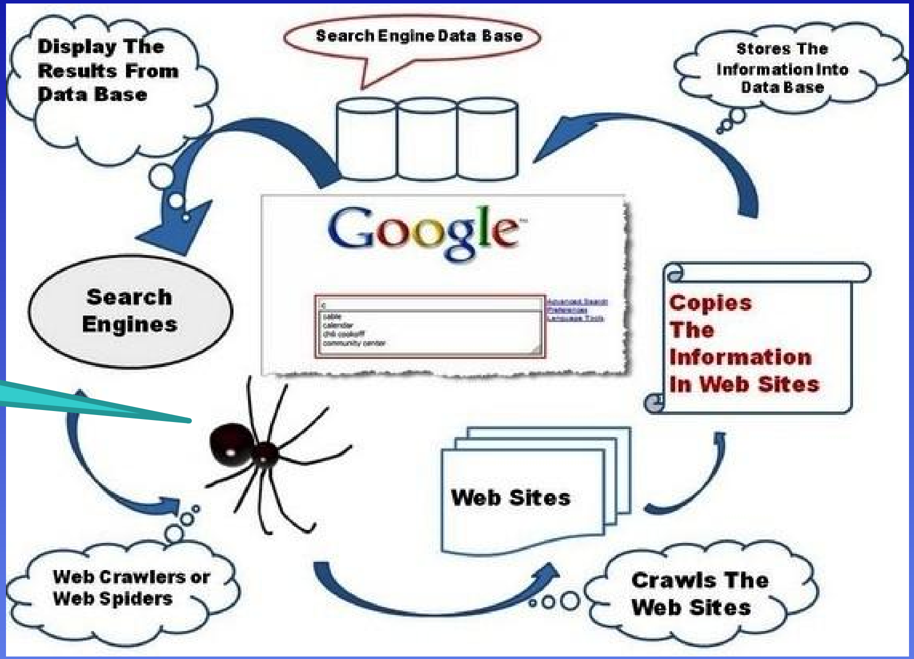
爬虫简介

大纲

- 爬虫是什么
- 基础知识
- 爬虫分类
- 爬虫架构
- 反抓取
- 网页解析
- 工具介绍

爬虫是什么

- 名字
  - Crawler、Spider、Bot
  - Web Scraping(verb)
- 定义(wiki)
  - A Web crawler, sometimes called a spider, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing

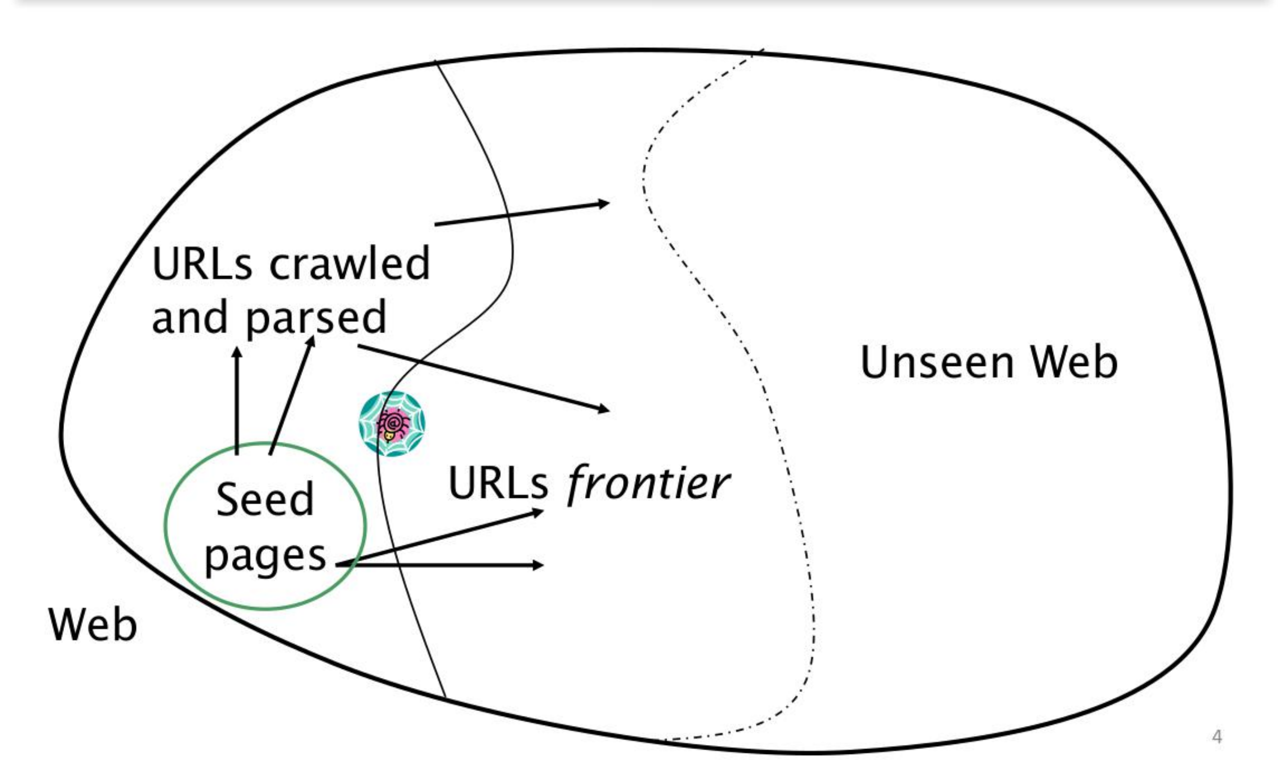

Display The Results From Data Base
Search Engine Data Base
Stores The Information Into Data Base
Google
Search Engines
Copies The Information In Web Sites
Web Sites
Web Crawlers or Web Spiders
Crawls The Web Sites

Spider

爬虫的抓取过程

- 从“种子”链接开始
  - 把它们加到待抓取“队列”

- While True:
  - 从队列中取任务
  - 抓取内容并且抽取链接
  - 把链接去重后加入队列

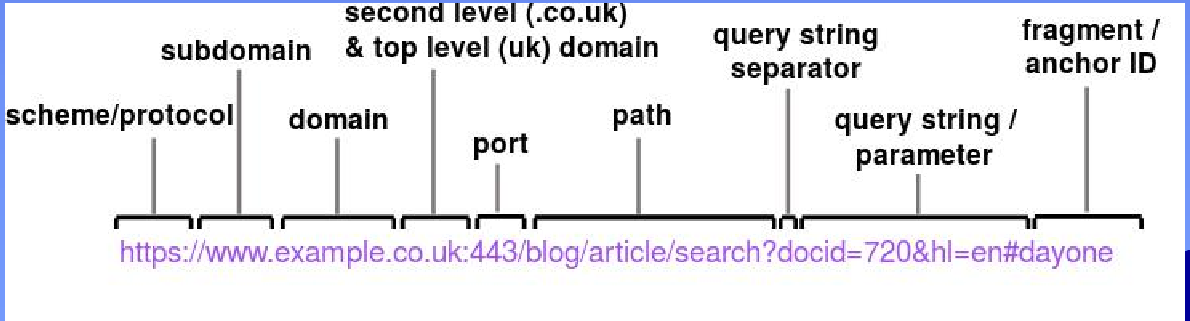


基础知识

- URL
- HTML/CSS/JS
- HTTP
- DNS

URL

- Uniform Resource Locator
  - colloquially termed a web address, is a reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it
- scheme:[//[user:password@]host[:port]][/]path[?query][#fragment]



https://www.example.co.uk:443/blog/article/search?docid=720&hl=en#dayone

URL

- Encoding/Decoding
  - https://www.baidu.com/s?ie=utf-8&f=8&rsv_bp=0&rsv_idx=1&tn=baidu&wd=%E5%A4%A9%E6%B0%94

