

Python及其应用

主讲人：钱惠敏

E-mail: amandaqian@hhu.edu.cn

第11讲 网络爬虫

11.1 网络爬虫简介

11.2 爬取网络小说_

11.3 爬取瀑布流图片

11.1 网络爬虫简介

➤网络爬虫

又被称为网页蜘蛛，网络机器人，是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。

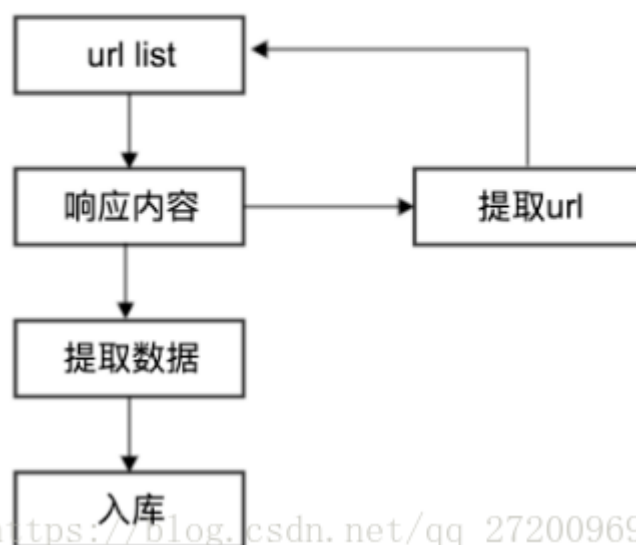
➤类型

- ✓通用爬虫：通常指搜索引擎的爬虫
- ✓聚焦爬虫：针对特定网站的爬虫

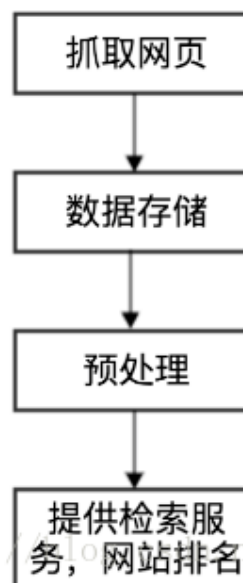
11.1 网络爬虫简介（续1）

➤ 爬虫的流程

聚焦爬虫流程



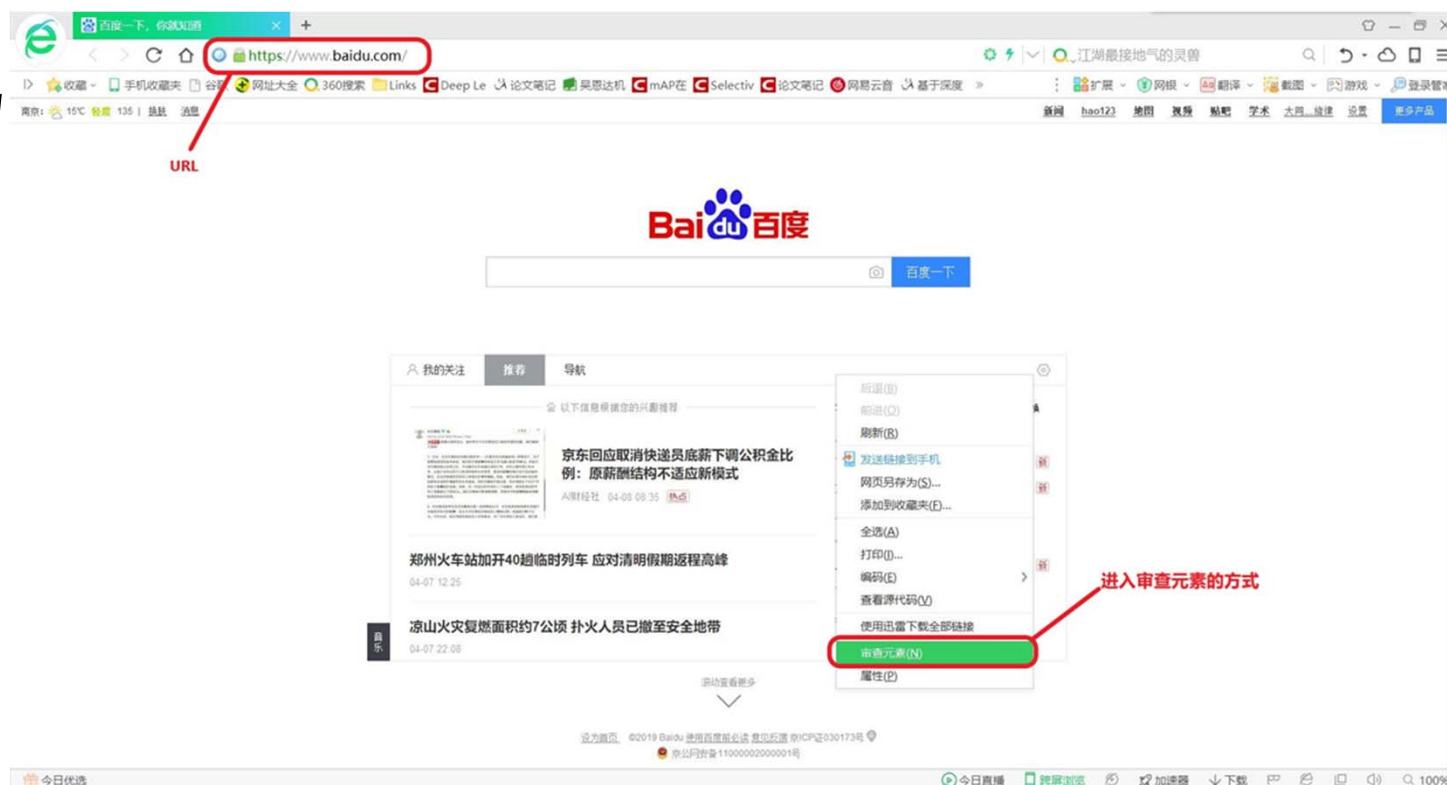
搜索引擎流程



11.2 爬取网络小说

➤ 预备知识

认识URL



11.2 爬取网络小说（续1）



11.2 爬取网络小说（续2）

➤ 安装requests库

| 方法 | 说明 |
|---------------------------------|---------------------------------|
| <code>requests.request()</code> | 构造一个请求，支撑以下各方法的基础方法 |
| <code>requests.get()</code> | 获取HTML网页的主要方法，对应于HTTP的GET |
| <code>requests.head()</code> | 获取HTML网页头信息的方法，对应于HTTP的HEAD |
| <code>requests.post()</code> | 向HTML网页提交POST请求的方法，对应于HTTP的POST |
| <code>requests.put()</code> | 向HTML网页提交PUT请求的方法，对应于HTTP的PUT |
| <code>requests.patch()</code> | 向HTML网页提交局部修改请求，对应于HTTP的PATCH |
| <code>requests.delete()</code> | 向HTML网页提交删除请求，对应于HTTP的DELETE |

11.2 爬取网络小说（续3）

➤ Beautiful Soup库

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml" class="sui-componentWrap">
<head>...</head>
<body class=" s-manhattan-index">
  <div id="s_is_index_css" style="display:none;">...</div>
  <textarea id="s_is_result_css" style="display:none;">...</textarea>
  <textarea id="s_index_off_css" style="display:none;">...</textarea>
  <div id="wrapper">
    <div class="s-skin-container s-isindex-wrap"> </div>
    <div id="head" class=">...</div>
    <div class="s_tab" id="s_tab">...</div>
    <div id="wrapper_wrapper"></div>
  </div>
  <input type="hidden" id="bsToken" name="bsToken" value="f2decc0656e40364e233e070337b6400">
  <script>...</script>
  <script>...</script>
  <script src="https://ss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/js/lib/jquery-1.10.2_1c4228b8.js"></script>
  <script>...</script>
  <script type="text/javascript">...</script>
  <script type="text/javascript">...</script>
