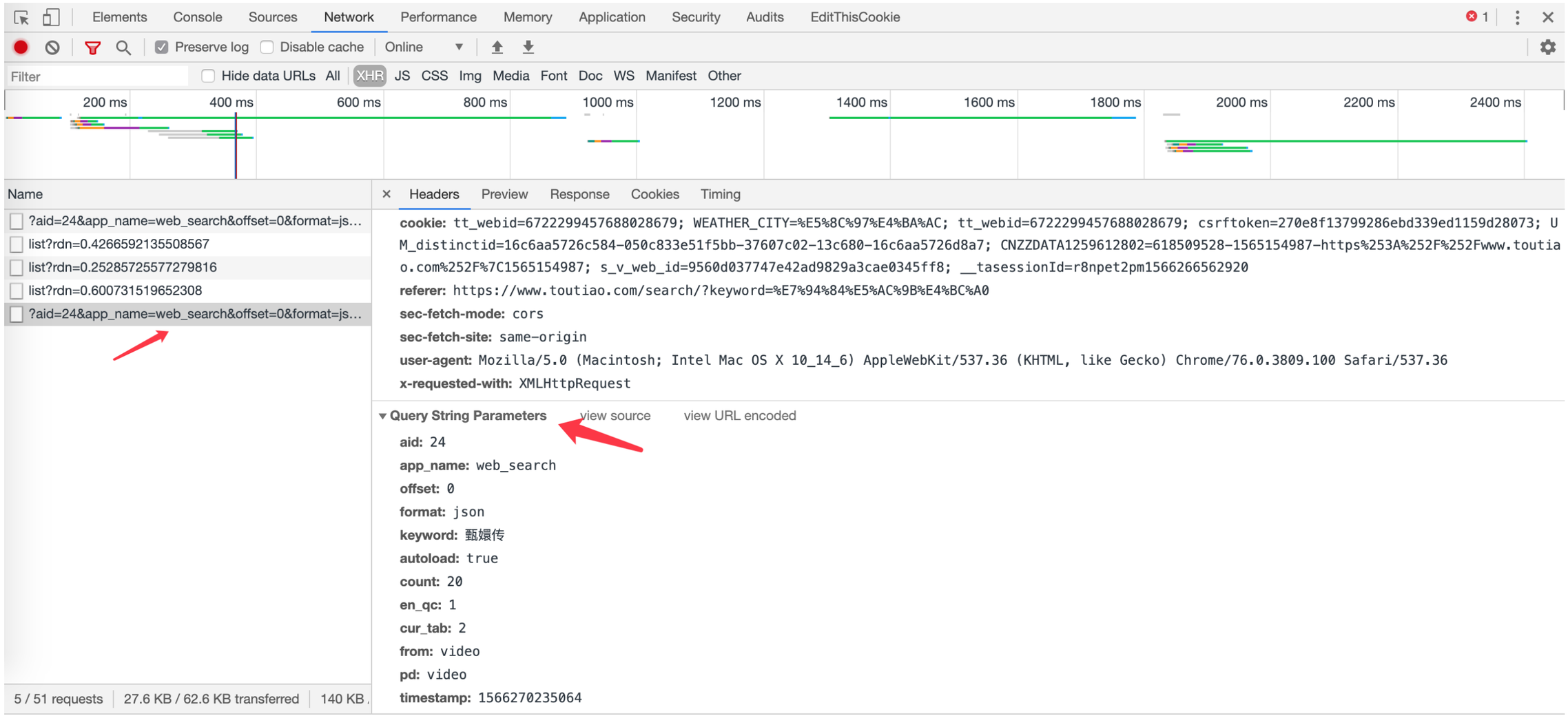
爬取今日头条Ajax新闻数据

本例在今日头条网页端中爬取指定的短视频．通过浏览器头条视频中搜索电视剧甄嬛传片段，可以得到以下内容：

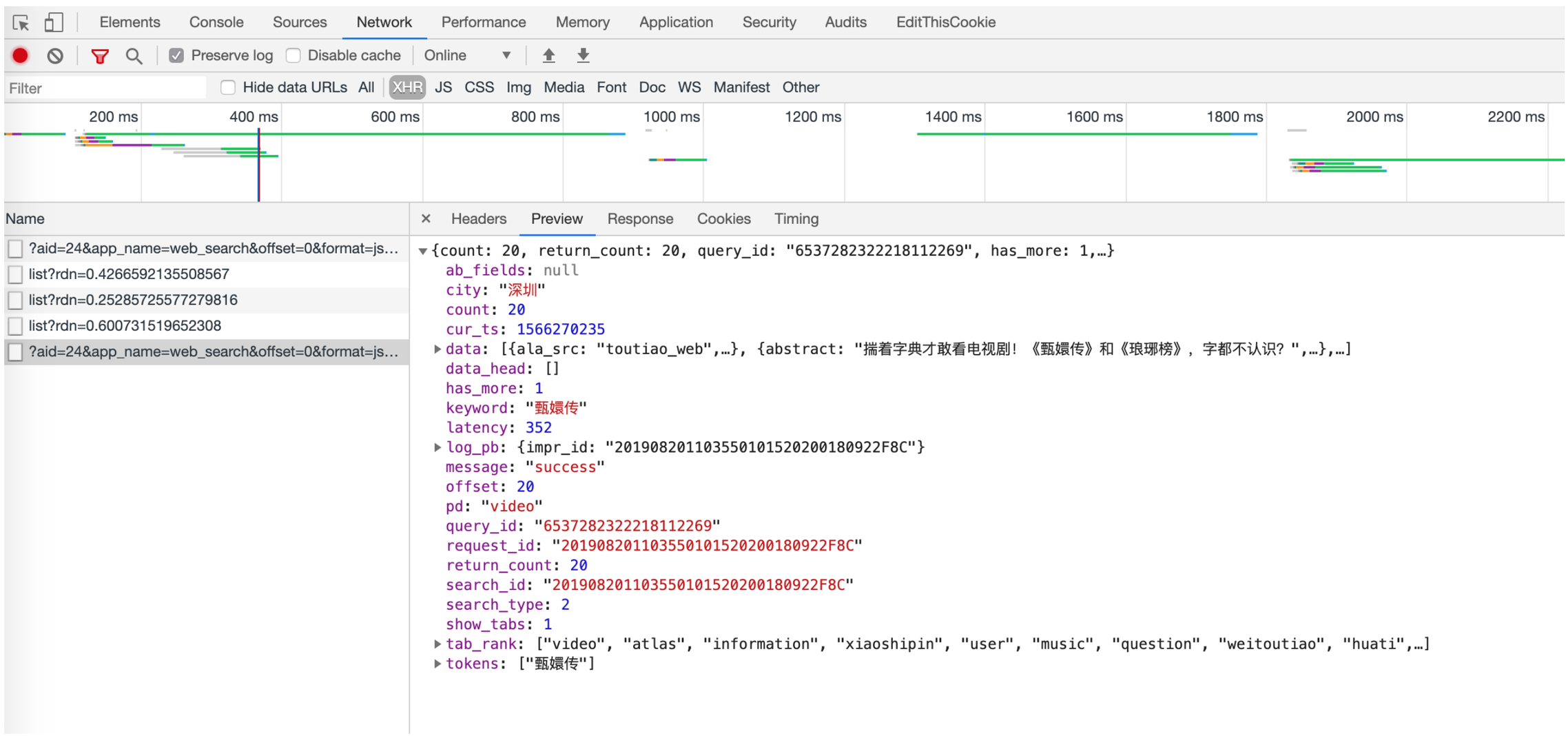
通过查看网页源代码，我们发现在页面的源代码中并没有短视频的播放链接等信息，如下：

```
src="//s3.pstatp.com/toutiao/resource/ntoutiao_web/static/js/common/lib_d2cb17c.js" crossorigin="anonymous"></script><link rel="stylesheet"
href="//s3b.pstatp.com/toutiao/resource/ntoutiao_web/page/search/search_840cf4b.css" /></head><body style="min-width: 768px; overflow-x: hidden;"><a
href="/articles_news_society/" style="display:none;"></a><div class="y-wrap"><div id="header"><div riot-tag="wtopbar" style="height: 34px; background: #222;"></div></div><div
riot-tag="login"></div><div id="toast"></div><div class="y-box container"><div class="y-left index-left"><a class="logo-wrap" href="//www.toutiao.com/"><img
src="//s3.pstatp.com/toutiao/resource/ntoutiao_web/static/image/logo_271e845.png" alt=""></a></div><div class="y-left index-middle"><div riot-tag="searchBar"></div><div riot-
