清華大學
Tsinghua University

小伙伴计划暑期学习营——零基础Python入门

第六讲：网络爬虫入门

张智帅 电子系

清华大学学生学业与发展指导中心

2019-2020学年夏季学期

第六讲：网络爬虫入门

- 爬虫基础
  - 宏观认识
  - 网页的组成
  - 正则表达式
- 常规爬虫
- 爬虫的法律问题
- Ajax爬虫
- 暑期学习营总结

宏观认识

- 什么是网络爬虫?
  - 按照一定的规则，自动从网络中提取信息的程序

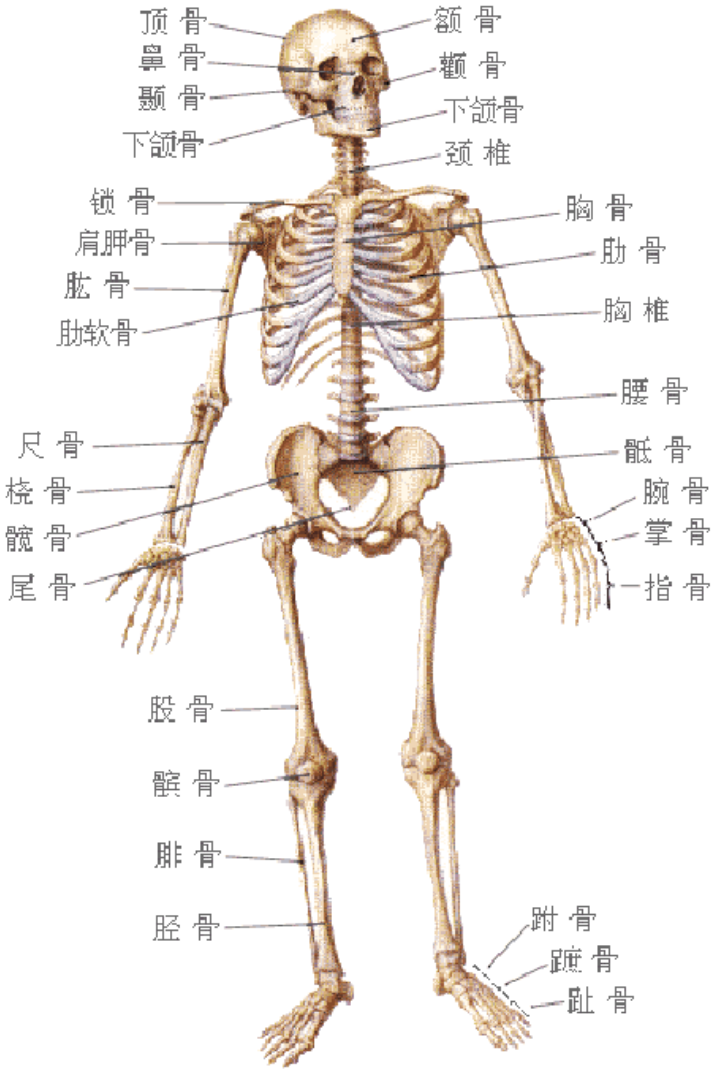
- 网络爬虫有什么用?
  - 获取信息
  - 自动操作、省时省力

- 为什么要用Python写网络爬虫?
  - 字符串处理方便
  - 第三方库完善

网页的组成

- **骨架：HTML**
  - Hyper Text Markup Language，超文本标记语言
  - 通过不同类型的标签，不同的元素相互嵌套和组合，形成网页的框架

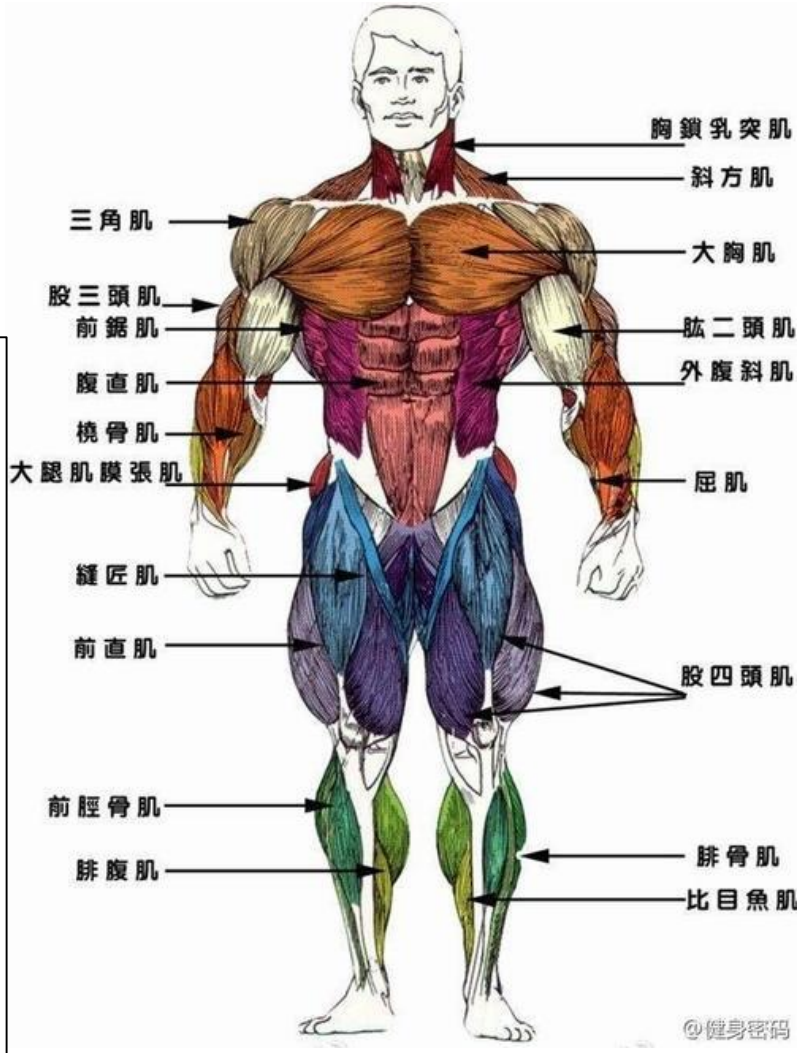
```
<h2 class="block-title">目录</h2>
<div class="catalog-list column-3">
<ol>
<li class="level1">
<span class="index">1</span>
<span class="text"><a href="#1">演艺经历</a></span>
</li>
<li class="level1">
<span class="index">2</span>
<span class="text"><a href="#2">个人生活</a></span>
</li>
<li class="level1">
<span class="index">3</span>
<span class="text"><a href="#3">主要作品</a></span>
</li>
```