  <script src="https://ssl1.bdstatic.com/5eN1bjq8AAUYm2zgoY3K/r/www/cache/static/protocol/https/global/js/all_async_search_04cd1bb.js"></script>
  <script>...</script>
  <script src="https://ss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/js/sbase_a65ca873.js"></script>
  <style type="text/css">...</style>
  <script type="text/javascript">...</script>
  <script type="text/javascript">...</script>
  <script src="https://ss0.bdstatic.com/5aV1bjqh_Q23odCf/static/mancard/js/config_1a9098e0.js"></script>
  <script id="s_js_mancard_main" data-src="https://ss0.bdstatic.com/5aV1bjqh_Q23odCf/static/mancard/js/mancard_2ee35e99.js">
    _manCard.asyncLoad("s_js_mancard_main");
  </script>
  <script src="https://ss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/js/min_super_fdb28b91.js"></script>
  <script>...</script>
  <script id="s_js_setting" data-src="https://ss0.bdstatic.com/5aV1bjqh_Q23odCf/static/superman/js/min_setting_7c10d417.js"></script>
  <script id="tipsplus-js" data-src="https://ss0.bdstatic.com/5aV1bjqh_Q23odCf/static/tipsplus/js/min_tips_e4616384.js" src="https://ss0.bdstatic.com/5aV1bjqh_Q23odCf/static/tipsplus/js/min_tips_e4616384.js"></script>
  <script data-onload="true" data-src="https://ss0.bdstatic.com/5aV1bjqh_Q23odCf/static/activity/js/activity_start_52498d2c.js" src="https://ss0.bdstatic.com/5aV1bjqh_Q23odCf/static/activity/js/activity_start_52498d2c.js"></script>
  <script>...</script>
  <span id="s_strpx_span1" style="visibility:hidden;position:absolute;bottom:0;left:0;font-weight:bold;font-size:12px;font-family:'arial';">中</span>
</html>
```

HTML信息

11.2 爬取网络小说（续4）

➤例子：爬取网站上的小说

(1) 小说URL: <https://www.qb50.com/>

契子URL: https://www.qb50.com/book_44997/49540588.html

(2) 运用之前介绍的requests.get()方法获取HTML信息

```
import requests
if __name__ == '__main__':
    target = 'https://www.qb50.com/book_44997/49540588.html'
    req = requests.get(url = target, verify = False)
    print(req.text)
```

11.2 爬取网络小说（续4）

➤例子：爬取网站上的小说

(1) 小说URL: <https://www.qb50.com/>

契子URL: https://www.qb50.com/book_44997/49540588.html

(2) 运用之前介绍的requests.get()方法获取HTML信息