tag="tabBar"></div><div riot-tag="feedBox"></div></div><div class="y-right index-right"></div></div><div riot-tag="verify"></div><div id="footer"><div><span class="fitem">©
2019 今日头条 </span><a class="fitem" href="http://www.12377.cn/" target="_blank">中国互联网举报中心</a><a class="fitem" href="http://www.miibeian.gov.cn/" target="_blank">京ICP证
140141号</a><a class="fitem" href="http://www.miibeian.gov.cn/" target="_blank">京ICP备12025439号-3</a><a class="fitem" href="http://www.beian.gov.cn/portal/registerSystemInfo?
recordcode=11000002002023" target="_blank" ga_event="click_about"><img src="//s3.pstatp.com/toutiao/resource/ntoutiao_web/static/image/gongan_d0289dc.png">京公网安备
11000002002023号
        </a></div><div><a class="fitem" href="/license/" class="icp" target="_blank">网络文化经营许可证</a><a class="fitem" href="/a3642705768/" target="_blank">跟帖评论自律管理承诺
书</a><span class="fitem">违法和不良信息举报电话：400-140-2108</span><span class="fitem">公司名称：北京字节跳动科技有限公司</span></div></div></div><script>
        window.globals = window.globals || {};

        window.taAnalysis && taAnalysis.setup({
          c: '/search/'
        });
        window.taAnalysis && taAnalysis.send('pageview', {});

        utils.on(document.getElementsByTagName('body')[0], 'click', function(e) {
            var target = utils.getTarget(e),
                ga_event,
                ga_category,
                ga_label,
                ga_value;

            while(target && target.nodeName.toUpperCase() !== 'BODY') {
                ga_event = utils.getAttribute(target, 'ga_event');
                ga_category = utils.getAttribute(target, 'ga_category') || '/search/';
                ga_label = utils.getAttribute(target, 'ga_label');
                ga_value = utils.getAttribute(target, 'ga_value') || 1;

                ga_event && window._czc && _czc.push(["_trackEvent", ga_category, ga_event, ga_label, ga_value]);
                ga_event && taAnalysis && taAnalysis.send('event', { ev: ga_event });
                target = target.parentNode;
            }
        });
```

