白雪爬虫和七个网站的故事

网络信息内容安全作业一

刘迅

2024/04/22

中国科学院大学

University of Chinese Academy of Sciences

目录

1 爬虫实验小结

实验目标

1 爬虫实验小结

	视频文件	视频简介	播放/点赞量	频道	标题	搜索	备注
ifeng	single	no	yes	no	yes	yes	
xiaodutv	single	no	yes	no	yes	yes	使用haokan搜索结果
thepaper	single	yes	yes	no	yes	yes	
haokan	single	no	yes	no	yes	yes	
ku6	single	no	no	no	yes	no	
cntv	segment	yes	yes	no	yes	yes	
bilibili	single	yes	yes	yes	yes	yes	

图: 目标网站所含功能一览

- 本次实验要求完成 7 个网站的视频内容爬取任务，功能涵盖视频标题、简介、点赞/播放量与频道获取，以及搜索功能，并实现单个/分段视频的下载。
- 由于网站设计差异，不同网站实现的功能有所差异。此处用红框列出与其他网站有较大不同的区别。

实验目标

1 爬虫实验小结

	视频文件	视频简介	播放/点赞量	频道	标题	搜索	备注
ifeng	single	no	yes	no	yes	yes	
xiaodutv	single	no	yes	no	yes	yes	使用haokan搜索结果
thepaper	single	yes	yes	no	yes	yes	
haokan	single	no	yes	no	yes	yes	
ku6	single	no	no	no	yes	no	
cntv	segment	yes	yes	no	yes	yes	
bilibili	single	yes	yes	yes	yes	yes	

图: 目标网站所含功能一览

1. cntv 的视频是分为多个 ts 文件传输，其他视频网站的视频都是单个媒体文件。
2. ku6 是七个网站中唯一一个没有播放/点赞量的。
3. ku6 还是七个网站中唯一一个没有搜索功能的。
4. xiaodutv 的搜索结果自动跳转到百度，聚合了 haokan, bilibili, weibo 多家搜索结果内容。为了方便起见，同时也不影响实验效果，xiaodutv 的搜索使用 haokan 的方法。

实验目标

1 爬虫实验小结

	视频文件	视频简介	播放/点赞量	频道	标题	搜索	备注
ifeng	single	no	yes	no	yes	yes	
xiaodutv	single	no	yes	no	yes	yes	使用haokan搜索结果
thepaper	single	yes	yes	no	yes	yes	
haokan	single	no	yes	no	yes	yes	
ku6	single	no	no	no	yes	no	
cntv	segment	yes	yes	no	yes	yes	
bilibili	single	yes	yes	yes	yes	yes	

图: 目标网站所含功能一览

“频道”项的处理

上课提到了没有频道标签的视频，如果需要分析其频道，应该如何处理，应该是引导大家向侧信道（通过弹幕、评论等信息分析）、聚类算法等角度思考。本次实验受时间所限，仅考虑有显式频道标签的 bilibili 网站，其他网站的频道留空处理。

实验成果

1 爬虫实验小结

本项目实现了实验要求对七个网站所有指定数据的抓取[1]，并能够下载单个/多段视频。

视频ID	视频标题	视频简介	播放/点赞量	频道	搜索词
BV1Vi421f7Go	《腐败的班级干部》	你行行好，就算再卡我，起码能不能明确告诉我哪里不行，这样我:	149.5万 播放	生活	-
BV1dS4y1P7h7	夜店式自习2	-	50.3万 播放	生活	-
BV1c341Q7XH	这一定就是原片吧9	视频剧情与人设皆为虚构 不要上升真羊！	5173.3万 播放	动画	-
BV1VE411j7wk	活0活现	微博	1842.1万 播放	生活	-
BV1GC41137fv	当 世 界 只 剩 一 个 00 后	弟长命百岁，永远不死！！！！！！！！！！！！！！！！！！	370.6万 播放	生活	-
BV1CP4y1W7WJ	官宣了，家人们	-	159.7万 播放	生活	-
BV16H4y1K7tJ	总裁！夫人她又跑了……	感谢王妈的客串！(> ㅁ <‡)	216.1万 播放	生活	-
BV1f8411c7Bx	别再问我做什么工作的了……	未找到简介	231.7万 播放	生活	-
BV1LF411f7TW	【柠×害羞7】起猛了，看到卡芙卡公主抱银狼	-	158.7万 播放	时尚	-
BV1fK4y1c7xT	10年HR教我的野路子，真的百试百灵诶！	这招真是妙啊…	23.5万 播放	生活	-
BV1T94y1T7hN	中东到底在争什么？一口气讲透中东乱局	突，5次中东战争，3次石油危机，中东都在争什么？一口气拿下中	4.4万 播放	知识	中东
BV1Bi4y147Y8	中东战争究竟在打什么？以色列的诞生与崛起【历史调研室08】	求关注求三连！	368.9万 播放	知识	中东
BV1Z8411G7dd	中东：“地缘政治熔炉”如何炼成？	中东：“地缘政治熔炉”如何炼成？	13.7万 播放	知识	中东
BV1UR4y1Q7DE	2022卡塔尔世界杯宣传片：赶紧来感受一下中东风情吧~	youtube	3.7万 播放	生活	中东
BV19w4m117oA	纳赛尔失败以后，中东反帝靠什么？【中东20】	一键三连推荐给更多朋友！	37.6万 播放	知识	中东
BV1ZM4m1X7B2	俄乌失控，中东升温，全球棋手争相“熬鹰”？【熬鹰纪元】	俄乌失控，中东升温，全球棋手争相“熬鹰”？	284.4万 播放	知识	中东
BV1T34y1Z74V	反战公益短片《决定》，中东小男孩的一天！	战乱中的普通人没有选择！深深的无力感	23.0万 播放	影视	中东
BV1H94y1R7Ua	【硬核】一口气了解石油和中东的百年之战｜油价 OPEC 中东…	20世纪，可以说石油是经济的关键，而中东是石油的关键。	242.6万 播放	知识	中东
BV1rZ4y1U7Zh	这么漂亮的地理课件，b站上可不好找，初中地理《中东》	课件获取方式，见评论区置顶评论	10.3万 播放	知识	中东
BV1UC4y137q2	什么是中东国家，中东战略位置有多重要？中东实力前5名都有谁？一支视频告诉你	东国家，中东战略位置有多重要？中东5国实力前5名都有谁？一支视	1.4万 播放	知识	中东