```
import requests
if __name__ == '__main__':
    target = 'https://www.qb50.com/book_44997/49540588.html'
    req = requests.get(url = target, verify = False)
    print(req.text)
```


11.2 爬取网络小说

[<div id="content"> 全本小说网 www.qb50.com, 最快更新深空彼岸最新章节!

 红日西坠, 列车远去, 在与铁轨的震动声中带起大片枯黄的落叶, 也带起秋的萧瑟。

 王煊注视, 直至列车渐消失, 他才收回目光, 又送走了几位同学。

 自此一别, 将天各一方, 不知道多少年后才能再相见, 甚至有些人再无重逢期。

 周围, 有人还在缓慢地挥手, 久久未曾放下, 也有人沉默着, 颇为伤感。

 大学四年, 一起走过, 积淀下的情谊总有些难以割舍。

 落日余晖斜照飘落的黄叶, 光影斑驳, 交织出几许岁月流逝之感。

 一位清秀的女生转过身去, 暗自擦去镜片后的眼泪。

 在这个特殊的年代, 毕业后他们将各自归去, 此生可能都不再相遇。

 秋风吹过, 黄叶凌乱, 纷纷扬扬。

 在这个季节, 有人失意, 有人得意。

 毕业四个月了, 有人留在了这座城市, 前程灿烂, 也有人在忐忑中等待, 坚守, 而更多的人则怅然离去, 将回故里。

 > 王煊走在回去的路上, 也在想自己将何去何从。

 街道陈旧, 路两旁的梧桐树大片地坠落叶子, 满地都是。

 有人与他并肩走在一起, 为他鸣不平: “留下的人没有你, 为什么会这样? 他们竟将你放弃!”

 身为同窗兼好友, 在秦诚看来, 但凡有名额都绕不开王煊, 他必然会被选中。

 结果出来后, 许多人心情复杂, 王煊居然落选。

 “不说我了, 你怎么样, 有结果了吗?” 王煊问他。

 秦诚小声告诉他, 家里托了关系, 可能要去新月。

 “新月, 深空对岸, 不知道以后我们还能不能再见。” 王煊停下脚步, 身边的好友都将远行了。

 他身材颀长, 并不单薄, 匀称有力, 在晚霞中, 身上有一层淡金光彩, 一双眼睛清澈而有神。

 “我会回来的, 肯定还能相见。” 秦诚是个感性的人, 难舍故土, 尤其是想到, 很难再见到好友, 心中有些不好受。

 “回来时喊我!” 王煊用力抱了抱他的肩头。

 风中有哽咽声传来, 王煊与秦诚回头, 看到一位男同学情绪很激动。

 他脸色苍白, 哭出声来, 用力喊着: “我真的很想留在这座城市, 想等到最后的机会, 我不想这样回老家!”

 在这里生活与学习了四年, 他已经很努力了, 拼搏、争取、规划自己的未来, 想找到自己的位置, 但终究留不下来。

 他失声痛哭。

 秋风带着凉意, 一些同窗跟着心情低落。

 另一边, 一对情侣停下脚步, 彼此相顾, 没有言语, 只是在无声地落泪。

 他们将分别, 从此以后相距不是数千里, 而是隔着一片星空, 此生或许再也见不到了。

 两人脸上满是泪水, 最后一次相拥, 而相对而言, 整座城市承接过去的风格, 在岁月中保存下来。

 其他地方, 有些旧时代留下的城市则废弃了, 久无人迹, 大面积的荒芜, 爬满藤蔓, 荆棘丛生, 渐渐被草木淹没。

 回到校区后, 秦诚还在为王煊不忿, 劝他去找人了解原因, 为什么被放弃, 讨一个说法。

 即便毕业了, 他们也被允许住宿在校区, 直到确定完最终的所有人选。

 这次机会难得, 被选中的人留在这座城市等待, 不久后将前往新星, 那边似乎有了某种非同寻常的发现。

 秦诚也没有被选中, 他家里人费尽力气也只是给了他进入深空的机会。

 他将前往新月, 那颗围绕新星转动的月亮, 是新星外最重要的基地。

 秦诚低声道: “你知道吗, 即便现在那边只有一鳞半爪的传闻, 也已经让提前得到小道消息的人热血沸腾。无论如何, 你都要得到一个名额!”

 月色下, 树影婆娑, 王煊在草坪上舒展身体, 他在演练旧时代的“散术”, 实战性极强, 将地面上大量的黄叶都带动的飞舞了起来, 漫天都是。

 他没有停下, 动作很快, 但呼吸平稳, 道: “我有名额了的结果。”

 王煊停下, 直起身来, 仰头望向天空, 喃喃道: “深空彼岸, 深空彼岸, 深空彼岸。”

]

```
if __name__ == '__main__':
    target= 'https://www.qb50.com/book_44997/49540588.html'
    req=requests.get(url = target,verify = False)
    html=req.text
    bf=BeautifulSoup(html,features='lxml')
    texts=bf.find_all('div',id='content')
    print(texts)
```




11.2 爬取网络小说

```
if __name__ == '__main__':
    target= 'https://www.qb50.com/book_44997/49540588.html'
    req=requests.get(url = target,verify = False)
    html=req.text
    bf=BeautifulSoup(html,'lxml')
    texts=bf.find_all('div',id='content')
    print(texts[0].text.replace('\xa0'*4,'\n\n'))
```

11.2 爬取网络小说

全本小说网 www.qb50.com，最快更新深空彼岸最新章节！

红日西坠，列车远去，在与铁轨的震动声中带起大片枯黄的落叶，也带起秋的萧瑟。

王焯注视，直至列车渐消失，他才收回目光，又送走了几位同学。

自此一别，将天各一方，不知道多少年后才能再相见，甚至有些人再无重逢期。

周围，有人还在缓慢地挥手，久久未曾放下，也有人沉默着，颇为伤感。

大学四年，一起走过，积淀下的情谊总有些难以割舍。

落日余晖斜照飘落的黄叶，光影斑驳，交织出几许岁月流逝之感。

一位清秀的女生转过身去，暗自擦去镜片后的眼泪。

在这个特殊的年代，毕业后他们将各自归去，此生可能都不再相遇。

秋风吹过，黄叶凌乱，纷纷扬扬。

在这个季节，有人失意，有人得意。

毕业四个月了，有人留在了这座城市，前程灿烂，也有人在忐忑中等待，坚守，而更多的人则怅然离去，将回故里。

王焯走在回去的路上，也在想自己将何去何从。

街道陈旧，路两旁的梧桐树大片地坠落叶子，满地都是。

有人与他并肩走在一起，为他鸣不平：“留下的人没有你，为什么会这样？他们竟将你放弃！”

11.2 爬取网络小说