网页的组成

- **肌肉：Javascript**
  - 通过JS脚本定义网页的行为
  - 例：交互、动画、播放音乐或视频……

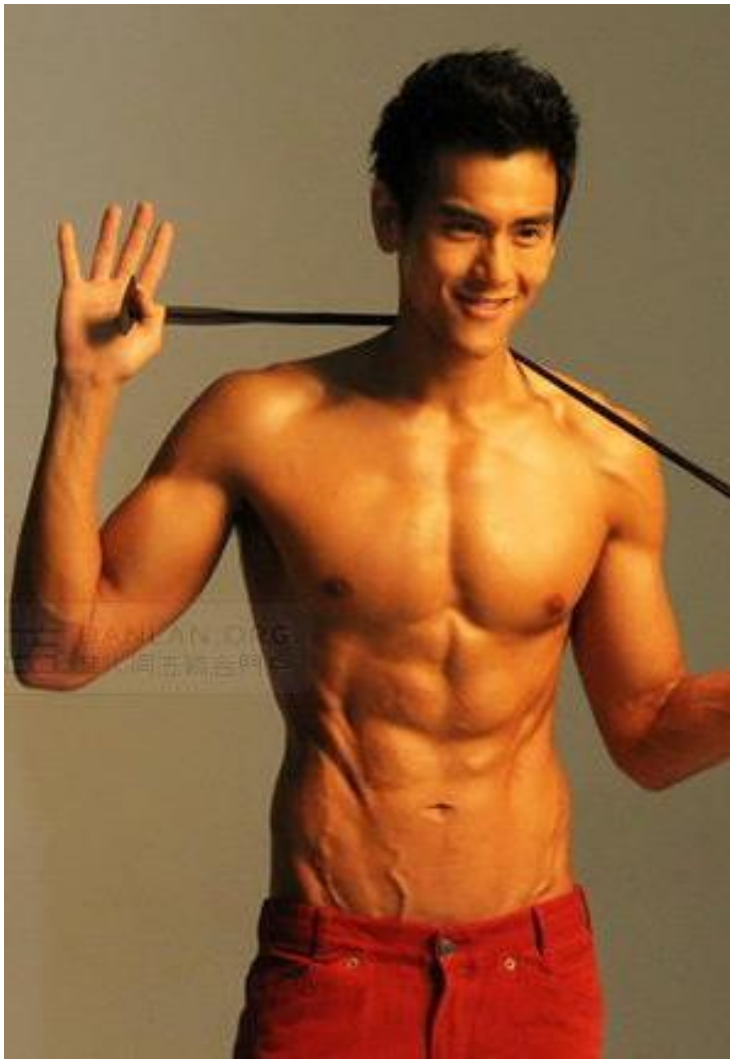
```
<script data-compress=strip>
    function h(obj){
        obj.style.behavior='url(#default#homepage)';
        var a = obj.setHomePage('//www.baidu.com/');
    }
</script>
<script>
    _manCard = {
        asynJs : [],
        asynLoad : function(id){
            _manCard.asynJs.push(id);
        }
    };
    window._sp_async = 1;

</script>
```



网页的组成

- **皮肤：CSS**
  - Cascading Style Sheets，层叠样式表
  - “样式”：网页中文字大小、颜色、元素间距、排列等格式
  - “层叠”：样式冲突时，根据顺序处理

```
.output_wrapper/*此属性为全局*/
{
    font-size: 16px;
    color: #3e3e3e;
    line-height: 1.6;
    word-spacing:0px;
    letter-spacing:0px;
    font-family: "Helvetica Neue",Helvetica,Arial,sans-serif;
}
p {/*段落*/
  margin: 1.5em 0px;
}
```



案例一：最简单的爬虫

- **第三方模块：requests**

[1]
```
import requests
```

[2]
```
url = "https://so.gushiwen.org/authors/authorvsw_b90660e3e492A1.aspx"
```