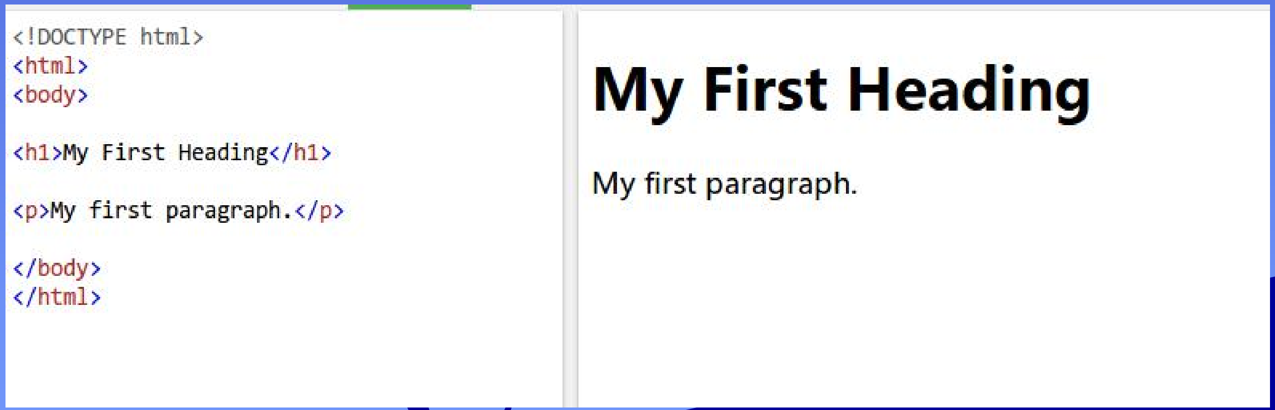
- Fragment
  - http://www.tagindex.net/html/link/anchor_example2.html#a003

URL Normalization

- HTTP://www.Example.com/ → http://www.example.com/
- http://www.example.com/a%c2%b1b → http://www.example.com/a%C2%B1b
- http://www.example.com:80/bar.html → http://www.example.com/bar.html
- Google Bot remove fragment
- https://en.wikipedia.org/wiki/URL_normalization

HTML

- Hypertext Markup Language (HTML) is the standard markup language for creating web pages and web applications. With Cascading Style Sheets (CSS) and JavaScript, it forms a triad of cornerstone technologies for the World Wide Web



```
<!DOCTYPE html>
<html>
<body>

<h1>My First Heading</h1>

<p>My first paragraph.</p>

</body>
</html>
```

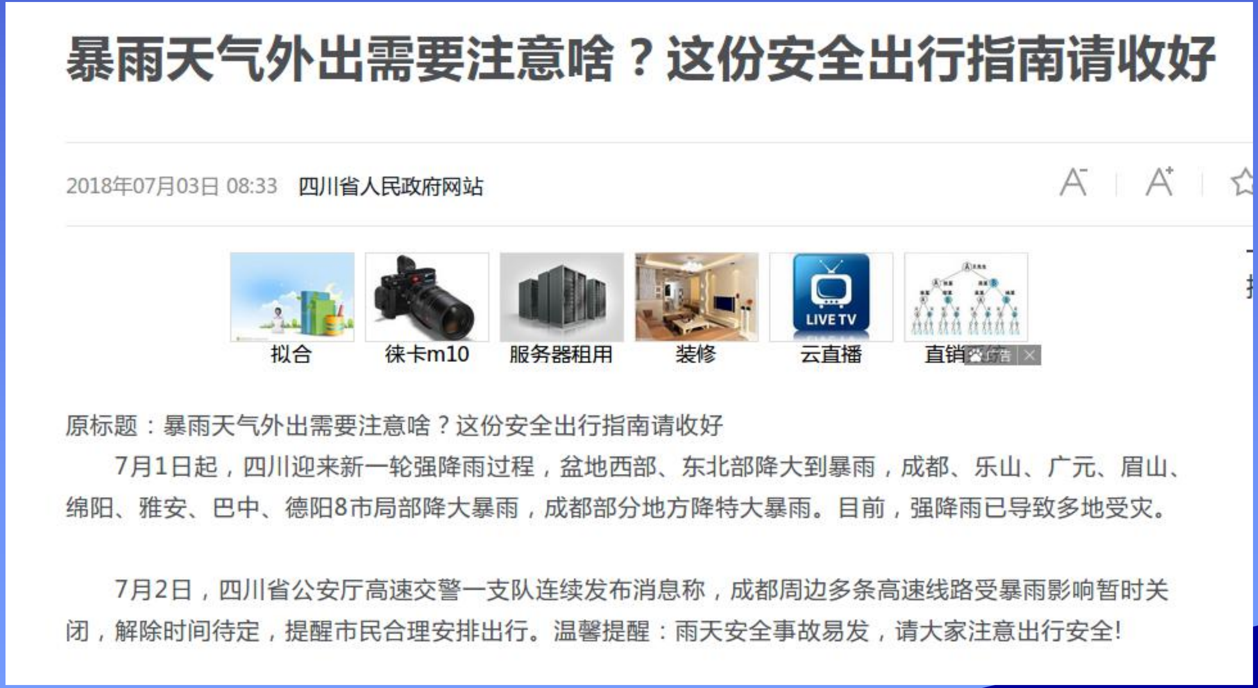
My First Heading

My first paragraph.