```
[<dl class="zjlist"><dt class="ttname"><h2>正文</h2></dt><dd><a href="49540588.html">第一章 旧土</a></dd><dd><a href="49540589.html">第二章 韶华易逝</a></dd><dd><a href="49540590.html">第三章 续命项目</a></dd><dd><a href="49540591.html">第四章 超自然</a></dd><dd><a href="49540592.html">第五章 弃若敝履</a></dd><dd><a href="49540593.html">第六章 女神</a></dd><dd></dd><dd></dd><dt class="ttname"><h2>正文卷</h2></dt><dd><a href="49540714.html">第七章 列仙不存</a></dd><dd><a href="49542133.html">第八章 聚会</a></dd><dd><a href="49542902.html">第九章 同窗</a></dd><dd><a href="49544872.html">第十章 新术</a></dd><dd><a href="49545847.html">第十一章 新旧争锋</a></dd><dd><a href="49547080.html">第十二章 温和俯视</a></dd><dd><a href="49547765.html">第十三章 旧术路的尽头</a></dd><dd><a href="49550616.html">第十四章 探险</a></dd><dd><a href="49551416.html">第十五章 羽化</a></dd><dd><a href="49553224.html">第十六章 银色兽皮书</a></dd><dd><a href="49553897.html">第十七章 截胡</a></dd><dd><a href="49555512.html">第十八章 偶遇</a></dd><dd><a href="49556362.html">第十九章 前女友</a></dd><dd><a href="49560293.html">第二十章 小王太猛</a></dd><dd><a href="49561229.html">第二十一章 不许成精</a></dd><dd><a href="49562079.html">第二十二章 与死亡擦肩</a></dd><dd><a href="49563139.html">第二十三章 超感</a></dd><dd><a href="49564527.html">第二十四章 先秦竹简的正确开启方式</a></dd><dd><a href="49565601.html">第二十五章 接触神秘</a></dd><dd><a href="49566451.html">第二十六章 不具普适应</a></dd><dd><a href="49567588.html">第二十七章 仙坟</a></dd><dd></dd><dd></dd></div></div>]
```

```
if __name__ == '__main__':
    target= 'https://www.qb50.com/book_44997/'
    req=requests.get(url = target,verify = False)
    html=req.text
    div_bf=BeautifulSoup(html,'lxml')
    div =div_bf.find_all('dl',class_='zjlist')
    print(div)
```

11.2 爬取网络小说

```
<div class="zjbox">
  <dl class="zjlist">
    <dt class="ttname">...</dt>
    <dd>
      <a href="49540588.html">第一章 旧土</a>
    </dd>
    <dd>
      <a href="49540589.html">第二章 韶华易逝</a>
    </dd>
    <dd> == $0
      <a href="49540590.html">第三章 续命项目</a>
    </dd>
    <dd>
      <a href="49540591.html">第四章 超自然</a>
    </dd>
    <dd>
      <a href="49540592.html">第五章 弃若敝屣</a>
    </dd>
    <dd>...</dd>
    <dd></dd>
    <dd></dd>
    <dd></dd>
  </dl>
</div>
```

```
from bs4 import BeautifulSoup
import requests

import urllib3
urllib3.disable_warnings()

if __name__ == '__main__':
    server='https://www.qb50.com/'
    target= 'https://www.qb50.com/book_44997/'
    req=requests.get(url = target,verify = False)
    html=req.text
    div_bf=BeautifulSoup(html,features='lxml')
    div =div_bf.find_all('dl',class_='zjlist')
    a_bf=BeautifulSoup(str(div),features='lxml')
    a=a_bf.find_all('a')
    for each in a:
        print(each.string,server+each.get('href'))
```

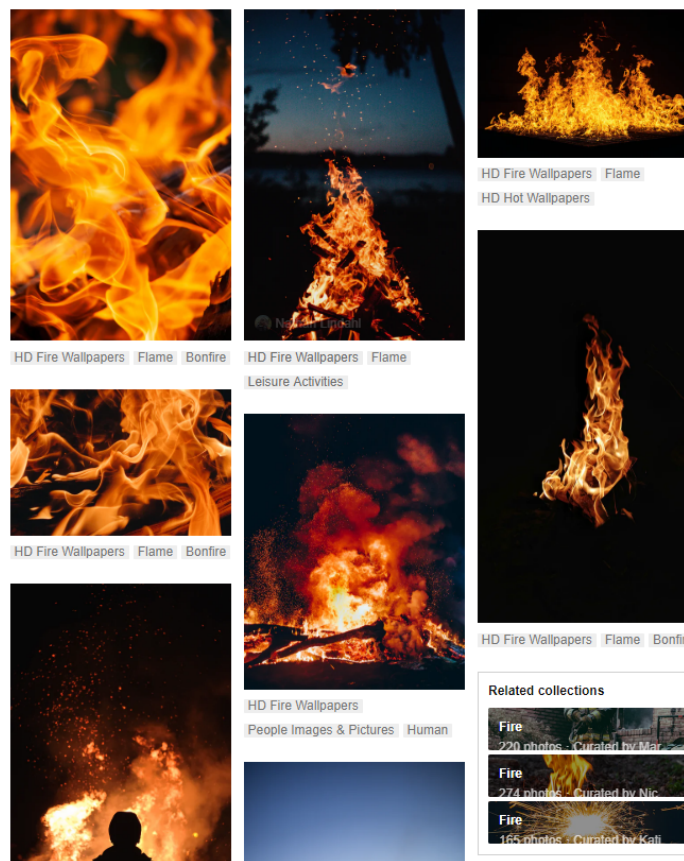
11.2 爬取网络小说

➤ 程序整合

noveldownload.py

11.3 爬取瀑布流图片

➤ 例子：Unsplash爬取fire图像



11.3 爬取瀑布流图片

➤例子：Unsplash爬取fire图像