[3]
```
r = requests.get(url)
r
```

```
<Response [200]>
```

[28]
```
r.text
```

怎么从这些文本中提取有用信息?

```
lt="译文" onclick="OnYiwen(\'62802abab937\')" id="btnYiwen62802abab937" />\n<script type
document.getElementById("btnShangxi62802abab937").style.display = "block";\r\n
t/javascript">\r\n                                document.getElementById("btnZhushi62802abab937
</script>\n<script type="text/javascript">\r\n                                    document.getElemen
splay = "block";\r\n                          </script>\n</div>\n<p><a style="font-size:18px;
```

正则表达式

回 囘 囲 圓

- **内建模块：re**
- 正则表达式(regular expression)：一种字符串匹配的模式（pattern）
- 用途：在一个字符串中搜索、替换、提取符合某个条件的子串

[1]
```
import re
```

[2]
```
# 简单日期匹配
text = "星期一，星期二，星期三，星期四，星期五，星期六，星期日，星期1，星期2，星期3，星期4，星期5，星期6，星期7"
```

[3]
```
pattern = re.compile('星期[^\d]')
titles = re.findall(pattern, text) # 查找所有能匹配的项
titles
```

```
['星期一', '星期二', '星期三', '星期四', '星期五', '星期六', '星期日']
```

[4]
```
titles = re.sub(pattern, r"星期八",text) # 替换所有能匹配的项
titles
```

```
'星期八，星期八，星期八，星期八，星期八，星期八，星期八，星期1，星期2，星期3，星期4，星期5，星期6，星期7'
```

- 廖雪峰教程、正则表达式查询表、正则表达式在线测试

案例一：最简单的爬虫（续）

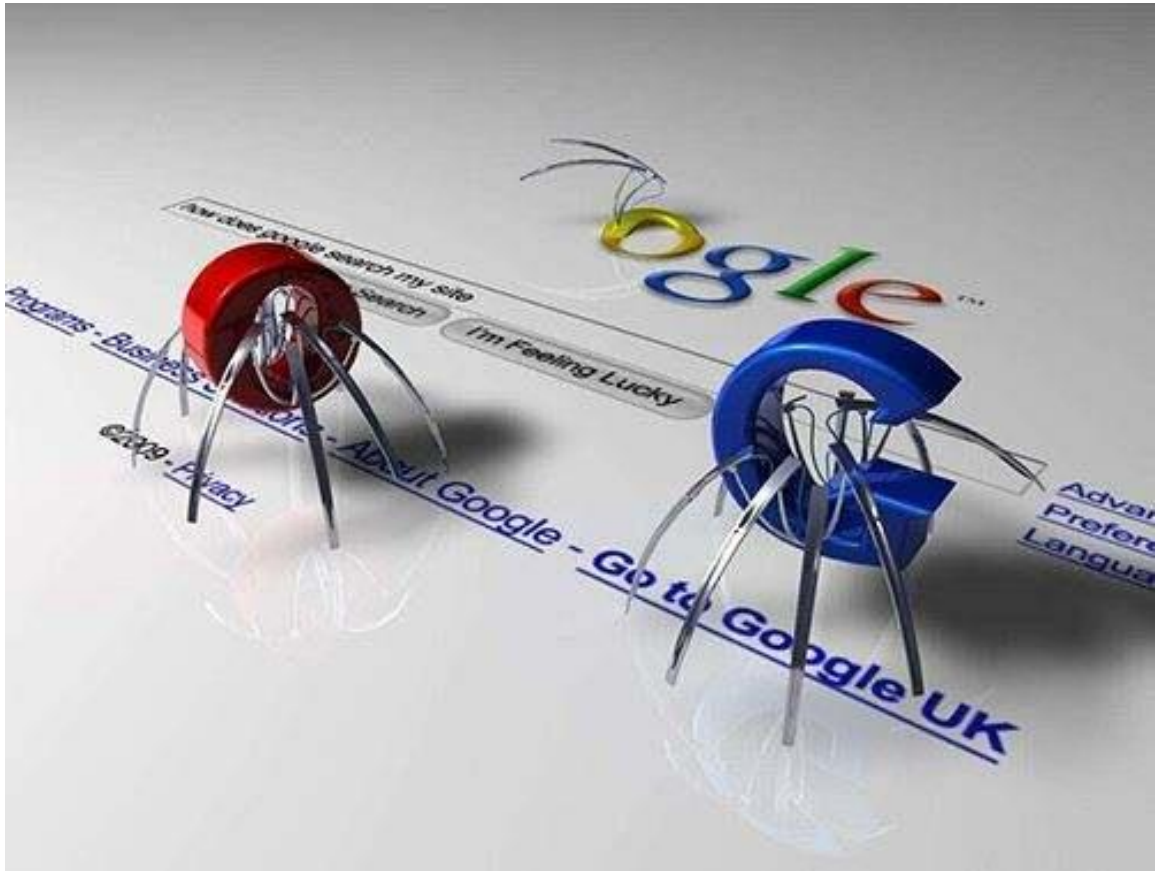
- **登录网页+提取信息**
  - **requests+re**

```
import re

# 匹配模板：首位分别为<div class="contson" xxxxxxxxxxx>和</div>的结构中间的部分
pattern = re.compile('<div class="contson" .*?>(.*?)</div>',re.S)
poetry = re.findall(pattern, r.text)
poetry
```

```
['\n君不见黄河之水天上来，奔流到海不复回。<br />君不见高堂明镜悲白发，朝如青丝暮成雪。<br />人
材必有用，千金散尽还复来。<br />烹羊宰牛且为乐，会须一饮三百杯。<br />岑夫子，丹丘生，将进酒，
(倾耳听 一作：侧耳听)<br />钟鼓馔玉不足贵，但愿长醉不愿醒。(不足贵 一作：何足贵；不愿醒 一作：
名。(古来 一作：自古；惟 通：唯)<br />陈王昔时宴平乐，斗酒十千恣欢谑。<br />主人何为言少钱，径
出换美酒，与尔同销万古愁。\n',
 '\n<p>噫吁嚱，危乎高哉！<br />蜀道之难，难于上青天！<br />蚕丛及鱼凫，开国何茫然！<br />尔来
有鸟道，可以横绝峨眉巅。<br />地崩山摧壮士死，然后天梯石栈相钩连。<br />上有六龙回日之高标，下
猿猱欲度愁攀援。(攀援 一作：攀缘)<br />青泥何盘盘，百步九折萦岩峦。<br />扪参历井仰胁息，以手抚
岩不可攀。<br />但见悲鸟号古木，雄飞雌从绕林间。<br />又闻子规啼夜月，愁空山。<br />蜀道之难，
不盈尺，枯松倒挂倚绝壁。<br />飞湍瀑流争喧豗，砯崖转石万壑雷。<br />其险也如此，嗟尔远道之人胡
阁峥嵘而崔嵬，一夫当关，万夫莫开。<br />所守或匪亲，化为狼与豺。<br />朝避猛虎，夕避长蛇，磨牙
```