- https://www.w3schools.com/html/tryit.asp?filename=tryhtml_basic_document

HTML

- http://news.sina.com.cn/c/2018-07-03/doc-ihevauxi3422393.shtml



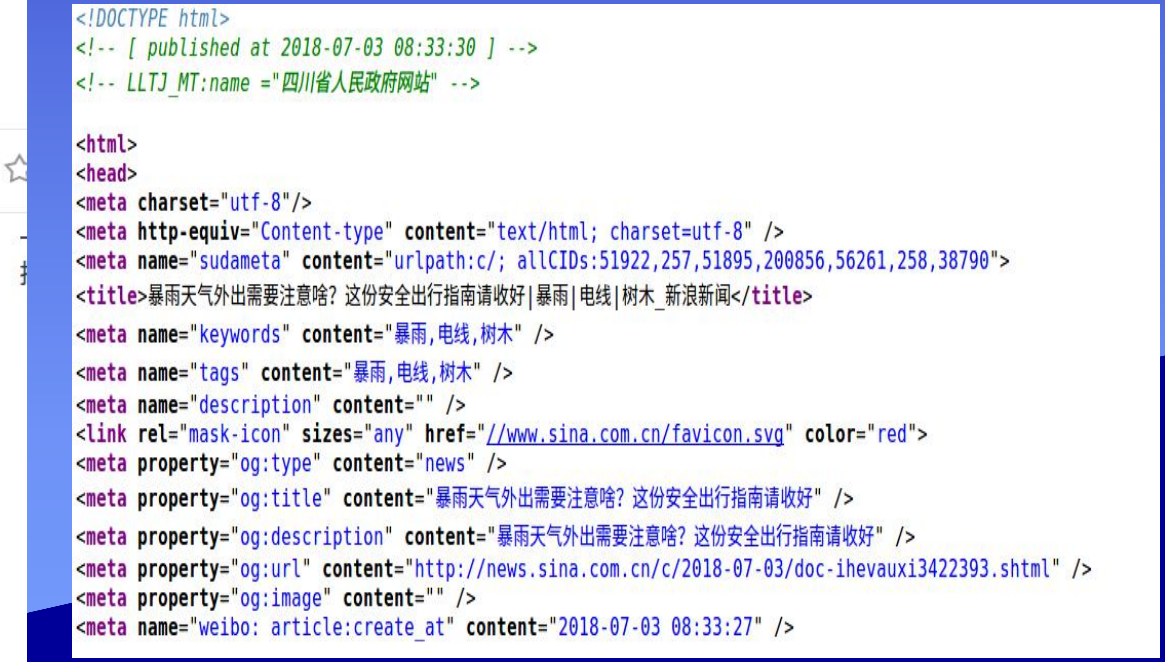
暴雨天气外出需要注意啥？这份安全出行指南请收好

2018年07月03日 08:33 四川省人民政府网站

原标题：暴雨天气外出需要注意啥？这份安全出行指南请收好

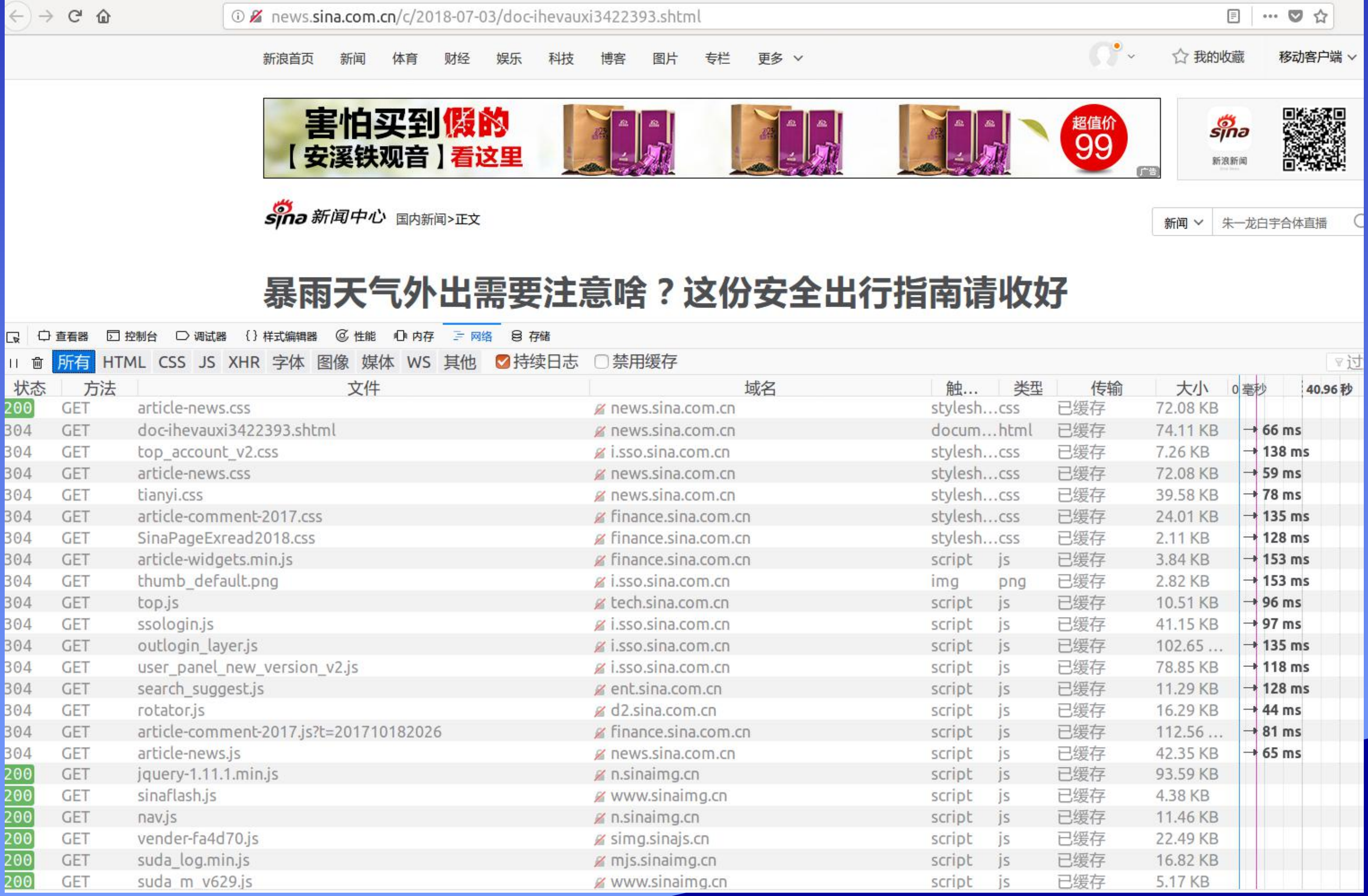
7月1日起，四川迎来新一轮强降雨过程，盆地西部、东北部降大到暴雨，成都、乐山、广元、眉山、绵阳、雅安、巴中、德阳8市局部降大暴雨，成都部分地方降特大暴雨。目前，强降雨已导致多地受灾。

7月2日，四川省公安厅高速交警一支队连续发布消息称，成都周边多条高速线路受暴雨影响暂时关闭，解除时间待定，提醒市民合理安排出行。温馨提醒：雨天安全事故易发，请大家注意出行安全!



```html
<!DOCTYPE html>
<!-- [ published at 2018-07-03 08:33:30 ] -->
<!-- LLTJ_MT:name ="四川省人民政府网站" -->

<html>
<head>
<meta charset="utf-8"/>
<meta http-equiv="Content-type" content="text/html; charset=utf-8" />
<meta name="sudameta" content="urlpath:c/; allCIDs:51922,257,51895,200856,56261,258,38790">
<title>暴雨天气外出需要注意啥? 这份安全出行指南请收好|暴雨|电线|树木_新浪新闻</title>
<meta name="keywords" content="暴雨,电线,树木" />
<meta name="tags" content="暴雨,电线,树木" />
<meta name="description" content="" />
<link rel="mask-icon" sizes="any" href="//www.sina.com.cn/favicon.svg" color="red">
<meta property="og:type" content="news" />
<meta property="og:title" content="暴雨天气外出需要注意啥? 这份安全出行指南请收好" />
<meta property="og:description" content="暴雨天气外出需要注意啥? 这份安全出行指南请收好" />
<meta property="og:url" content="http://news.sina.com.cn/c/2018-07-03/doc-ihevauxi3422393.shtml" />
<meta property="og:image" content="" />
<meta name="weibo: article:create_at" content="2018-07-03 08:33:27" />
```