```
<img class="_2zEKz" style="background-color: rgb(242, 199, 176);" alt="close-up photo of fire at nighttime" src="http  
s://images.unsplash.com/photo-1496483353456-90997957cf99?ixlib=rb-1.2.1&ixid=eyJhcHBfaWQiOjEyMDd9&w=1000&q=80" sizes  
="(min-width: 1335px) 416px, (min-width: 992px) calc(calc(100vw - 72px) / 3), (min-width: 768px) calc(calc(100vw - 48  
px) / 2), 100vw" data-test="photo-grid-multi-col-img" itemprop="thumbnailUrl" srcset="https://images.unsplash.com/pho
```


11.3 爬取瀑布流图片

➤ Flidder

Progress Telerik Fiddler Web Debugger

File Edit Rules Tools View Help

WinConfig WinConfig Replay X Go Stream Decode Keep: All sessions Any Process Find Save Browse Clear Cache TextWizard Tearoff MSDN Search... Online

| # | Result | Protocol | Host | URL | Body | Caching | Content-Type |
|-----|--------|----------|-----------------------|--------------------------------|---------|-------------|-----------------|
| 262 | 200 | HTTP | curl.f.360.cn | /vinfo.php | 850 | no-cache | application/... |
| 263 | 200 | HTTP | Tunnel to | settings-win.data.microso... | 0 | | |
| 264 | 200 | HTTP | Tunnel to | settings-win.data.microso... | 0 | | |
| 265 | 200 | HTTP | Tunnel to | settings-win.data.microso... | 0 | | |
| 266 | 200 | HTTP | Tunnel to | senry.io:443 | 0 | | |
| 267 | 200 | HTTP | Tunnel to | unsplash.com:443 | 0 | | |
| 268 | 200 | HTTPS | unsplash.com | /search/photos/fire | 51,374 | | text/html; c... |
| 269 | 200 | HTTP | Tunnel to | cdn.speedcurve.com:443 | 0 | | |
| 270 | 200 | HTTPS | unsplash.com | /a/third-party/snow-2.9.1.js | 25,990 | max-ag... | application/... |
| 271 | 200 | HTTP | Tunnel to | www.google-analytics.co... | 0 | | |
| 272 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 273 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 274 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 275 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 276 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 277 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 278 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 279 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 280 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 281 | 200 | HTTPS | www.google-analyti... | /analytics.js | 17,543 | public, ... | text/javasc... |
| 282 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 283 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 284 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 285 | 200 | HTTPS | images.unsplash.com | /profile-1543004110636-e... | 4,207 | public,... | image/jpeg |
| 286 | 200 | HTTPS | images.unsplash.com | /photo-1517594422361-5... | 60,856 | public,... | image/jpeg |
| 287 | 200 | HTTPS | images.unsplash.com | /profile-1510932839800-f... | 4,240 | public,... | image/jpeg |
| 288 | 200 | HTTPS | images.unsplash.com | /profile-1484718924463-2... | 4,050 | public,... | image/jpeg |
| 289 | 200 | HTTPS | images.unsplash.com | /photo-1543005472-1b1d... | 197,102 | public,... | image/jpeg |
| 290 | 200 | HTTPS | images.unsplash.com | /photo-1496483353456-9... | 267,260 | public,... | image/jpeg |
| 291 | 200 | HTTPS | images.unsplash.com | /profile-155144932987-d... | 4,054 | public,... | image/jpeg |
| 292 | 200 | HTTPS | images.unsplash.com | /profile-1496509391479-7... | 2,326 | public,... | image/png |
| 293 | 200 | HTTPS | images.unsplash.com | /profile-1523277534788-b... | 3,949 | public,... | image/jpeg |
| 294 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 295 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 296 | 200 | HTTPS | images.unsplash.com | /profile-fb-1542840351-f0... | 4,200 | public,... | image/jpeg |
| 297 | 200 | HTTPS | unsplash.com | /a/search-route.a8060.js | 4,565 | max-ag... | application/... |
| 298 | 200 | HTTPS | unsplash.com | /a/vendors-collection-rou... | 1,000 | max-ag... | application/... |
| 299 | 200 | HTTPS | images.unsplash.com | /profile-1506580101242-d... | 4,175 | public,... | image/jpeg |
| 300 | 200 | HTTPS | unsplash.com | /a/collection-route-editori... | 11,465 | max-ag... | application/... |

Request Headers

GET /search/photos/fire HTTP/1.1

Client

Accept: text/html, application/xhtml+xml, image/jxr, */*

Accept-Encoding: gzip, deflate

Accept-Language: zh-CN

User-Agent: Mozilla/5.0 (Windows NT 10.0; WOW64; Trident/7.0; rv:11.0) like Gecko

Cookies

Cookie

_ga=GA1.2.357469433.1552980642

_gid=GA1.2.981967733.1554771481

_sp_id.0295=bd876883-c561-42da-b634-67bee592170a.1552980642.10.1554778562.1554772952.1c084b1a-b4c1-4db3-8080-046dcd2fcd9f

Transformer Headers TextView SyntaxView ImageView HexView WebView Auth Caching Cookies Raw JSON XML

