

Python爬虫实战-今日头条热点新闻

开发工具：Pycharm Professional

这是目标网页，进入热点，可以看到新闻内容



首先试试能不能获取网页源代码

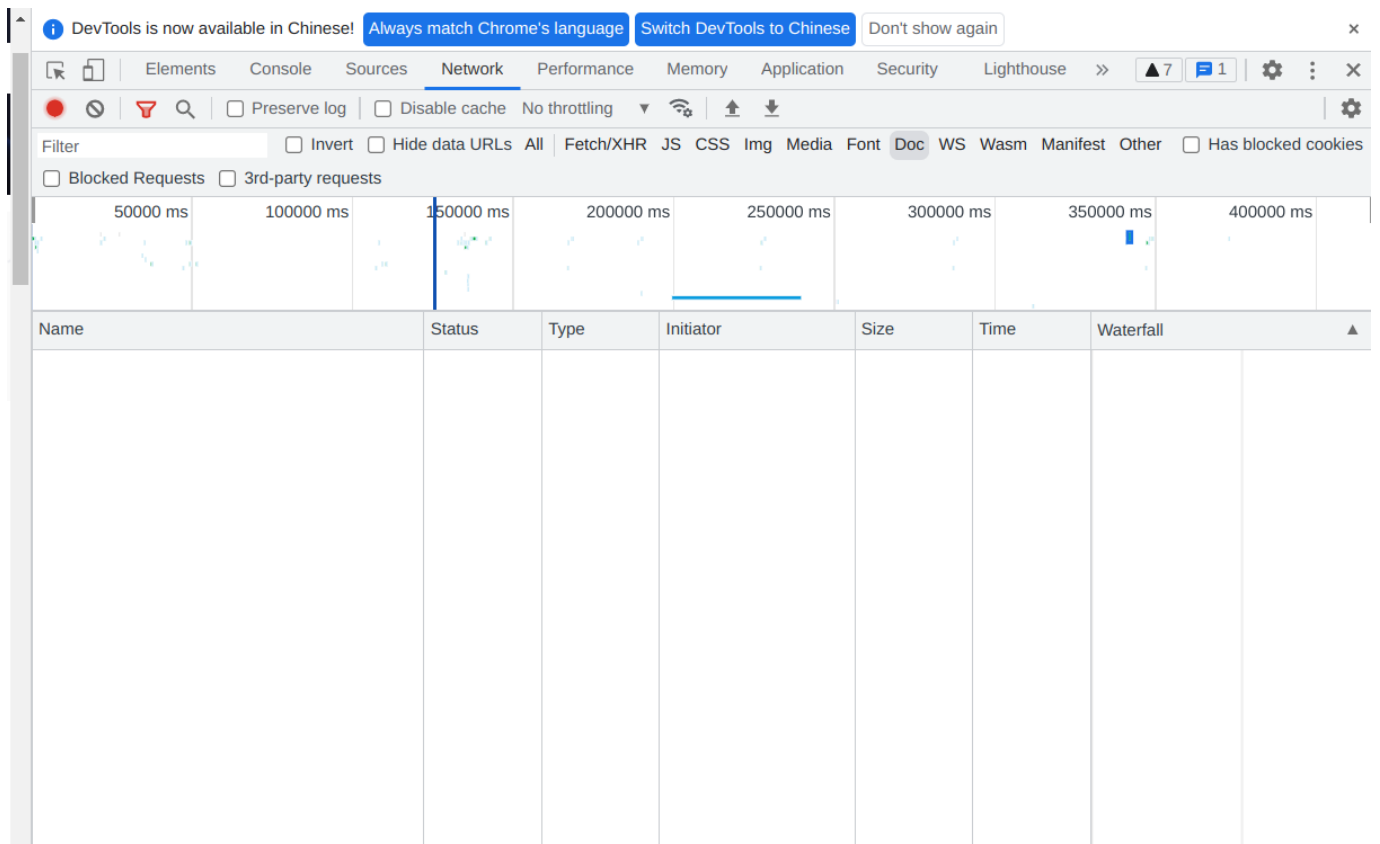
```
import requests

headers = {"user-agent": "Mozilla/5.0 (X11; Linux x86_64)"}
url = "https://www.toutiao.com/"
res = requests.get(url=url, headers=headers)
print(res.status_code)
print(res.text)
```