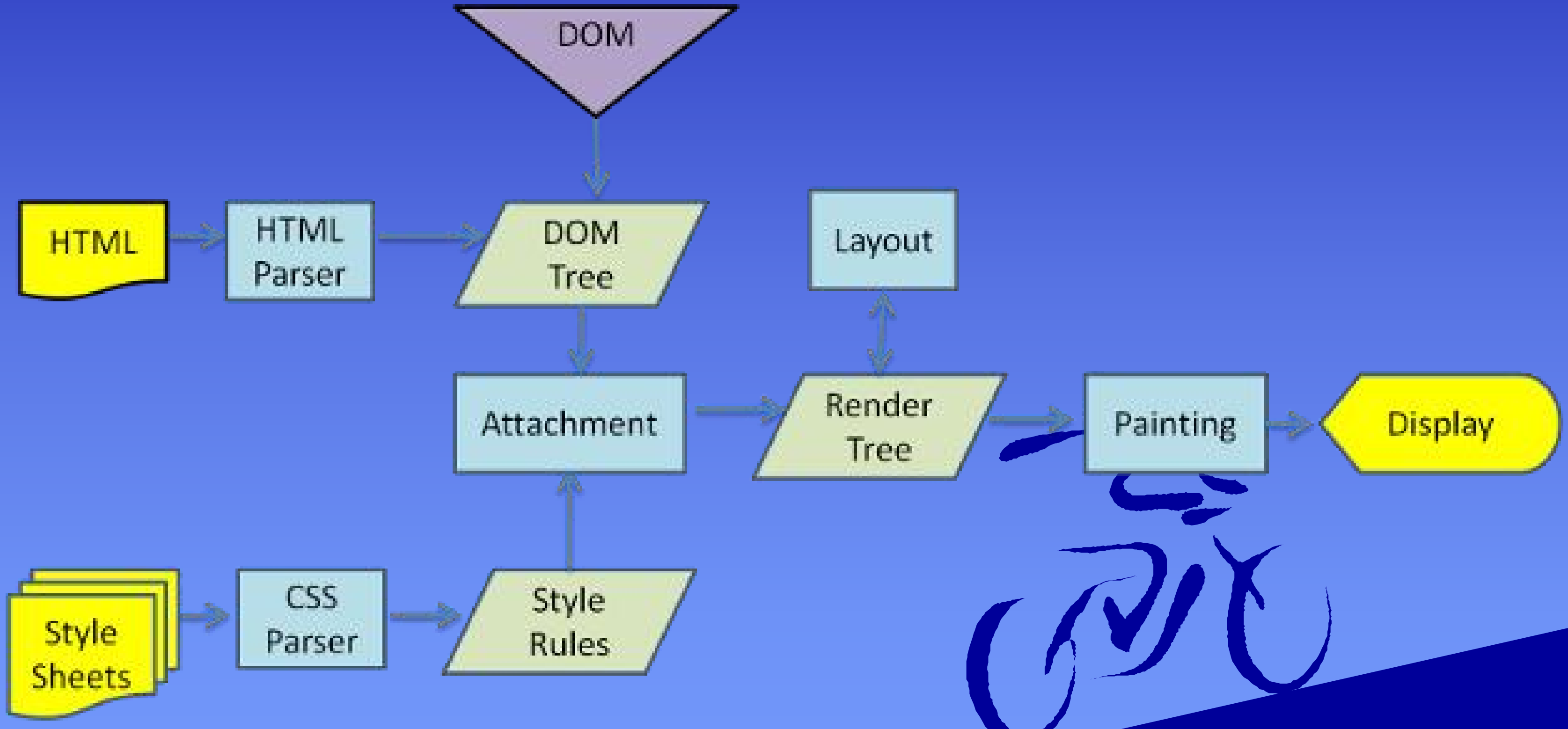
不只是一个HTML!



☑ 持续日志 □ 禁用缓存

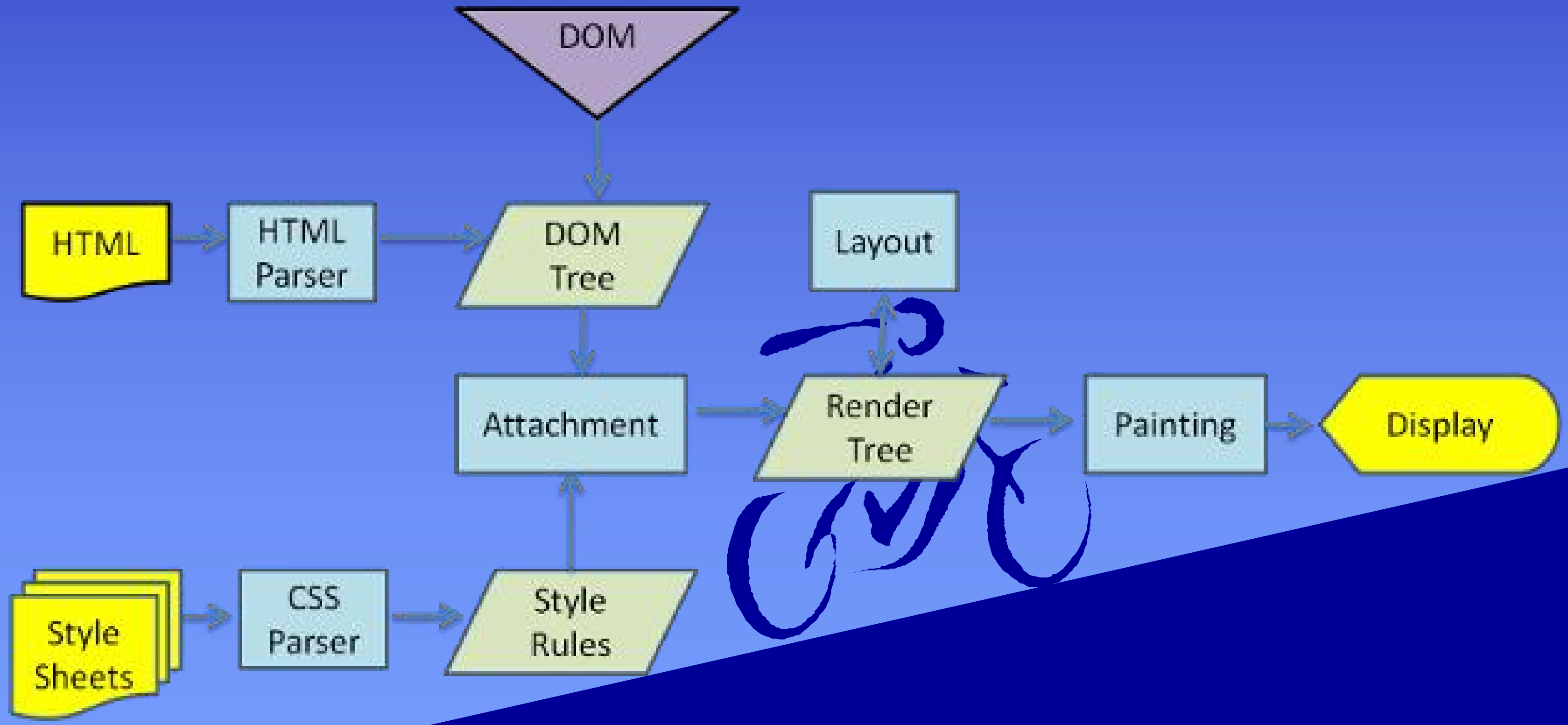
状态	方法	文件	域名	触…	类型	传输	大小	0 毫秒 40.96 秒
200	GET	article-news.css	news.sina.com.cn	stylesh…	css	已缓存	72.08 KB	
304	GET	doc-ihevauxi3422393.shtml	news.sina.com.cn	docum…	html	已缓存	74.11 KB	→ 66 ms
304	GET	top_account_v2.css	i.sso.sina.com.cn	stylesh…	css	已缓存	7.26 KB	→ 138 ms
304	GET	article-news.css	news.sina.com.cn	stylesh…	css	已缓存	72.08 KB	→ 59 ms
304	GET	tianyi.css	news.sina.com.cn	stylesh…	css	已缓存	39.58 KB	→ 78 ms
304	GET	article-comment-2017.css	finance.sina.com.cn	stylesh…	css	已缓存	24.01 KB	→ 135 ms
304	GET	SinaPageExread2018.css	finance.sina.com.cn	stylesh…	css	已缓存	2.11 KB	→ 128 ms
304	GET	article-widgets.min.js	finance.sina.com.cn	script	js	已缓存	3.84 KB	→ 153 ms
304	GET	thumb_default.png	i.sso.sina.com.cn	img	png	已缓存	2.82 KB	→ 153 ms
304	GET	top.js	tech.sina.com.cn	script	js	已缓存	10.51 KB	→ 96 ms
304	GET	ssologin.js	i.sso.sina.com.cn	script	js	已缓存	41.15 KB	→ 97 ms
304	GET	outlogin_layer.js	i.sso.sina.com.cn	script	js	已缓存	102.65 …	→ 135 ms
304	GET	user_panel_new_version_v2.js	i.sso.sina.com.cn	script	js	已缓存	78.85 KB	→ 118 ms
304	GET	search_suggest.js	ent.sina.com.cn	script	js	已缓存	11.29 KB	→ 128 ms
304	GET	rotator.js	d2.sina.com.cn	script	js	已缓存	16.29 KB	→ 44 ms
304	GET	article-comment-2017.js?t=201710182026	finance.sina.com.cn	script	js	已缓存	112.56 …	→ 81 ms
304	GET	article-news.js	news.sina.com.cn	script	js	已缓存	42.35 KB	→ 65 ms
200	GET	jquery-1.11.1.min.js	n.sinaimg.cn	script	js	已缓存	93.59 KB	
200	GET	sinaflash.js	www.sinaimg.cn	script	js	已缓存	4.38 KB	
200	GET	nav.js	n.sinaimg.cn	script	js	已缓存	11.46 KB	
200	GET	vender-fa4d70.js	simg.sinajs.cn	script	js	已缓存	22.49 KB	
200	GET	suda_log.min.js	mjs.sinaimg.cn	script	js	已缓存	16.82 KB	
200	GET	suda_m_v629.js	www.sinaimg.cn	script	js	已缓存	5.17 KB	