```
};
LUX.label = 'Search-Photos';
LUX.sampleRate = 0;
LUX.auto = false;
</script><script src="https://cdn.speedcurve.com/js/lux.js?id=140493345" async="" defer=""></script><link rel="preload"
as="script" href="/a/search-route.a8060.js"><link rel="preload" as="script" href="/a/vendors-collection-
route-editorial-route-following-route-keyword-landing-page-route-photos-route-sea-ef460f74.ed168.js"><link rel="
preload" as="script" href="/a/collection-route-editorial-route-following-route-keyword-landing-page-photos-
route-search-phot-f2df1778.7c293.js"><link rel="preload" as="script" href="/a/search-photos-route.78453.js"><link
rel="preload" as="script" href="/a/vendors-main.acfad.js"><link rel="preload" as="script" href="/a/main.82291.js"><
link rel="prefetch" href="/a/photos-route.86735.css"><link rel="prefetch" as="script" href="/a/photos-route.86735.js"/
><script>window._TRACING_ = {<script><link data-react-helmet="true" rel="apple-touch-icon" sizes="180x180" href="
https://unsplash.com/apple-touch-使用查询选择器或简单文本在 DOM 中搜索"="true" rel="icon" type="image/png" sizes="32x32"
href="https://unsplash.com/favicon-32x32.png"><link data-react-helmet="true" rel="icon" type="image/png" sizes="16x16"
href="https://unsplash.com/favicon-16x16.png"><link data-react-helmet="true" rel="mask-icon" href="https://unsplash.
com/safari-pinned-tab.svg" color="#000000"><link data-react-helmet="true" rel="manifest" href="/site-v2.webmanifest"/>
<link data-react-helmet="true" rel="canonical" href="https://unsplash.com/search/photos/fire"><meta data-react-helmet=
"true" name="charset" content="UTF8"><meta data-react-helmet="true" name="viewport" content="width=device-width,
initial-scale=1.0, maximum-scale=1.0, minimal-ui"><meta data-react-helmet="true" name="mobile-web-app-capable"
content="yes"><meta data-react-helmet="true" name="apple-mobile-web-app-capable" content="yes"><meta data-react-
helmet="true" name="apple-mobile-web-app-title" content="Unsplash"><meta data-react-helmet="true" name="application-
name" content="Unsplash"><meta data-react-helmet="true" name="author" content="Unsplash"><meta data-react-helmet="
true" name="msapplication-config" content="browserconfig.xml"><meta data-react-helmet="true" name="msapplication-
TitleColor" content="ffffff"><meta data-react-helmet="true" name="msapplication-TileImage" content="https://unsplash.
com/ms144x144.png"><meta data-react-helmet="true" name="theme-color" content="ffffff"><meta data-react-helmet="
true" http-equiv="Accept-CH" content=" DPR, Width, Viewport-Width"><meta data-react-helmet="true" name="description"
```


11.3 爬取瀑布流图片

➤ Flidder

Progress Telerik Fiddler Web Debugger

File Edit Rules Tools View Help

WinConfig Replay X Go Stream Decode Keep: All sessions Any Process Find Save Browse Clear Cache TextWizard Tearoff MSDN Search... Online

| # | Result | Protocol | Host | URL | Body | Caching | Content-Type |
|-----|--------|----------|------------------------|-----------------------------|---------|-----------|-----------------|
| 304 | 200 | HTTPS | unsplash.com | /a/vendors-main.acfad.js | 220,799 | max-ag... | application/... |
| 305 | 200 | HTTPS | cdn.speedcurve.com | /js/fux.js?id=140493345 | 5,509 | max-ag... | application/... |
| 306 | 200 | HTTPS | unsplash.com | /a/main.82291.js | 103,375 | max-ag... | application/... |
| 307 | 200 | HTTP | Tunnel to | logger.unsplash.com:443 | 0 | | |
| 308 | 200 | HTTP | Tunnel to | www.google-analyt... | 0 | | |
| 309 | 302 | HTTPS | www.google-analyt... | /r/collect?v=1&v=j738a... | 416 | no-cac... | text/html; c... |
| 310 | 200 | HTTP | Tunnel to | stats.g.doubleclick... | 0 | | |
| 311 | 200 | HTTPS | stats.g.doubleclick... | /r/collect?v=1&v=j738a... | 35 | no-cac... | image/gif |
| 312 | 200 | HTTP | Tunnel to | secure.insightexpressai... | 0 | | |
| 313 | 200 | HTTP | Tunnel to | secure.insightexpressai... | 0 | | |
| 314 | 200 | HTTP | Tunnel to | secure.insightexpressai... | 0 | | |
| 315 | 200 | HTTPS | images.unsplash.com | /photo-1505017791108-7... | 331,679 | public... | image/jpeg |