第六讲：网络爬虫入门

- 爬虫基础
- **常规爬虫**
- 爬虫的法律问题
- Ajax爬虫
- 暑期学习营总结



常规爬虫

- 适用范围
  - 静态网页
  - HTML源代码中包含了所有的信息
- 基本操作
  - **登录网页，解析文本**
- 设计逻辑
  - 并不是一串代码能解决一切问题
  - 理解基本逻辑，进行针对性的设计

案例二：爬取百度百科图片

- 步骤：**词条→图集地址**→图片地址列表→读取图片并保存

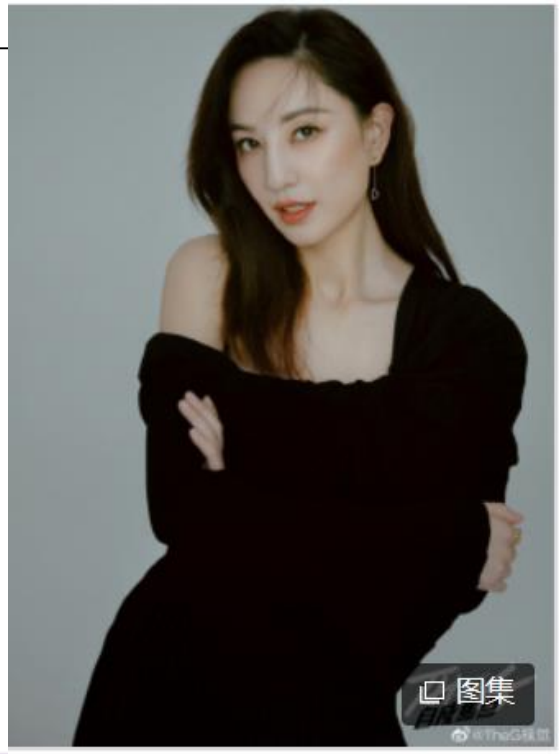
```
<div class="summary-pic">
<a nslog-type="10002401"
                    href="/pic/%E9%83%91%E5%B8%8C%E6%80%A1/805307/1/314e251f95cad1c8a786ca5
                >
<img src="https://bkimg.cdn.bcebos.com/pic/314e251f95cad1c8a786ca56e2777009c93d70cffb23?x-b
<button class="picAlbumBtn"><em></em><span>图集</span></button>
<div>郑希怡的概述图（43张）</div>
```

```
from bs4 import BeautifulSoup
```

```
# lxml是一种HTML解析器，也可以用html.parser等等
bs = BeautifulSoup(r.text, "lxml")
```

```
# 选择 class="summary-pic" 的标签里的 a 标签
pic_range = bs.select(".summary-pic a")
pic_range
```

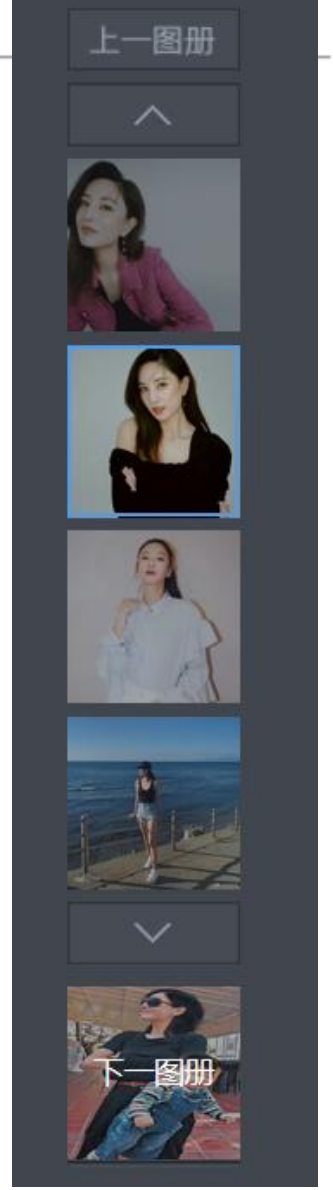
```
[<a href="/pic/%E9%83%91%E5%B8%8C%E6%80%A1/805307/1/314e251f95cad1c8
emma&amp;ct=single" nslog-type="10002401" target="_blank">
 <img src="https://bkimg.cdn.bcebos.com/pic/314e251f95cad1c8a786ca56
s=image/resize,m_lfit,w_268,limit_1/format,f_jpg"/>
 <button class="picAlbumBtn"><em></em><span>图集</span></button>
 <div>郑希怡的概述图（43张）</div>
 </a>]
```