浏览器渲染过程


DOM
HTML
HTML Parser
DOM Tree
Layout
Attachment
Render Tree
Painting
Display
Style Sheets
CSS Parser
Style Rules

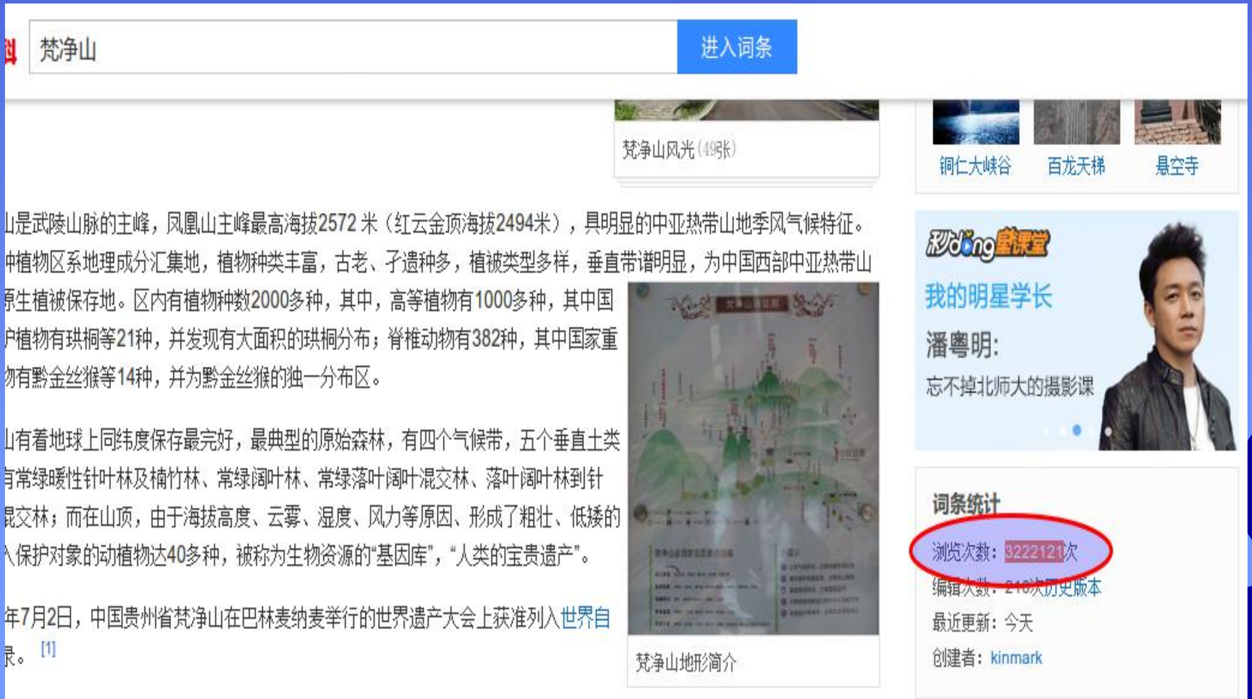
浏览器渲染过程

- https://www.html5rocks.com/en/tutorials/internals/howbrowserswork/



Ajax

- Ajax ("Asynchronous JavaScript And XML") is a set of Web development techniques using many Web technologies on the client side to create asynchronous Web applications



梵净山 进入词条

梵净山风光(49张)

铜仁大峡谷 百龙天梯 悬空寺

山是武陵山脉的主峰，凤凰山主峰最高海拔2572 米（红云金顶海拔2494米），具明显的中亚热带山地季风气候特征。中植物区系地理成分汇集地，植物种类丰富，古老、孑遗种多，植被类型多样，垂直带谱明显，为中国西部中亚热带山原生植被保存地。区内有植物种数2000多种，其中，高等植物有1000多种，其中国护植物有珙桐等21种，并发现有大面积的珙桐分布；脊椎动物有382种，其中国家重物有黔金丝猴等14种，并为黔金丝猴的独一分布区。

山有着地球上同纬度保存最完好，最典型的原始森林，有四个气候带，五个垂直土类有常绿暖性针叶林及楠竹林、常绿阔叶林、常绿落叶阔叶混交林、落叶阔叶林到针混交林；而在山顶，由于海拔高度、云雾、湿度、风力等原因、形成了粗壮、低矮的入保护对象的动植物达40多种，被称为生物资源的"基因库"，"人类的宝贵遗产"。

年7月2日，中国贵州省梵净山在巴林麦纳麦举行的世界遗产大会上获准列入世界自录。[1]

梵净山地形简介

秒dong星课堂

我的明星学长

潘粤明:

忘不掉北师大的摄影课

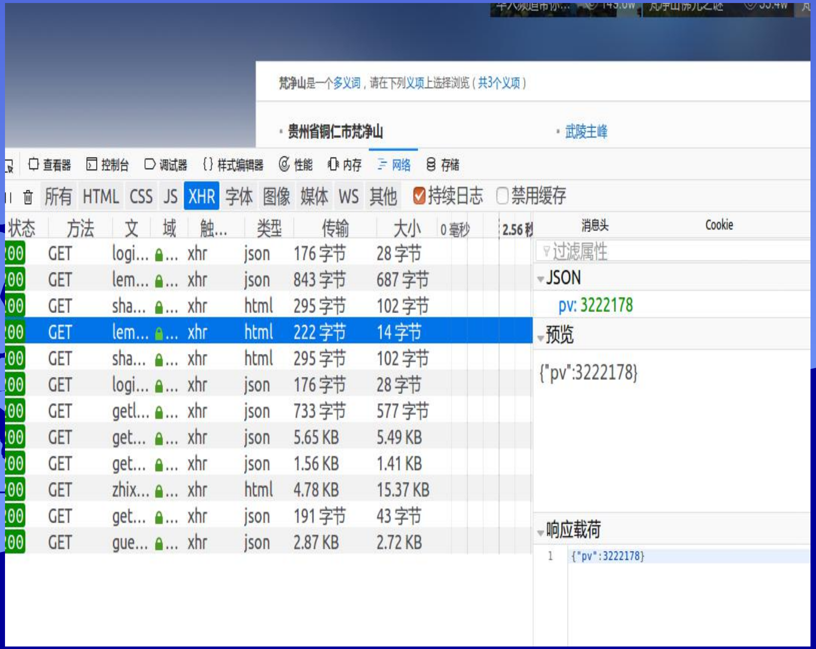
词条统计

浏览次数：3222121次

编辑次数：210次历史版本

最近更新：今天

创建者：kinmark



梵净山是一个多义词，请在下列义项上选择浏览（共3个义项）

· 贵州省铜仁市梵净山 · 武陵主峰

查看器 控制台 调试器 样式编辑器 性能 内存 网络 存储
所有 HTML CSS JS XHR 字体 图像 媒体 WS 其他 ☑持续日志 □禁用缓存
0 毫秒 2.56 秒
消息头 Cookie
过滤属性