| 316 | 200 | HTTPS | images.unsplash.com | /photo-1546182208-1e70... | 115,469 | public... | image/jpeg |
| 317 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 318 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 319 | 200 | HTTP | Tunnel to | unsplash.com:443 | 0 | | |
| 320 | 200 | HTTP | Tunnel to | unsplash.com:443 | 0 | | |
| 321 | 200 | HTTP | Tunnel to | unsplash.com:443 | 0 | | |
| 322 | 200 | HTTPS | secure.insightpre... | /adServer/adServerESL.a... | 35 | max-ag... | image/gif |
| 323 | 200 | HTTPS | secure.insightpre... | /adServer/adServerESL.a... | 35 | max-ag... | image/gif |
| 324 | 200 | HTTPS | secure.insightpre... | /adServer/adServerESL.a... | 35 | max-ag... | image/gif |
| 325 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 326 | 200 | HTTP | Tunnel to | images.unsplash.com:443 | 0 | | |
| 327 | 200 | HTTP | Tunnel to | unsplash.com:443 | 0 | | |
| 328 | 200 | HTTPS | unsplash.com | /api/search/photos?quer... | 9,125 | no-cac... | application/... |
| 329 | 200 | HTTPS | images.unsplash.com | /photo-1504470695779-7... | 177,331 | public... | image/jpeg |
| 330 | 200 | HTTPS | images.unsplash.com | /photo-1497906539264-e... | 323,249 | public... | image/jpeg |
| 331 | 200 | HTTPS | unsplash.com | /favicon-32x32.png | 144 | public... | image/png |
| 332 | 200 | HTTPS | unsplash.com | /a/photos-route.86735.js | 20,277 | max-ag... | application/... |
| 333 | 200 | HTTPS | images.unsplash.com | /photo-1495467033336-2... | 143,744 | public... | image/jpeg |
| 334 | 200 | HTTPS | images.unsplash.com | /photo-1510772314292-9... | 21,267 | public... | image/jpeg |
| 335 | 200 | HTTPS | unsplash.com | /a/photos-route.86735.css | 2,210 | max-ag... | text/css |
| 336 | 200 | HTTPS | logger.unsplash.com | /r?stm=1554787535627b... | 43 | | image/gif |
| 337 | 200 | HTTPS | logger.unsplash.com | /r?stm=1554787540971b... | 43 | | image/gif |
| 338 | 200 | HTTPS | logger.unsplash.com | /r?stm=1554787545638b... | 43 | | image/gif |
| 339 | 200 | HTTP | Tunnel to | v10.events.data.microsof... | 0 | | |
| 340 | 200 | HTTP | tgccfg.qq.com | /tas/pattern/5600200040... | 20,100 | max-ag... | application/... |
| 341 | 200 | HTTP | Tunnel to | logger.unsplash.com:443 | 0 | | |
| 342 | 200 | HTTPS | logger.unsplash.com | /r?stm=1554787735642b... | 43 | | image/gif |

Request Headers

GET /api/search/photos?query=fire&per_page=20&page=2 HTTP/1.1

Client

Accept: */*
Accept-Encoding: gzip, deflate
Accept-Language: zh-CN
User-Agent: Mozilla/5.0 (Windows NT 10.0; WOW64; Trident/7.0; rv:11.0) like Gecko

Cookies

_ga=GA1.2.357469433.1552980642
_gat=1
_gid=GA1.2.981967733.1554771481

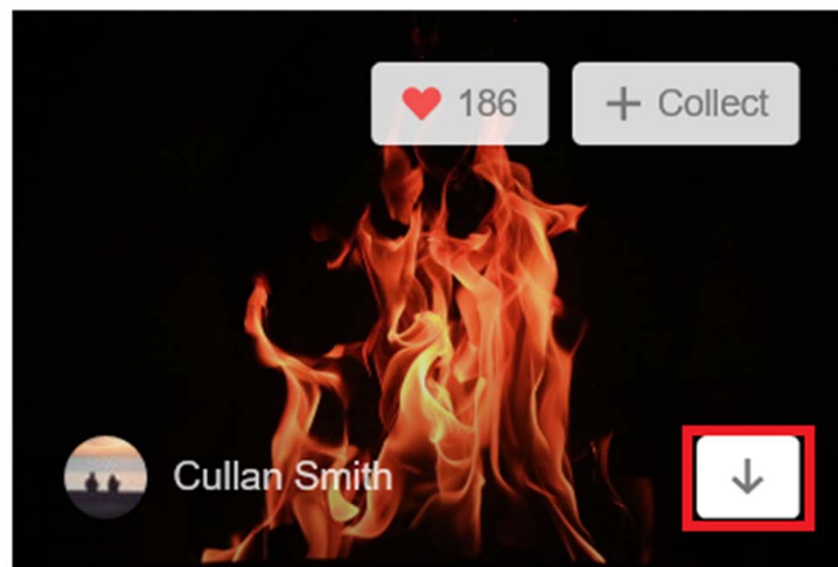
Transformer Headers TextView SyntaxView ImageView HexView Webview Auth Caching Cookies Raw JSON XML

height=1536
id=1UAI5_PQg_E
liked_by_user=False
likes=68
links
download=https://unsplash.com/photos/1UAI5_PQg_E/download
download_location=https://api.unsplash.com/photos/1UAI5_PQg_E/download
html=https://unsplash.com/photos/1UAI5_PQg_E
self=https://api.unsplash.com/photos/1UAI5_PQg_E
photo_tags
sponsored=False
sponsored_by=(null)
sponsored_impressions_id=(null)
tags
updated_at=2019-03-27T18:07:38-04:00
urls
full=https://images.unsplash.com/photo-1518229351380-11aef3ffc358?xlib=rb-1.2.1&q=85&fm=jpg&crop=entropy&cs=tiny&rgb&w=1080&f=max&id=eyJhZDhfaWQOeYMDd9
raw=https://images.unsplash.com/photo-1518229351380-11aef3ffc358?xlib=rb-1.2.1&id=eyJhZDhfaWQOeYMDd9
regular=https://images.unsplash.com/photo-1518229351380-11aef3ffc358?xlib=rb-1.2.1&q=80&fm=jpg&crop=entropy&cs=tiny&rgb&w=1080&f=max&id=eyJhZDhfaWQOeYMDd9
small=https://images.unsplash.com/photo-1518229351380-11aef3ffc358?xlib=rb-1.2.1&q=80&fm=jpg&crop=entropy&cs=tiny&rgb&w=400&f=max&id=eyJhZDhfaWQOeYMDd9
thumb=https://images.unsplash.com/photo-1518229351380-11aef3ffc358?xlib=rb-1.2.1&q=80&fm=jpg&crop=entropy&cs=tiny&rgb&w=200&f=max&id=eyJhZDhfaWQOeYMDd9
user
accepted_tos=True
bio=View Images curated by Cullan

Expand All Collapse JSON parsing completed.

11.3 爬取瀑布流图片

➤JSON数据



<https://unsplash.com/photos/BdTtvBRh0ng/download?force=true>
<https://unsplash.com/photos/x3S1aGQNgro/download?force=true>