案例二：爬取百度百科图片（续）

- 步骤：词条→**图集地址→图片地址列表**→读取图片并保存

```
<div class="pic-list">

<a class="pic-item " title="" data-index="0" data-sign="0ff41bd5ad6eddc45
/1/0ff41bd5ad6eddc451dad6865e92a1fd5266d016953f">
<img src="https://bkimg.cdn.bcebos.com/pic/0ff41bd5ad6eddc451dad6865e92a1f
block;width:80px;height:96px;top:0px;left:0px;">
</a>

<a class="pic-item " title="" data-index="1" data-sign="7dd98d1001e939019
/1/7dd98d1001e939019d00003f73ec54e736d196b5">
<img src="https://bkimg.cdn.bcebos.com/pic/7dd98d1001e939019d00003f73ec54e
block;width:121px;height:80px;top:0px;left:-20px;">
</a>

<a class="pic-item " title="" data-index="2" data-sign="d043ad4bd11373f08
/1/d043ad4bd11373f082023b9ed0465cfbfbedab64aed3">
<img src="https://bkimg.cdn.bcebos.com/pic/d043ad4bd11373f082023b9ed0465cf
block;width:80px;height:95px;top:0px;left:0px;">
</a>

<a class="pic-item " title="" data-index="3" data-sign="728da9773912b31bb
/1/728da9773912b31bb051ab271b51217adab44aed4b21">
```



案例二：爬取百度百科图片（续）

- 步骤：词条→图集地址→**图片地址列表**→读取图片并保存

```
    # 选择 class="pic-list" 的标签里的 img 标签
    pic_list_html = bs.select(".pic-list img")

    pic_urls = []
    for pic_html in pic_list_html:
        # 获取每个img标签中的src属性，即每张图片的地址
        pic_url = pic_html.get("src")
        pic_urls.append(pic_url)
    pic_urls
```

```
['https://bkimg.cdn.bcebos.com/pic/0ff41bd5ad6eddc451dad6865e92a1fd5266d016953f?x-bce-pr
esize,m_lfit,h_160,limit_1',
 'https://bkimg.cdn.bcebos.com/pic/7dd98d1001e939019d00003f73ec54e736d196b5?x-bce-proces
e,m_lfit,w_150,limit_1',
 'https://bkimg.cdn.bcebos.com/pic/d043ad4bd11373f082023b9ed0465cfbfbedab64aed3?x-bce-pr
esize,m_lfit,h_160,limit_1',
 'https://bkimg.cdn.bcebos.com/pic/728da9773912b31bb051ab271b51217adab44aed4b21?x-bce-pr
esize,m_lfit,w_150,limit_1',
 'https://bkimg.cdn.bcebos.com/pic/a8773912b31bb051f819d226ab33cdb44aed2e734a21?x-bce-pr
esize,m_lfit,w_150,limit_1',
 'https://bkimg.cdn.bcebos.com/pic/3812b31bb051f8198618634447fd5ded2e738bd44921?x-bce-pr
esize,m_lfit,w_150,limit_1'
```

案例二：爬取百度百科图片（续）

- 步骤：词条→图集地址→图片地址列表→**读取图片并保存**

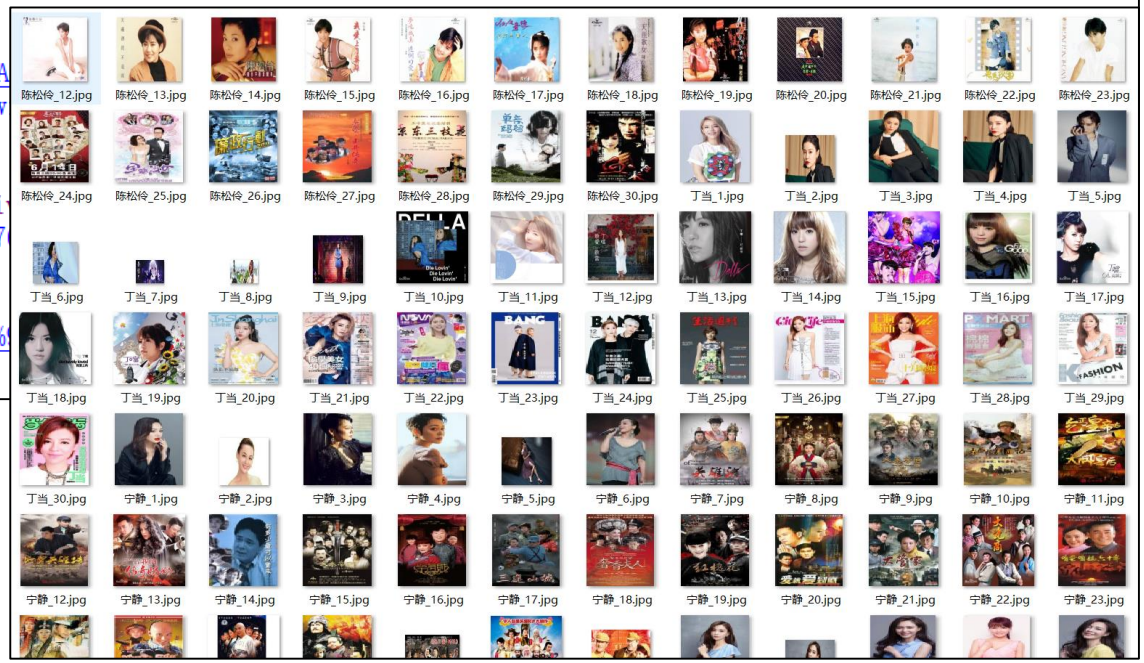
```
    for i, pic_url in enumerate(pic_urls):
        pic = requests.get(pic_url, timeout=15)

        # 合成图片的文件名
        pic_name = name + "_" + str(i + 1) + ".jpg"
        filename = os.path.join(path, pic_name)

        with open(filename, "wb") as f:
            f.write(pic.content)
            print("成功下载: %s" % (filename))
```

```
成功下载: xjj\郑希怡_1.jpg
成功下载: xjj\郑希怡_2.jpg
成功下载: xjj\郑希怡_3.jpg
成功下载: xjj\郑希怡_4.jpg
```