状态	方法	文	域	触...	类型	传输	大小
00	GET	logi...	🔒...	xhr	json	176 字节	28 字节
00	GET	lem...	🔒...	xhr	json	843 字节	687 字节
00	GET	sha...	🔒...	xhr	html	295 字节	102 字节
00	GET	lem...	🔒...	xhr	html	222 字节	14 字节
00	GET	sha...	🔒...	xhr	html	295 字节	102 字节
00	GET	logi...	🔒...	xhr	json	176 字节	28 字节
00	GET	getl...	🔒...	xhr	json	733 字节	577 字节
00	GET	get...	🔒...	xhr	json	5.65 KB	5.49 KB
00	GET	get...	🔒...	xhr	json	1.56 KB	1.41 KB
00	GET	zhix...	🔒...	xhr	html	4.78 KB	15.37 KB
00	GET	get...	🔒...	xhr	json	191 字节	43 字节
00	GET	gue...	🔒...	xhr	json	2.87 KB	2.72 KB

JSON

pv: 3222178

预览

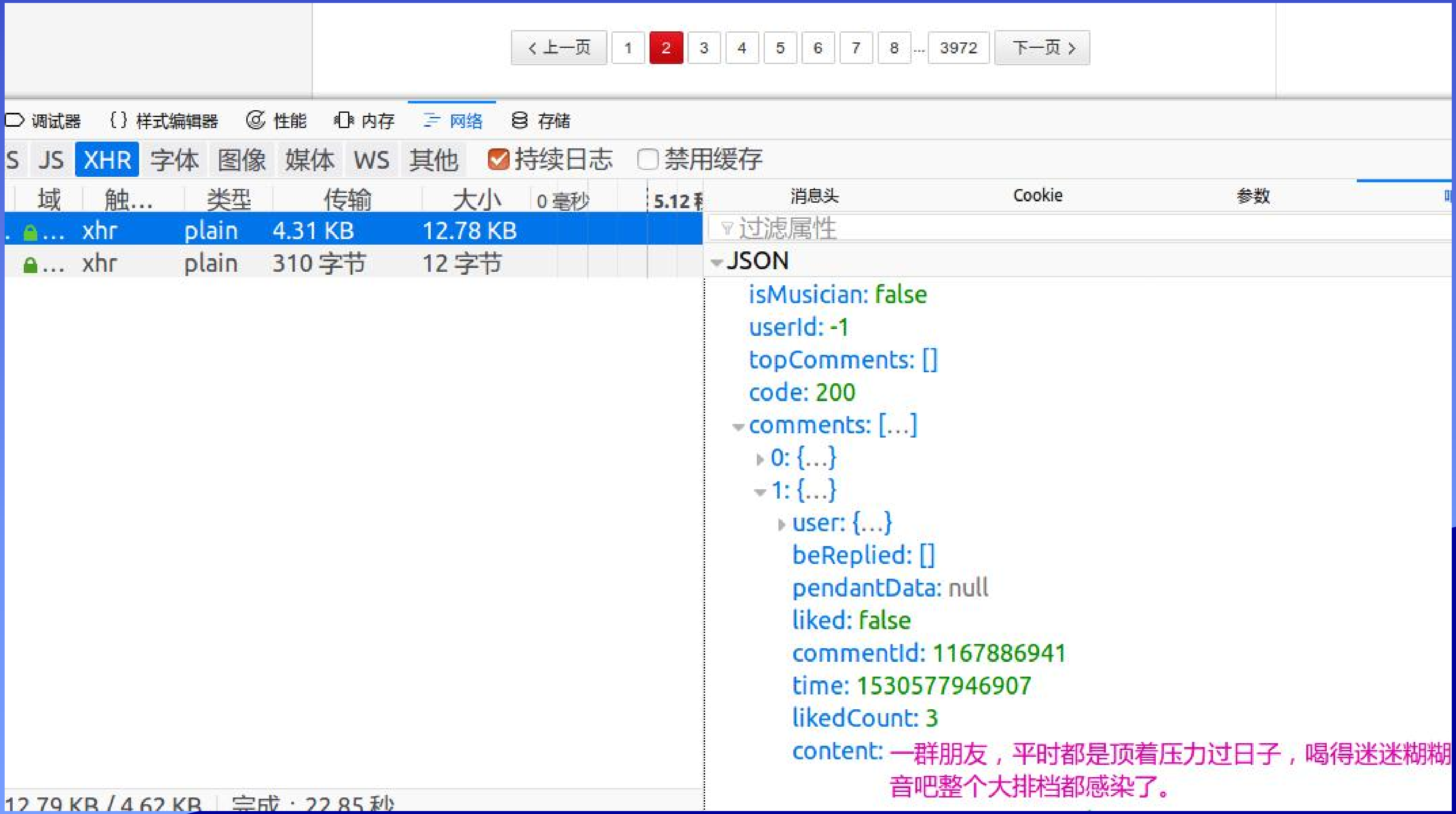
{"pv":3222178}

响应载荷

```
{"pv":3222178}
```

Ajax

- https://music.163.com/#/song?id=168091



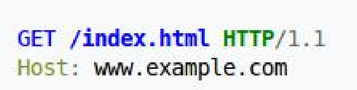
域	触…	类型	传输	大小	0 毫秒	5.12 秒
🔒…	xhr	plain	4.31 KB	12.78 KB		
🔒…	xhr	plain	310 字节	12 字节		

12.79 KB / 4.62 KB 完成：22.85 秒

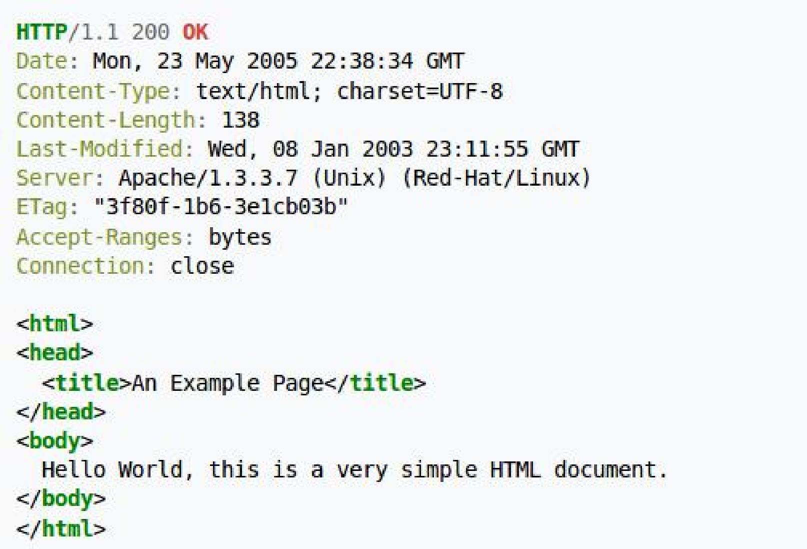
```
JSON
  isMusician: false
  userId: -1
  topComments: []
  code: 200
  comments: […]
    0: {…}
    1: {…}
      user: {…}
      beReplied: []
      pendantData: null
      liked: false
      commentId: 1167886941
      time: 1530577946907
      likedCount: 3
      content: 一群朋友，平时都是顶着压力过日子，喝得迷迷糊糊
               音吧整个大排档都感染了。
```