可以看到，网站返回的是js代码，并非HTML网页源代码，真实的源代码不可能这么少

```
/usr/bin/python3.7 /home/hwx/PycharmProjects/work/main.py
200
<html><head><meta charset="UTF-8" /></head><body></body><script src='https://lf3-cdn-tos.bytescm.com/obj/rc-web-sdk/acrawler.js'></script><script>function _f1(e,t){if("string"!
Process finished with exit code 0
```

查看网页源代码后，发现Doc这一栏是空的，因为对方服务器并未返回静态文本

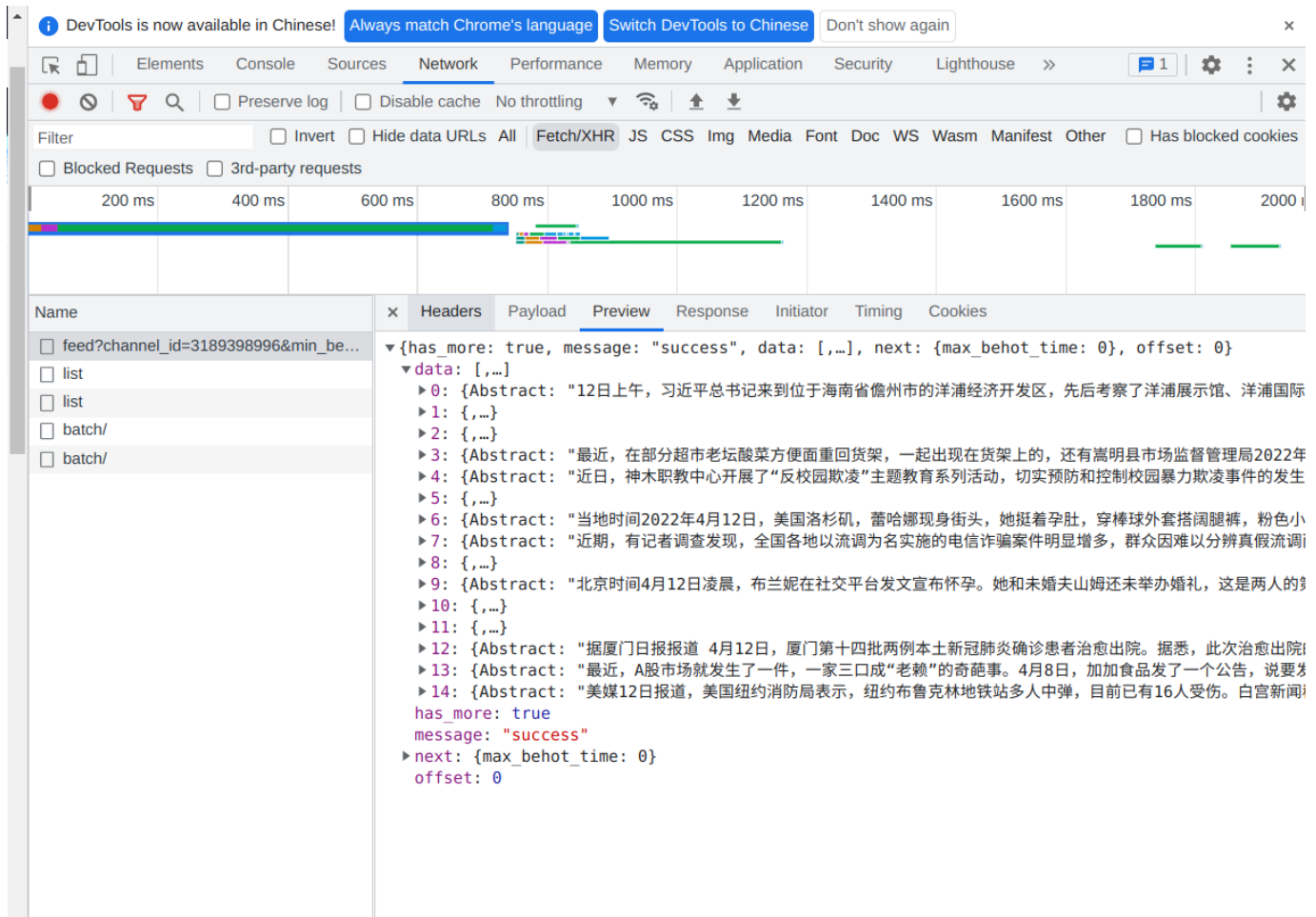


这说明网页是动态的，使用了Ajax

什么是Ajax？

全程Asynchronous JavaScript and XML，即异步的JavaScript和XML。利用了JavaScript保证页面不被刷新、页面链接不变的情况下与服务器交换数据并更新部分网页内容。