案例二扩展：爬取更多图片

- 爬取《乘风破浪的姐姐》的各位参赛嘉宾的图片
- 步骤：**母词条→子词条**→图集地址→图片地址列表→读取图片并保存

```
itle-text"><span class="title-prefix">乘风破浪的姐姐</span>参赛嘉宾</h3>

para" label-module="para">*（按姓氏首字母排序）</div><table log-set-param="table_view"  data-sort="sortDisabled"><tr><td valign=
style="width:100%;height:auto;max-width:129.96575342466px;"

age-link" nslog-type="9317"
href="/pic/%E4%B9%98%E9%A3%8E%E7%A0%B4%E6%B5%AA%E7%9A%84%E5%A7%90%E5%A7%90/49998987/0/f7246b600c338744ebf8fa670046cef9d72a60590adc
title=""

"lazy-img" src="
                style="width:100%;height:100%;max-w
        />

</td><td valign="top" align="left" width="174"><di
style="width:100%;height:auto;max-width:125.8706467

age-link" nslog-type="9317"
href="/pic/%E4%B9%98%E9%A3%8E%E7%A0%B4%E6%B5%AA%E7%
title=""
```



第六讲：网络爬虫入门

- 爬虫基础
- 常规爬虫
- **爬虫的法律问题**
- Ajax爬虫
- 暑期学习营总结

爬虫的法律问题

- 爬虫玩得好，监狱进的早；数据玩的溜，牢饭吃个够
- **侵犯公民个人信息罪**：采集公民的姓名、身份证件号码、通信通讯联系方式、住址、账号密码、财产状况、行踪轨迹等个人信息，并将之**用于非法途径**的
  - 现金贷风控、不合规的P2P、赌博类游戏、黑五类产品……
  - 友情提示：爬虫可能违法了
  - 新三板挂牌公司涉窃取30亿条个人信息，非法牟利超千万元
- **非法获取计算机信息系统数据罪**：爬虫程序规避网站经营者设置的反爬虫措施或者破解服务器防抓取措施，或强行访问正常情况不能到达的页面；非法获取相关信息
- **破坏计算机信息系统罪**：爬虫程序干扰被访问的网站正常运营（近乎DDOS 的请求频率，造成服务器瘫痪）

- 法律依据：中华人民共和国网络安全法

爬虫的法律问题

- 灰色产业
- 12306刷票软件
  - “最高峰时1天内页面浏览量达813.4亿次，1小时最高点击量59.3亿次，平均每秒164.8万次。”

- 社交平台僵尸粉、直播平台刷量、电商刷单、偷取社区内容……

- 爬虫技术本身中立，法律不禁止爬虫
- 搜索引擎、聚合导航、数据分析、人工智能等业务，都需要基于爬虫技术
- **不要爬取个人信息，不要利用爬虫非法获利，不要爬取网站的付费内容**

第六讲：网络爬虫入门

- 爬虫基础
- 常规爬虫
- 爬虫的法律问题
- **Ajax爬虫**
- 暑期学习营总结

案例三：学堂在线字幕

- 视频、文字在哪里？

```
<body ondragstart="return false">
  <div id="app"></div>

  <script>
    var hostName = window.location.hostname;

    var _mtac = {"performanceMonitor":1,"senseQuery":1};
    (function () {
      var mta = document.createElement("script");
      mta.src = "//pingjs.qq.com/h5/stats.js?v2.0.4";
      mta.setAttribute("name", "MTAH5");
      if((hostName.indexOf("jsmh.xuetangx.com") !=-1 || hostName.indexOf("www.bnuonline.com") !=-1)){
        mta.setAttribute("sid", "500693653");
      }else{
        mta.setAttribute("sid", "500676615");
        mta.setAttribute("cid", "500679396");
      }
      var s = document.getElementsByTagName("script")[0];
      s.parentNode.insertBefore(mta, s);
    })();
  </script>
<script src="https://res.wx.qq.com/open/js/jweixin-1.6.0.js" type="text/javascript"></script>
<script src="https://static-cdn.xuetangx.com/xtassets/manifest_5eddc375c98dde9265f4.js"></script><scri
</script><script src="https://static-cdn.xuetangx.com/xtassets/10_5be8339c0d985546f870.js"></script><s
/101_f43c980e761d8a2f5621.js"></script></body>
```