HTTP

- Hypertext Transfer Protocol
  - The Hypertext Transfer Protocol (HTTP) is an application protocol for distributed, collaborative, and hypermedia information systems.[1] HTTP is the foundation of data communication for the World Wide Web.



```
GET /index.html HTTP/1.1
Host: www.example.com
```

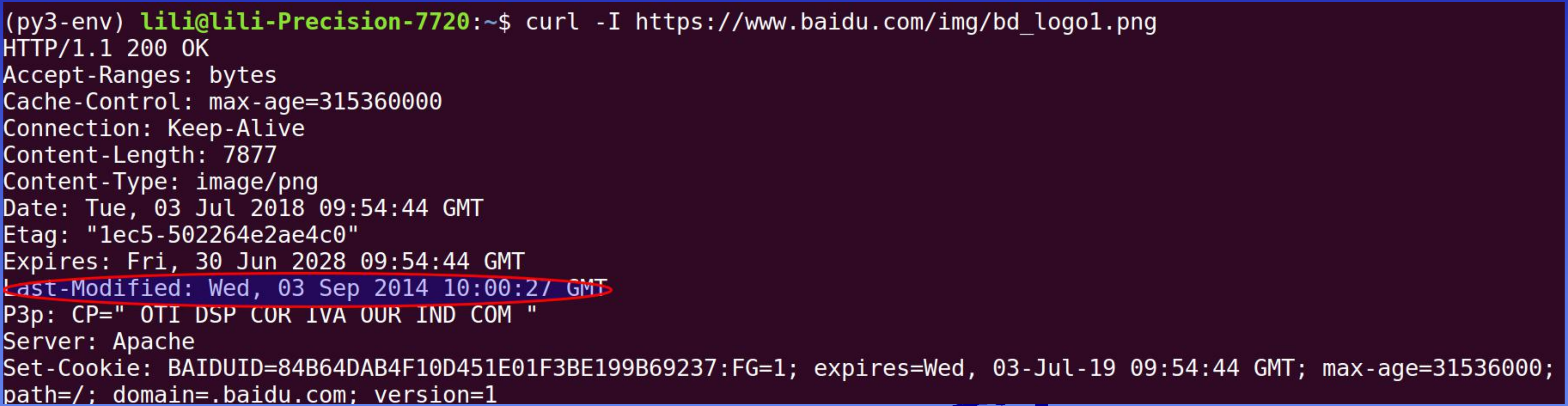


```
HTTP/1.1 200 OK
Date: Mon, 23 May 2005 22:38:34 GMT
Content-Type: text/html; charset=UTF-8
Content-Length: 138
Last-Modified: Wed, 08 Jan 2003 23:11:55 GMT
Server: Apache/1.3.3.7 (Unix) (Red-Hat/Linux)
ETag: "3f80f-1b6-3e1cb03b"
Accept-Ranges: bytes
Connection: close

<html>
<head>
  <title>An Example Page</title>
</head>
<body>
  Hello World, this is a very simple HTML document.
</body>
</html>
```

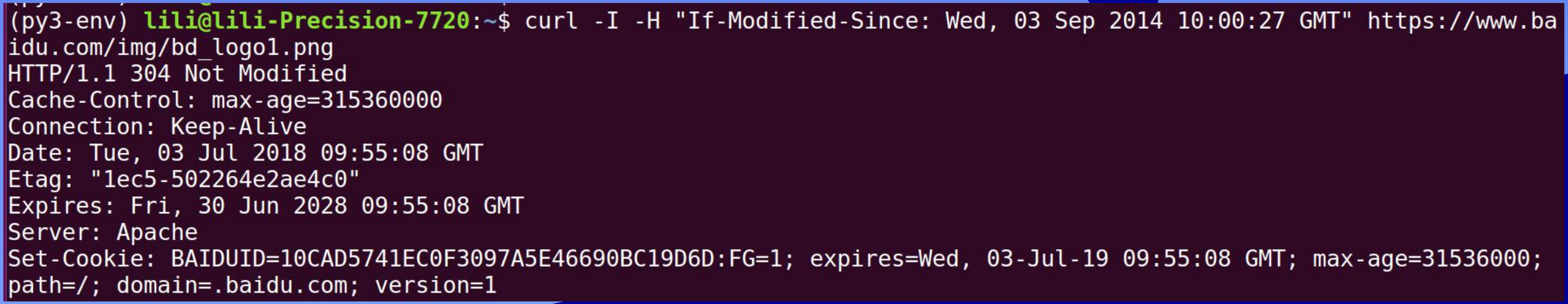
HTTP Headers

- 请求
  - Referer
  - Cookie
  - User-Agent
  - Accept-Encoding
- 响应
  - Cache-Control/Expires
  - Last-Modified/If-Modified-Since
  - ETag/If-None-Match

HTTP Header Example



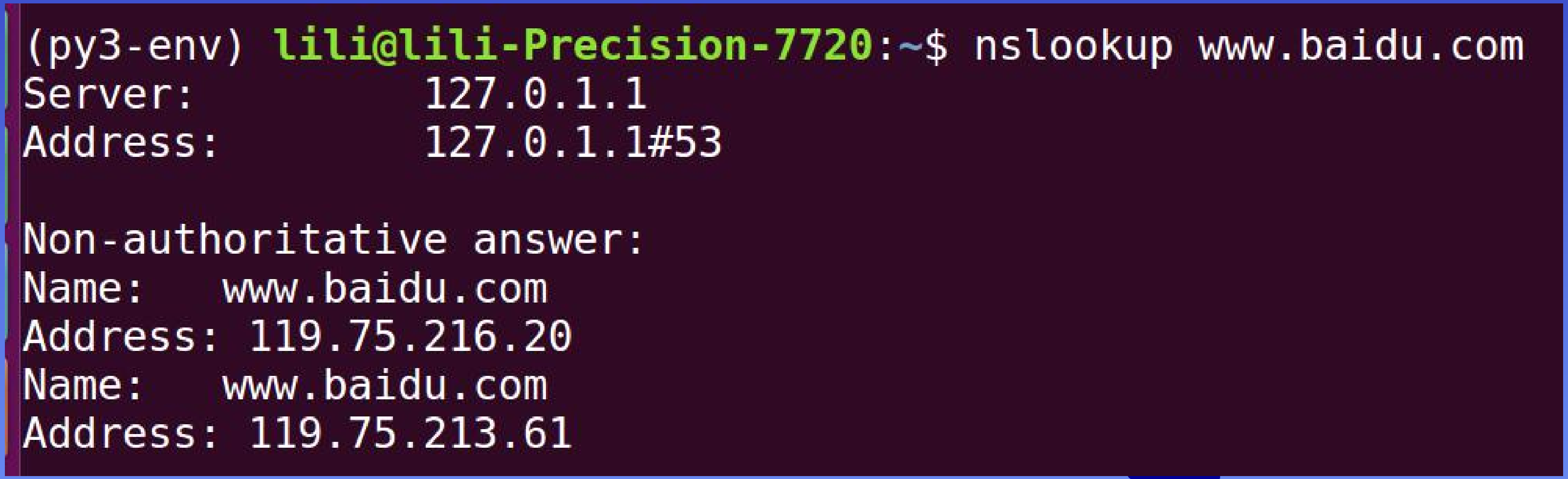
```
(py3-env) lili@lili-Precision-7720:~$ curl -I https://www.baidu.com/img/bd_logo1.png
HTTP/1.1 200 OK
Accept-Ranges: bytes
Cache-Control: max-age=315360000
Connection: Keep-Alive
Content-Length: 7877
Content-Type: image/png
Date: Tue, 03 Jul 2018 09:54:44 GMT
Etag: "1ec5-502264e2ae4c0"
Expires: Fri, 30 Jun 2028 09:54:44 GMT
Last-Modified: Wed, 03 Sep 2014 10:00:27 GMT
P3p: CP=" OTI DSP COR IVA OUR IND COM "
Server: Apache
Set-Cookie: BAIDUID=84B64DAB4F10D451E01F3BE199B69237:FG=1; expires=Wed, 03-Jul-19 09:54:44 GMT; max-age=31536000;
path=/; domain=.baidu.com; version=1
```



