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

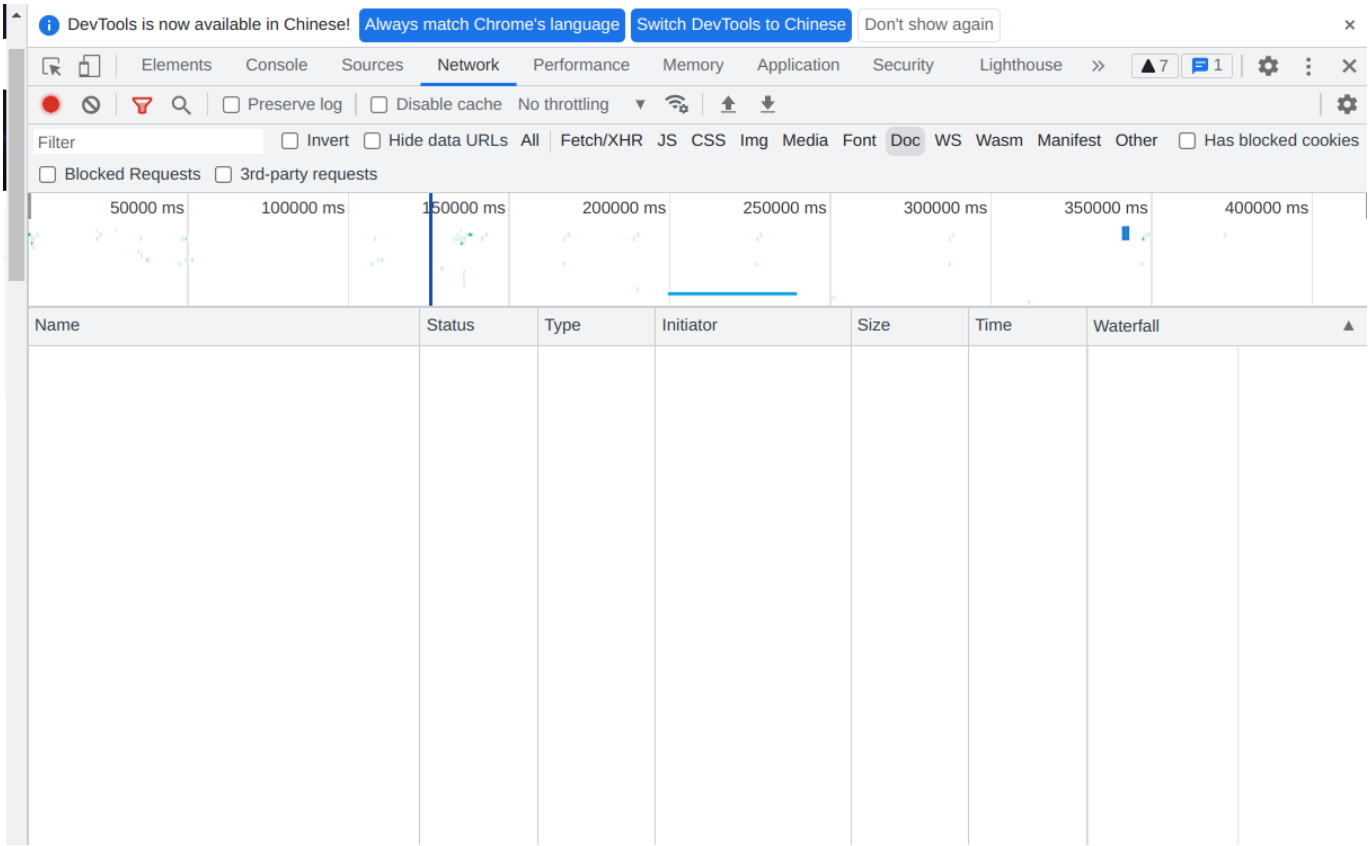
首先试试能不能获取网页源代码

```
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

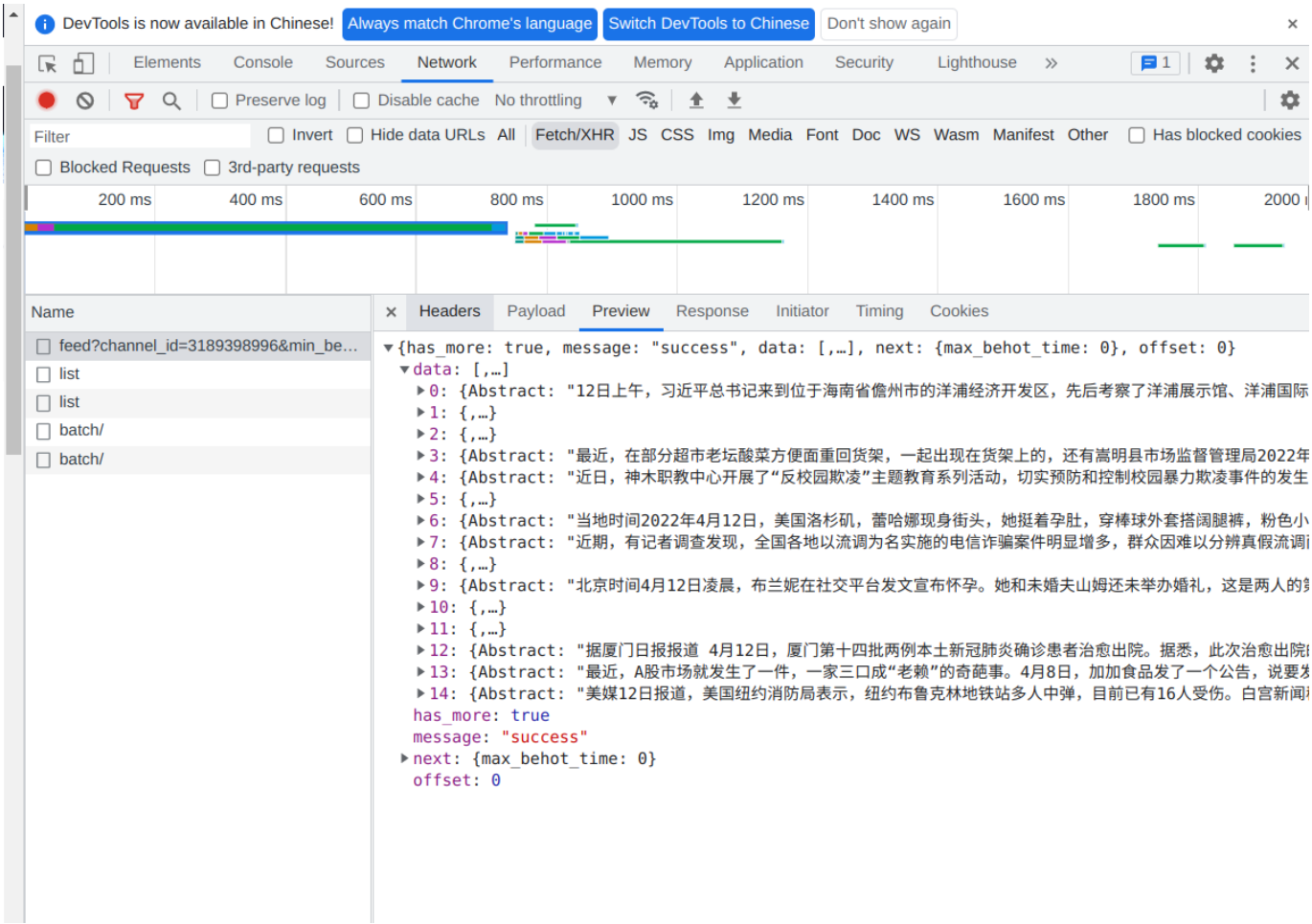


这说明网页是动态的，使用了Ajax

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

```python
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":
"_02B4Z6wo0010180lHZAAAIDAqGitE9CKQX.NARkAAJEoW3dqKEHE9AEoz9seBI77-
wOFjuO8tqa2VMt.IDatj.5l6I2ddsOnN5iw6nT.dbeeagQGlVchRFsUNsrTiMyiaJnOgERnkTx-
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

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

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: "2022041310474001015107320205750 2A9
▶ 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.vpLjZK1OlK
mIWDx87Kf.MWg4wUeXaPq.eULBrlBPj8d39ecjDuxkHRc8gSlXtTfnmUMSiNiubOx9rc1"
}
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

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