Ajax有其特殊的请求类型，叫作XHR。打开检查，点击XHR，找到一个链接，单击Preview，就能看到响应的内容，这些内容是JSON格式的，这里浏览器自动作了解析，可以发现这其中就是要爬取的内容，15条热点询问按顺序排列



点击Payload就能看到传递的参数



代码如下，获取到的数据要用json解析

```
import requests

headers = {"user-agent": "Mozilla/5.0 (X11; Linux x86_64)"}
params = {
    "channel_id": "3189398996",
    "min_behot_time": "1649817465",
    "refresh_count": "3",
    "category": "pc_profile_channel",
    "aid": "24",
    "app_name": "toutiao_web",
    "_signature":
    "_02B4Z6wo00101801HZAAIDAqGitE9CKQX.NARkAAJEoW3dqKEHE9AEoz9seBI77-
w0Fju08tqa2VMt.IDatj.5l6I2dds0nN5iw6nT.dbeeagQG1VchRFsUNsrTiMyiaJn0gERnkTx-
Jsa154"
}
url = "https://www.toutiao.com/api/pc/list/feed"
res = requests.get(url=url, headers=headers, params=params).json()
print(type(res))
print(res)
```

如下就获取到了数据

```
/usr/bin/python3.7 /home/hwx/PycharmProjects/work/main.py
<class 'dict'>
{'has_more': True, 'message': 'success', 'data': [{'Abstract': '12日上午，习近平总书记来到位于海南省儋州市的洋浦经济开发区，先后考察了洋浦展示馆、洋浦国际集装箱码头小铲滩港区，了解洋浦经济开发区发展和中国特色自由贸易港建设等情况。'}], 'next': {'max_behot_time': 0}, 'offset': 0}
Process finished with exit code 0
```

在浏览器中查看json数据，发现新闻的简介在Abstract字段下

```
▼ {has_more: true, message: "success", data: [...], next: {max_behot_time: 0}, offset: 0}
  ▼ data: [...],
    ▼ 0: {Abstract: "12日上午，习近平总书记来到位于海南省儋州市的洋浦经济开发区，先后考察了洋浦展示馆、洋浦国际集装箱码头小铲滩港区，了解洋浦经济开发区发展和中国特色自由贸易港建设等情况。"}
      Abstract: "12日上午，习近平总书记来到位于海南省儋州市的洋浦经济开发区，先后考察了洋浦展示馆、洋浦国际集装箱码头小铲滩港区，了解洋浦经济开发区发展和中国特色自由贸易港建设等情况。"
      action_extra: {"channel_id": 3189398996}
      ▼ action_list: [{action: 1}, {action: 3}, {action: 7}, {action: 9}]
        ▶ 0: {action: 1}
        ▶ 1: {action: 3}
        ▶ 2: {action: 7}
        ▶ 3: {action: 9}
      aggr_type: 1
      article_alt_url: ""
      article_sub_type: 0
      article_type: 0
      article_url: "https://toutiao.com/group/7085883390372790818/"
      article_version: 0
      ban_bury: 1
      ban_comment: false
      ban_danmaku: true
      behot_time: 1649818060
      bury_count: 0
      bury_style_show: 0
      cell_flag: 0
      cell_layout_style: 1
      cell_type: 0
      city: ""
      comment_count: 1451
      cursor: 1649818060999
      digg_count: 2595
      diqq_icon_key: ""
```

链接在article_url下，标题在title字段下，平台名称在media__name字段下

```
    feed_title: "习近平在海南洋浦经济开发区考察调研"
  ▶ filter_words: {action_extra: {"channel_id": 3189398996},...}
  ▶ forward_info: {forward_count: 101}
    group_flags: 0
    group_id: "7085883390372790818"
    group_source: 2
    group_type: 0
    has_image: false
    has_m3u8_video: false
    has_mp4_video: false
    has_video: false
    hot: 0
    image_type: "none"
    info_desc: ""
    insert_ads: ""
    is_stick: true
    item_id: "7085883390372790818"
    item_version: 0
    keywords: ""
    label: "置顶"
  ▶ label_extra: {icon_url: {}, is_redirect: false, redirect_url: "", style_type: 0}
    label_style: 1
    level: 0
    like_count: 2595
  ▶ log_pb: {author_id: "4492956276", group_source: "2", impr_id: "202204131047400101510732020575C2A9}
  ▶ media_info: {...}
    media name: "央视新闻"
```