通过网页调试工具我们可以看到，视频数据的获取是通过调用API传递相关参数得到的．



In [ ]: 接口的输出信息如下：



In [ ]: 通过处理接口返回的JSON数据，可以从中提取出短视频的名称以及播放地址等信息.

示例代码：

In [2]:
```
#coding:utf8
import json
import requests

def get_toutiao_movie(offset, content):
    data = {
        "aid": 24,
        'app_name': "web_search",
        "offset": offset,
        "format": "json",
        "keyword": content,
        "autoload": "true",
        "count": 20,
        "en_qc": 1,
        "cur_tab": 1,
        "from": "video",
        "pd": "video",
        "timestamp": 1566266654534
    }


    headers = {
        "authority": "www.toutiao.com",
        "method": "GET",
        "path": "/api/search/content/?aid=24&app_name=web_search&offset=0&format=json&keyword=%E7%94%84%E5%AC%9B%E4%BC%A0&autoload=true&count=20&en_qc=1&cur_tab=1&from
        "scheme": "https",
        "accept": "application/json",
        "accept-encoding": "gzip, deflate, br",
        "accept-language": "zh-CN,zh;q=0.9,en;q=0.8",
        "content-type": "application/x-www-form-urlencoded",
        "cookie": "tt_webid=6722299457688028679; WEATHER_CITY=%E5%8C%97%E4%BA%AC; tt_webid=6722299457688028679; csrftoken=270e8f13799286ebd339ed1159d28073; UM_distinct
        "referer": "https://www.toutiao.com/search/?keyword=%E7%94%84%E5%AC%9B%E4%BC%A0",
        "sec-fetch-mode": "cors",
        "sec-fetch-site": "same-origin",
        "user-agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/76.0.3809.100 Safari/537.36",
        "x-requested-with": "XMLHttpRequest",
        }

    url = "https://www.toutiao.com/api/search/content/"
    resp = requests.get(url, params=data, headers=headers)

    if resp.status_code == 200:
        return resp.json()
    else:
        return {}

def get_movie_result(message):
    result = {}
    for items in message.get("data", []):
        title = items.get("abstract")
        url = items.get("article_url")
        if not any([title is None, url is None]):
            result[title] = url
    return result

def main():
    api_json = get_toutiao_movie(0, "甄嬛传")
    result = get_movie_result(api_json)
    for (k, v) in result.items():
        print(k, v)

if __name__ == '__main__':
    main()
```

```
揣着字典才敢看电视剧！《甄嬛传》和《琅琊榜》，字都不认识？ http://toutiao.com/group/6726018994840535555/ (http://toutiao.com/group/6726018994840535555/)
甄嬛传：甄嬛到死不知，雍正为啥要把淳儿养到17岁，才开始宠幸她 http://toutiao.com/group/6726872061970481672/ (http://toutiao.com/group/6726872061970481672/)
甄嬛传：玉娆到底有多瞧不起纯元？雍正不会想到，可甄嬛心里明白 http://toutiao.com/group/6726872034191606280/ (http://toutiao.com/group/6726872034191606280/)
甄嬛传：敬妃背叛甄嬛的事，谁知甄嬛竟将胧月交给她抚养，好计谋 http://toutiao.com/group/6726796105180250631/ (http://toutiao.com/group/6726796105180250631/)
甄嬛传：甄嬛生下龙凤胎，苏培盛主动讨赏，不料皇上给了他惊喜 http://toutiao.com/group/6726805488249356814/ (http://toutiao.com/group/6726805488249356814/)
甄嬛传：甄嬛看见祺嫔脖子上的东西，才知那是麝香，还是皇后赏的 http://toutiao.com/group/6726798870673621507/ (http://toutiao.com/group/6726798870673621507/)
甄嬛传：祺贵人为何是光着脚逃出冷宫的？ http://toutiao.com/group/6726850264453612039/ (http://toutiao.com/group/6726850264453612039/)
甄嬛传谁看懂了华妃这几粒药下去，为何这三人的表情都变 http://toutiao.com/group/6726854892121489932/ (http://toutiao.com/group/6726854892121489932/)
甄嬛传：皇上早就在甄嬛身边安插亲信，甄嬛却不知，幸好有苏培盛 http://toutiao.com/group/6726402118988595726/ (http://toutiao.com/group/6726402118988595726/)
甄嬛传：苏培盛是真心对槿汐，在皇上面前落泪，皇上的反应也真逗 http://toutiao.com/group/6726799666609963527/ (http://toutiao.com/group/6726799666609963527/)
甄嬛传：剧中的她，得罪了甄嬛和皇后，却还能一直笑到最后 http://toutiao.com/group/6726447207362331143/ (http://toutiao.com/group/6726447207362331143/)
甄嬛传：甄嬛一回宫就痛杀后宫不良风气，此举连太后都说好，霸气 http://toutiao.com/group/6726801008762028551/ (http://toutiao.com/group/6726801008762028551/)
《甄嬛传》“娘娘们”集体亮相，蒋欣的话成亮点，你最喜欢哪一个 http://toutiao.com/group/6726527455898632711/ (http://toutiao.com/group/6726527455898632711/)
甄嬛传：甄嬛得知流朱死讯，悲痛不已，竟对自己下这样的狠手 http://toutiao.com/group/6726394780093252104/ (http://toutiao.com/group/6726394780093252104/)
甄嬛传：果郡王甄嬛死别时，皇上不派人偷听，这2点原因很可悲！ http://toutiao.com/group/6727057577714123271/ (http://toutiao.com/group/6727057577714123271/)
甄嬛传：槿汐苏培盛对食一事被东窗事发，甄嬛无奈，只得向她求助 http://toutiao.com/group/6726720035768762888/ (http://toutiao.com/group/6726720035768762888/)
```

In [ ]: