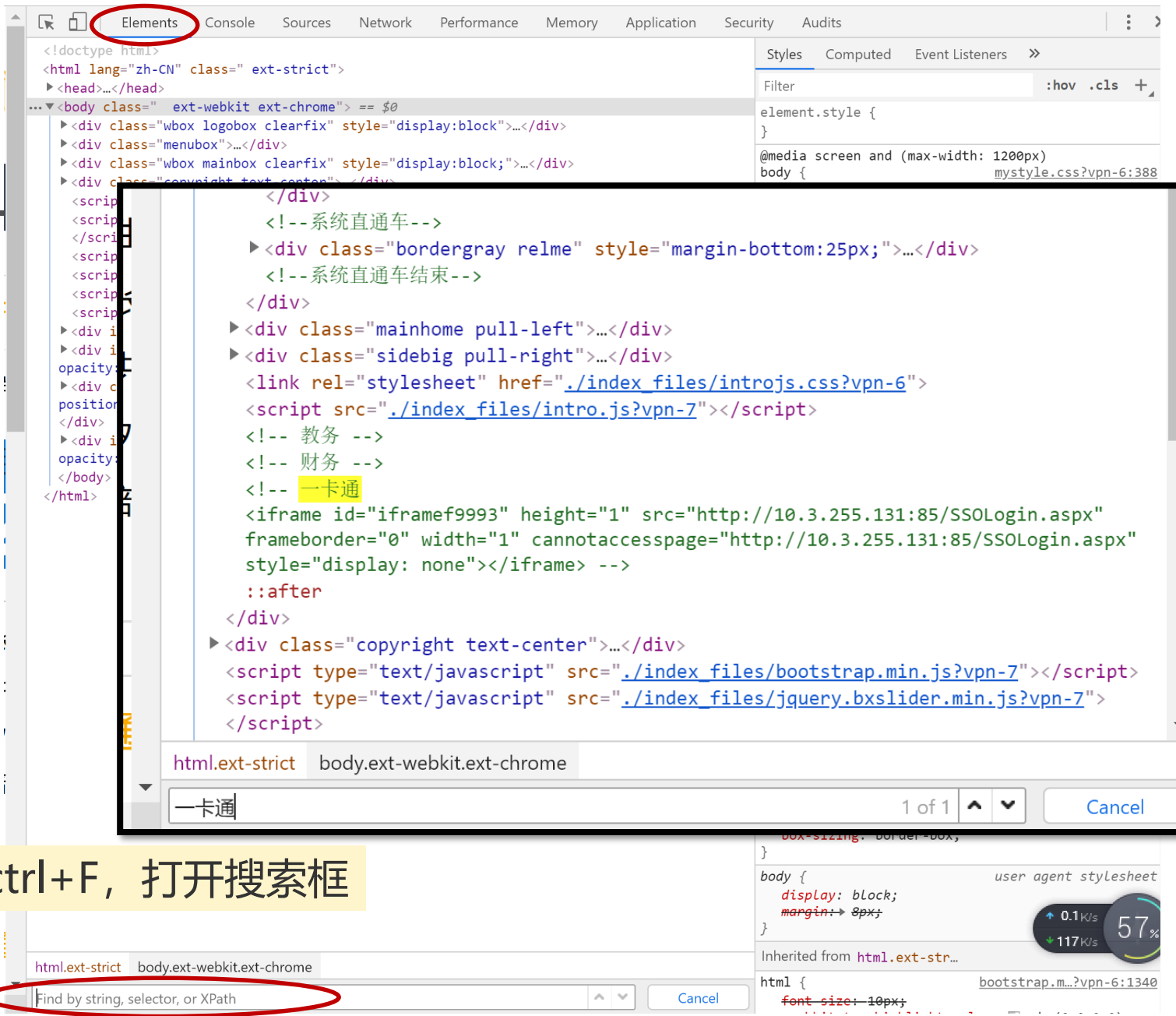
X-Frame-Options: SAMEORIGIN

X-Ua-Compatible: IE=edge,chrome=1

Content-Encoding: gzip

开发者工具 查看、分析 网页源码

<https://jingyan.baidu.com/article/046a7b3e92c3f4f9c37fa948.html>



使用查找工具ctrl+F，打开搜索框

■ 在源代码上滑动鼠标，代码对应的页面部分会被选中加深

教师首页-欢迎访问信息服务门户 × 开发者工具如何查看源码_百度搜 × 如何查看网页源码-百度经验 × 简 如何用Chrome开发者工具查看 × +

webvpn.bupt.edu.cn/http/77726476706e69737468656265737421fdee0f9e32207c1e7b0c9ce29b5b/t_index.jsp?urltype=tree.TreeTempUrl&wbtreei

应用 默认文件夹 网址收藏 我的工作 我的生活 2020会议 百度 北邮信息 邮箱 科研网站 教学 娱乐 译 http://fanyi.baidu....