于是可以很简单取出数据

```

import requests

headers = {"user-agent": "Mozilla/5.0 (X11; Linux x86_64)"}
params = {
    "channel_id": "3189398996",
    "min_behot_time": "1649822510",
    "refresh_count": "3",
    "category": "pc_profile_channel",
    "aid": "24",
    "app_name": "toutiao_web",
    "_signature":
    "_02B4Z6wo00101lpfs4gAAIDDycxytizPE35ae7cAAPTQfdPdBniixGRckbmmoCyxFb.vpLjZK10lK
mIWDx87Kf.MWg4wUeXaPq.eULBr1BPj8d39ecjDuxkHRC8gSLXtTfnmUMSiNiub0x9rc1"
}
url = "https://www.toutiao.com/api/pc/list/feed"
res = requests.get(url=url, headers=headers, params=params)
items = res.json()["data"]

f = open("result.txt", "w+")
for item in items:
    title = item["title"]
    link = item["article_url"]
    abstract = item["Abstract"]
    media = item["media_name"]
    f.write(f"{title}\t{link}\t{media}\t{abstract}\n")
    print(title)
f.close()

```

结果如下

1	独家视频！习近平在海南洋浦经济开发区考察调研	https://toutiao.com/group/7085883390372790818/	央视新闻	12日上午，习近平总书记来到位于海南省儋州市的	✓
2	习近平海南行！走进洋浦经济开发区	https://toutiao.com/group/7085887253330903567/	央视新闻	△视频 走进洋浦经济开发区2022年4月12日 习近平总书记来到	
3	坚持“动态清零”总方针不犹豫不动摇	http://m2.people.cn/news/toutiao.html?s=MV8zXzE1NTMyNDc0XzQwODFfMTY0OTgwMjYyMA==	人民网	3月17日，习近	
4	防控奥密克戎必须戴N95口罩？专家辟谣：熔喷布口罩就有效	http://m.q11d.com/new/general/18942003	齐鲁壹点	近日，有传言称：新冠病毒变异后，医用口罩对奥密克	
5	辅警夜查酒驾致男子坠河溺亡，法院判决交警执法行为违法	http://jmwap.ctdsb.net/jmshare/news/detail_index.html?contentType=5&contentId=1396258			
6	美国纽约市警方锁定布鲁克林地铁枪击案嫌疑人	http://ex.chinadaily.com.cn/exchange/partners/77/rss/channel/cn/columns/32tlvc/stories/WS625626			
7	稻田里满眼绿！横州市葛汶村以机械化生产助力乡村振兴	http://www.gxnews.com.cn/staticpages/20220412/newgx6254e571-20713342.shtml	广西新闻网		
8	法国已检出新冠XD和XE感染者 正密切监测多种毒株	https://m.haiwainet.cn/ttc/3541926/2022/0413/content_32388911_1.html	海外网	据《巴黎人报》	
9	习近平在海南洋浦经济开发区考察调研	http://my-h5news.app.xinhuanet.com/h5/article.html?articleId=8037ea38735aa20f1107ab2368495084	新华网		
10	新手机刚买就降了580元，退差价还要被商家扣124元赠品钱……	https://toutiao.com/group/7085923439260779049/	海曙检察	人生最心塞的事之一刚买完就降价还	
11	韩国：国际婚姻家庭中小学在校生近9年增240.8%	https://m.jiemian.com/article/7326995_ttkx.html	界面快讯	韩联社报道，据韩国女性家庭部和教育部4月13日	
12	小辣椒 大产业	https://toutiao.com/group/7085914355715293735/	阿拉尔发布	人间四月天，春种正当时。眼下，在第一师阿拉尔市八团塔门镇，多台农用机械正	
13	拒不主动申报旅居史！昆明警方对阳性人员艾某某开展调查	https://toutiao.com/group/7085916171198792207/	北京日报客户端	4月13日，记者从昆明市公安局获悉，	
14	海关总署：一季度我国进出口增长10.7%，外贸开局平稳	http://m.q11d.com/new/general/18941811	齐鲁壹点	今天（13日）上午，国务院新闻办召开新闻发布会，海	

须要注意的是，这里的传参有一定时效性，一旦过期则无法获取到数据