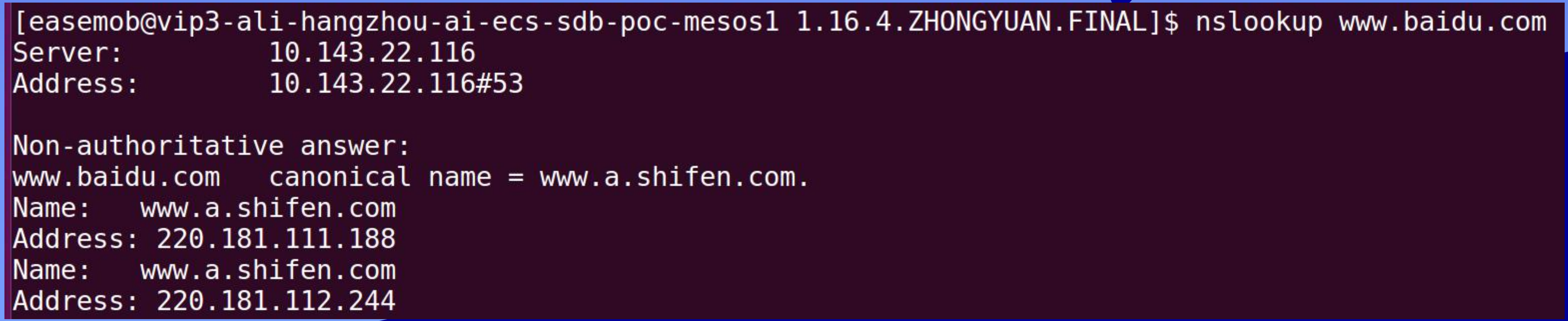
```
(py3-env) lili@lili-Precision-7720:~$ curl -I -H "If-Modified-Since: Wed, 03 Sep 2014 10:00:27 GMT" https://www.ba
idu.com/img/bd_logo1.png
HTTP/1.1 304 Not Modified
Cache-Control: max-age=315360000
Connection: Keep-Alive
Date: Tue, 03 Jul 2018 09:55:08 GMT
Etag: "1ec5-502264e2ae4c0"
Expires: Fri, 30 Jun 2028 09:55:08 GMT
Server: Apache
Set-Cookie: BAIDUID=10CAD5741EC0F3097A5E46690BC19D6D:FG=1; expires=Wed, 03-Jul-19 09:55:08 GMT; max-age=31536000;
path=/; domain=.baidu.com; version=1
```

DNS

- The Domain Name System (DNS) is a hierarchical decentralized naming system for computers, services, or other resources connected to the Internet or a private network.



```
(py3-env) lili@lili-Precision-7720:~$ nslookup www.baidu.com
Server:         127.0.1.1
Address:        127.0.1.1#53

Non-authoritative answer:
Name:   www.baidu.com
Address: 119.75.216.20
Name:   www.baidu.com
Address: 119.75.213.61
```



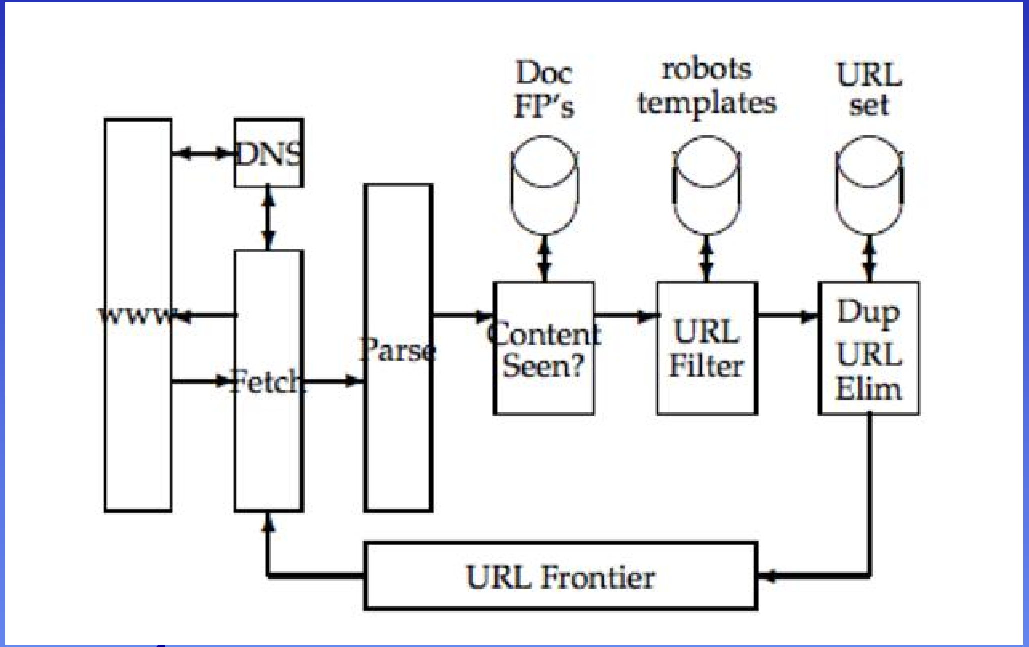
```
[easemob@vip3-ali-hangzhou-ai-ecs-sdb-poc-mesos1 1.16.4.ZHONGYUAN.FINAL]$ nslookup www.baidu.com
Server:         10.143.22.116
Address:        10.143.22.116#53

Non-authoritative answer:
www.baidu.com   canonical name = www.a.shifen.com.
Name:   www.a.shifen.com
Address: 220.181.111.188
Name:   www.a.shifen.com
Address: 220.181.112.244
```

爬虫分类

- 通用爬虫
  - 通用搜索引擎
- 垂直(Vertical)爬虫
  - 垂直搜索引擎
- 定向爬虫
  - Web Scraping

通用爬虫架构

- URL frontier
- DNS resolution
- Fetching module
- Parsing module
- Duplicate elimination



通用爬虫的挑战

- 海量数据
  - Google索引了上千亿网页
- 实时
  - 新的文章能在几分钟搜到
- 深度网络
  - 社交媒体、自媒体
- 反作弊
  - 爬虫陷阱

定向爬虫

反爬虫

- robots.txt https://www.baidu.com/robots.txt
- 封ip/验证码
- headers检验
- 登录
- js "加密"

网页解析

- 抽取网页的重要数据
  - 标题
  - 正文
  - 发表时间
- 方法
  - 手写模板/规则
  - 通用方法

XPath

- https://www.w3schools.com/xml/xpath_intro.asp

- https://doc.scrapy.org/en/xpath-tutorial/topics/xpath-tutorial.html

工具介绍

- Firebug
- Tcpflow
- Xpath插件

练习

- 在终端打印百度首页的DOM树
  - 递归遍历DOM树
- 抓取一个网站
  - 使用介绍的框架
- 根据关键词抓取搜狗微信文章
  - Referer、速度控制
- 抓取https://music.163.com/#/song?id=168091 的所有评论
  - url前端加密

练习

- 通用的标题和正文抽取器
  - anchor text ratio
- 网页截图，超过一屏的截屏，某个元素(验证码)的截屏
  - AShot
- 抓取qunar明天北京到上海所有航班的票价
  - 前端视觉混淆