北京邮电大学
Beijing University of Posts and Telecommunications

服务

门户首页 规章制度 资源中心

img 300 × 182.78

校内新

04月

2

系统直通车

研究生

Elements Console Sources Network Performance Memory Application Se

```
<div class="sidesmall pull-left">
  <div class="relme" style="margin-bottom:25px;">
    <div style=" width:300px; height:180px; margin-bottom:25px;">
      <script language="javascript" src="/http/7772647.../system/resource/js/
      dynclicks.js"></script>
      <a href="http://77726476706e69737468656265737421fdee0f9e32207c1e7b0c9ce29b5b/t_index.jsp?urltype=tree.TreeTempUrl&wbtreei=00d8db9d6562d/" target=_
      blank" onclick="addDynclick(298 × 180 pixels (intrinsic: 300 × 182 pixels))" return vpn_return">
        
      </a>
    </div>
  </div>
  <!-- 系统直通车 -->
  <div class="bordergray relme" style="margin-bottom:25px;">...</div>
  <!-- 系统直通车结束 -->
</div>
<div class="mainhome pull-left">...</div>
<div class="sidebig pull-right">...</div>
<link rel="stylesheet" href="/index_files/introjs.css?vpn-6">
<script src="/index_files/intro.js?vpn-7"></script>
<!-- 教务 -->
<!-- 财务 -->
<!-- 一卡通 -->
<iframe id="iframef9993" height="1" src="http://10.3.255.131:85/SSOLogin.aspx"
frameborder="0" width="1" cannotaccesspage="http://10.3.255.131:85/SSOLogin.aspx"
style="display: none"></iframe> -->
::after
</div>
<div class="copyright text-center">...</div>
```

- 开发者工具查找代码（2021年加）

http://www.360doc.com/content/19/0701/12/164769_846028343.shtml（在画面元素上点击右键-->选择检查）



The screenshot shows a web browser displaying a real estate advertisement for '天健云山府' (Tianjian Yunshan). The advertisement features a large image of a modern building complex and a smaller inset image of a cityscape at night. The text on the page includes '天健云山府' and '66000元/平 白云/白云大道'. A right-click context menu is open over the main image, showing options such as '返回(B)', '前进(F)', '重新加载(R)', '另存为(A)...', '打印(P)...', '投射(C)...', '翻成中文 (简体) (T)', '查看网页源代码(V)', and '检查(N)'. The '检查(N)' option is highlighted. The developer tools are open, showing the 'Network' tab and the 'Performance' tab. The 'Performance' tab is selected, and the 'Timeline' view is visible. The 'Timeline' view shows a list of events, including 'load', 'DOMContentLoaded', and 'loadComplete'. The 'load' event is selected, and the 'Timeline' view shows a list of events, including 'load', 'DOMContentLoaded', and 'loadComplete'. The 'load' event is selected, and the 'Timeline' view shows a list of events, including 'load', 'DOMContentLoaded', and 'loadComplete'. The 'load' event is selected, and the 'Timeline' view shows a list of events, including 'load', 'DOMContentLoaded', and 'loadComplete'.

天健云山府

66000元/平
白云/白云大道

返回(B) Alt+向左箭头
前进(F) Alt+向右箭头
重新加载(R) Ctrl+R
另存为(A)... Ctrl+S
打印(P)... Ctrl+P
投射(C)...
翻成中文 (简体) (T)
查看网页源代码(V) Ctrl+U
检查(N) Ctrl+Shift+I

Network Performance Memory Application

```
<a href="/xinfang/lp-11098/?home_xf_post1" target="_blank" class="img" ...  
 == $0
```

开发者工具查找代码（2021年加）

代码查看工具_F12 - 开发者工具详解（2020-11-27）

https://blog.csdn.net/weixin_40006133/article/details/111237724

查看元素的代码

点击左上角的箭头图标（或按快捷键 Ctrl+Shift+C）进入选择元素模式，从页面中选择需要查看的（Elements）一栏中定位到该元素源代码的具体位置。

查看元素的属性：定位到元素的源代码之后，可以从源代码中读出改元素的属性。



2、选择元素