- HTML 代码仅是一个空壳，整个网页都是由 JavaScript 渲染出来的

Ajax抓取

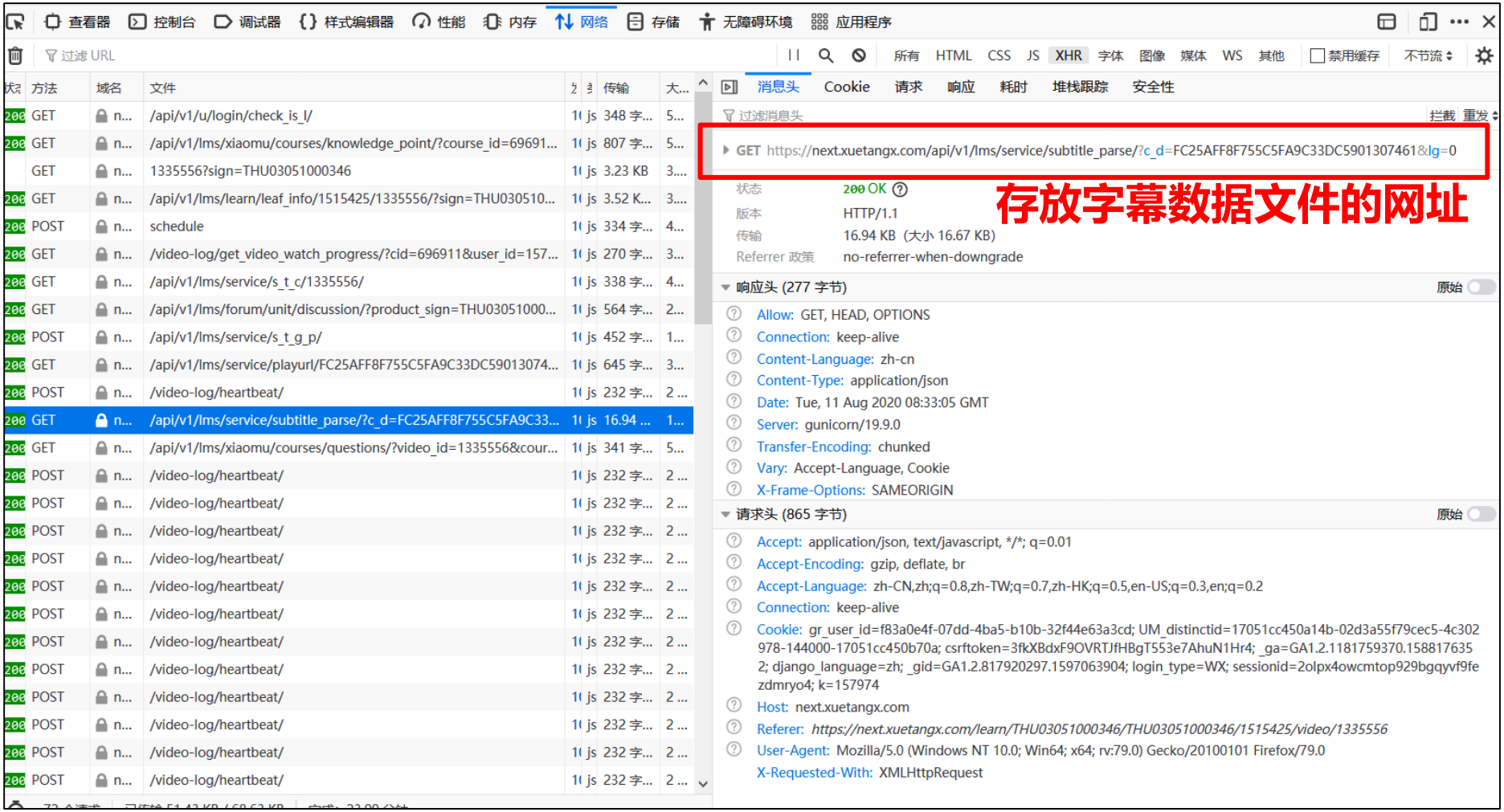
- Ajax（Asynchronous JavaScript and XML，异步 JavaScript 和 XML）
  - 采用 Ajax、前端模块化工具来构建
  - 其 API 接口通常采用 JSON 格式