```
import requests
if __name__ == '__main__':
    target = 'http://unsplash.com/napi/search'
    req = requests.get(url=target, verify=False)
    print(req.text)
```

11.3 爬取瀑布流图片

➤ JSON数据

File Edit Format Run Options Window Help

```
import requests
if __name__ == '__main__':
    target = 'http://unsplash.com/napi/search/photos?query=fire&xp=&per_page=20'
    req = requests.get(url=target, verify=False)
    print(req.text)
```

File Edit Shell Debug Options Window Help

```
st/advanced-usage.html#ssl-warnings
{"total":23947,"total_pages":1198,"results":[{"id":"1UAI5_PQg_E","created_at":"2018-02-09T21:23:45-05:00","updated_at":"2019-03-27T18:07:38-04:00","width":2304,"height":1536,"color":"#FEF393","description":"Burning Cold","alt_description":"fire illustration","urls":{"raw":"https://images.unsplash.com/photo-1518229351380-11aef3ffc358?ixlib=rb-1.2.1&ixid=eyJhcnBfaWQiOjE5MDd9","full":"https://images.unsplash.com/photo-1518229351380-11aef3ffc358?ixlib=rb-1.2.1&q=85&fm=jpg&crop=entropy&cs=srgb&ixid=eyJhcnBfaWQiOjE5MDd9","regular":"https://images.unsplash.com/photo-1518229351380-11aef3ffc358?ixlib=rb-1.2.1&q=80&fm=jpg&crop=entropy&cs=tiny&w=1080&fit=max&ixid=eyJhcnBfaWQiOjE5MDd9","small":"https://images.unsplash.com/photo-1518229351380-11aef3ffc358?ixlib=rb-1.2.1&q=80&fm=jpg&crop=entropy&cs=tinysrgb&w=400&fit=max&ixid=eyJhcnBfaWQiOjE5MDd9","thumb":"https://images.unsplash.com/photo-1518229351380-11aef3ffc358?ixlib=rb-1.2.1&q=80&fm=jpg&crop=entropy&cs=tinysrgb&w=200&fit=max&ixid=eyJhcnBfaWQiOjE5MDd9"},"links":{"self":"https://api.unsplash.com/photos/1UAI5_PQg_E","html":"https://unsplash.com/photos/1UAI5_PQg_E","download":"https://unsplash.com/photos/1UAI5_PQg_E/download","download_location":"https://api.unsplash.com/photos/1UAI5_PQg_E/download"},"categories":[],"sponsored":false,"sponsored_by":null,"sponsored_impressions_id":null,"likes":68,"liked_by_user":false,"current_user_collections":[],"user":{"id":"Edg-hHn0oT4","updated_at":"2019-03-04T21:01:04-05:00","username":"cullansmith","name":"Cullan Smith","first_name":"Cullan","last_name":"Smith","twitter_username":"CullanSmithYT","portfolio_url":"https://twitter.com/CullanSmithYT?s=09","bio":"View Images curated by Cullan","location":null,"links":{"self":"https://api.unsplash.com/users/cullansmith","html":"https://unsplash.com/@cullansmith","photos":"https://api
```

```
import requests
if __name__ == '__main__':
    target = 'http://unsplash.com/napi/search'
    req = requests.get(url=target, verify=False)
    print(req.text)
```

11.3 爬取瀑布流图片

➤ JSON数据

```
File Edit Format Run Option: File Edit Shell Debug Options Window Help
import requests, json
if __name__ == '__main__':
    target = 'http://unsplash.com/napi/search'
    req = requests.get(url=target)
    html = json.loads(req.text)
    for each in html['results']:
        print('图片ID:', each['id'])
```

Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: C:/Users/dell/Desktop/课用/代码/抓包2.py =====
=====
Warning (from warnings module):
File "C:/Users/dell/AppData/Local/Programs/Python/Python37/lib/site-packages/urllib3/connectionpool.py", line 847
InsecureRequestWarning)
InsecureRequestWarning: Unverified HTTPS request is being made. Adding certificate verification is strongly advised. See: https://urllib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warnings
图片ID: 1UAI5_PQg_E
图片ID: t8T_yUgCKSM
图片ID: aTbrv4d8wIw
图片ID: Igct8iZucFI
图片ID: eHAQRwMrWDE
图片ID: icrhAD-qidc
图片ID: rWotMddrvUM
图片ID: AtPbmIH97mA
图片ID: AhUqPgCEMhA
图片ID: abkEA0jnY0s
图片ID: qqEpMbG1Kyg
图片ID: XGMUmv_KoH8
图片ID: W_XpzHlPVqM
图片ID: uNs17UOr_Ec
图片ID: UKX_DwNKXSA
图片ID: VzLlMT6ZYn0
图片ID: rRg049i8w2s
图片ID: 7IlaJn7GTFE
图片ID: 00LAAOW-u8s
图片ID: -98jVaVuGv0

11.3 爬取瀑布流图片

➤ 程序整合

imagedownload.py