- 爬取方法
  - **分析其后台 Ajax 接口**
  - 使用 Selenium、Splash 等库来实现模拟 JavaScript 渲染

- 查看网络传输数据流的方法
  - 右键→查看元素→点击网络
  - 快捷键：F12
  - 菜单→Web开发者→网络

案例三：学堂在线字幕

□ 分析其后台 Ajax 接口



状	方法	域名	文件			传输	大...
200	GET	n...	/api/v1/u/login/check_is_l/	1	js	348 字...	5...
200	GET	n...	/api/v1/lms/xiaomu/courses/knowledge_point/?course_id=69691...	1	js	807 字...	5...
	GET	n...	1335556?sign=THU03051000346	1	js	3.23 KB	3....
200	GET	n...	/api/v1/lms/learn/leaf_info/1515425/1335556/?sign=THU030510...	1	js	3.52 K...	3....
200	POST	n...	schedule	1	js	334 字...	4...
200	GET	n...	/video-log/get_video_watch_progress/?cid=696911&user_id=157...	1	js	270 字...	3...
200	GET	n...	/api/v1/lms/service/s_t_c/1335556/	1	js	338 字...	4...
200	GET	n...	/api/v1/lms/forum/unit/discussion/?product_sign=THU03051000...	1	js	564 字...	2...
200	POST	n...	/api/v1/lms/service/s_t_g_p/	1	js	452 字...	1...
200	GET	n...	/api/v1/lms/service/playurl/FC25AFF8F755C5FA9C33DC59013074...	1	js	645 字...	3...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	GET	n...	/api/v1/lms/service/subtitle_parse/?c_d=FC25AFF8F755C5FA9C33...	1	js	16.94 ...	1...
200	GET	n...	/api/v1/lms/xiaomu/courses/questions/?video_id=1335556&cour...	1	js	341 字...	5...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...
200	POST	n...	/video-log/heartbeat/	1	js	232 字...	2 ...

GET https://next.xuetangx.com/api/v1/lms/service/subtitle_parse/?c_d=FC25AFF8F755C5FA9C33DC5901307461&lg=0

存放字幕数据文件的网址

状态	200 OK
版本	HTTP/1.1
传输	16.94 KB（大小 16.67 KB）
Referrer 政策	no-referrer-when-downgrade

响应头 (277 字节) 原始

```
Allow: GET, HEAD, OPTIONS
Connection: keep-alive
Content-Language: zh-cn
Content-Type: application/json
Date: Tue, 11 Aug 2020 08:33:05 GMT
Server: gunicorn/19.9.0
Transfer-Encoding: chunked
Vary: Accept-Language, Cookie
X-Frame-Options: SAMEORIGIN
```

请求头 (865 字节) 原始

```
Accept: application/json, text/javascript, */*; q=0.01
Accept-Encoding: gzip, deflate, br
Accept-Language: zh-CN,zh;q=0.8,zh-TW;q=0.7,zh-HK;q=0.5,en-US;q=0.3,en;q=0.2
Connection: keep-alive
Cookie: gr_user_id=f83a0e4f-07dd-4ba5-b10b-32f44e63a3cd; UM_distinctid=17051cc450a14b-02d3a55f79cec5-4c302
978-144000-17051cc450b70a; csrftoken=3fkXBdxF9OVRTJfHBgT553e7AhuN1Hr4; _ga=GA1.2.1181759370.158817635
2; django_language=zh; _gid=GA1.2.817920297.1597063904; login_type=WX; sessionid=2olpx4owcmtop929bgqyvf9fe
zdmryo4; k=157974
Host: next.xuetangx.com
Referer: https://next.xuetangx.com/learn/THU03051000346/THU03051000346/1515425/video/1335556
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:79.0) Gecko/20100101 Firefox/79.0
X-Requested-With: XMLHttpRequest
```

案例三：学堂在线字幕

- **读取JSON**

```
subtitle_url = r"https://next.xuetangx.com/api/v1/lms/service/subtitle_parse/?c_d=AAE95227B6C09E6A9C33DC5901307461&lg=0"
```

```
r = requests.get(subtitle_url)
data = r.json()
```

```
data["text"]
```

```
['我们再来看第二个问题',
 '毛泽东为什么提出',
 '“马克思主义中国化”',
 '毛泽东之所以提出要实现',
 '马克思主义的中国化',
 '源于对中国革命进程中',
 '正反两个方面的实践经验的科学总结',
 '在第一、二次国内',
 '革命战争时期',
 '中国共产党经历过',
 '两次胜利和两次失败',
 '大家知道是哪两次吗?',
 '第一次其实指的是',
 '北伐战争的胜利与国民革命的失败',
```

进阶方向*

- 提升爬虫效率
  - 多线程、多进程
  - 分布式、集群化
  - 大数据存储

- 提升反爬技术
  - 浏览器控制
    - Selenium、Splash
  - **APP/JavaScript逆向**
  - **代理池**
  - **验证码破解**

不是说问题不大么

课后练习

1. **（二选一）**爬取某个百度百科的系列图片

2. **（二选一）**爬取古诗文网的某位诗人的前五页作品
   https://so.gushiwen.org/authors/

课程回顾

- **培训目标：**
  - 上手一个好用、主流的工具 ⟹ 在哪写、在哪运行、在哪看结果
  - 了解最主要的语法特征 ⟹ 对于细节问题，知道去哪查、怎么查
  - 了解python与外界的交互 ⟹ 可以怎么与实用场景联系起来

周次	课程大纲	课程内容
第一周	认识Python	课程介绍；安装与使用
第二周	Python的基本语法	基础语法元素与结构
		类与对象、模块
第三周	Python的简单应用	文档处理入门
		数据处理入门
		网络爬虫入门

清華大學
Tsinghua University

关心学习，更关心你

清华大学学生学习与发展指导中心

答疑坊

只为你答疑解惑

一对一咨询

一心为你

唐仲英计划

全面提升你的公共领导力

写作助理

最贴心，最有针对性的写作辅导

官方微信公众号：乐学

扫码关注，获取一手资讯和资源

关注乐学，参与学协活动！

也欢迎加入学协大家庭！

清华大学学生学业发展协会

清华大学学生学习与发展中心指导

清华大学十佳社团协会

一个关心学习，更关心你的社团

小伙伴计划

线上打卡、线下交流

工作坊

“私人订制”的小型讲座

全清华GPA最高的社团

全清华最关心学习的社团

反馈问卷