图: 例子： bilibili 爬取结果整理

[1] 只要有的话

实验成果

1 爬虫实验小结

1+X+1

(X=7)

代码框架，更具可拓展性的贡献之处

目录

2 创新点与主要贡献

目录

2 创新点与主要贡献

1+X+1 代码框架

2 创新点与主要贡献

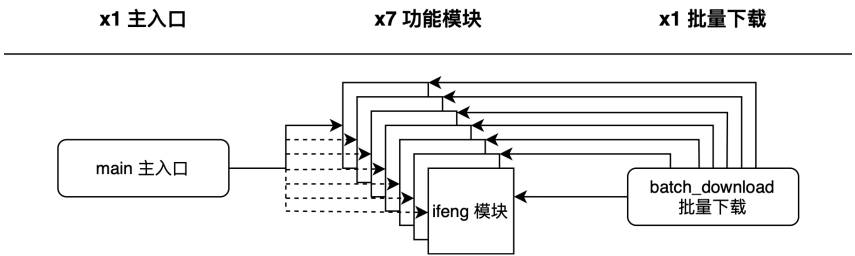

图: 1+X+1 代码框架

项目构建了一套统一的代码框架，用户只需要在 `main` 中选择目标视频网站，后续操作即可屏蔽不同网站的差异，通过功能一致的模块对外交互。

1+X+1 代码框架：入口

2 创新点与主要贡献

```python
src > main.py > ...
if __name__ == '__main__':
    print('==========欢迎来到聚合视频下载系统! ==========')
    website = input("请输入视频网站: ")
    if website == 'haokan':
        import haokan as website
    elif website == 'thepaper':
        import thepaper as website
    elif website == 'ifeng':
        import ifeng as website
    elif website == 'xiaodutv':
        import xiaodutv as website
    elif website == 'ku6':
        import ku6 as website
    elif website == 'cntv':
        import cntv as website
    elif website == 'bilibili':
        import bilibili as website
    id = input("请输入视频ID: ")
    keyword = input("请输入搜索词: ")
    if id:
        website.get_video_info(id)
    elif keyword:
        website.search_video(keyword)
    else:
        print("请输入视频ID或搜索词...")
```

图: 主入口示例

```
==========欢迎来到聚合视频下载系统! ==========
请输入视频网站: ifeng
请输入视频ID:
请输入搜索词:    小米
视频URL: https://video19.ifeng.com/video09/2024/04/21/p178781361381651718O-102-222416.mp4
        正在保存ifeng视频: 吉利、智己、极越包围小米，围剿还是蹭流量？.mp4
        成功保存ifeng视频: 吉利、智己、极越包围小米，围剿还是蹭流量？.mp4
------根据 关键词 获取视频信息------
视频标题是: 吉利、智己、极越包围小米，围剿还是蹭流量？
ifeng 没有简介
88 点赞
ifeng 没有频道
视频存储路径: data/ifeng/吉利、智己、极越包围小米，围剿还是蹭流量？.mp4
ifeng视频且去存储: logs/ifeng/吉利、智己、极越包围小米，围剿还是蹭流量？.txt
视频URL: https://video19.ifeng.com/video09/2024/04/20/p178722970732730846l-102-080538.mp4
        正在保存ifeng视频: 苹果全球销量降10%，小米传音却猛增为何？.mp4
        成功保存ifeng视频: 苹果全球销量降10%，小米传音却猛增为何？.mp4
------根据 关键词 获取视频信息------
视频标题是: 苹果全球销量降10%，小米传音却猛增为何？
ifeng 没有简介
1102 点赞
ifeng 没有频道
视频存储路径: data/ifeng/苹果全球销量降10%，小米传音却猛增为何？.mp4
ifeng视频且去存储: logs/ifeng/苹果全球销量降10%，小米传音却猛增为何？.txt
视频URL: https://video19.ifeng.com/video09/2024/04/16/p7185921135341478844-102-170536.mp4
        正在保存ifeng视频: 小米SU7细节，对比才有差距#小米su7 #新能源汽车 #小米汽车.mp4
        成功保存ifeng视频: 小米SU7细节，对比才有差距#小米su7 #新能源汽车 #小米汽车.mp4
------根据 关键词 获取视频信息------
视频标题是: 小米SU7细节，对比才有差距#小米su7 #新能源汽车 #小米汽车
ifeng 没有简介
```

图: 运行结果示例

1+X+1 代码框架：功能模块

2 创新点与主要贡献

图: 功能模块框架示例-1

图: 功能模块框架示例-2

目录

2 创新点与主要贡献

项目代码流程图分析

2 创新点与主要贡献

使用 pycallgraph 工具绘制代码调用关系图，更好地分析逻辑关系和性能瓶颈。

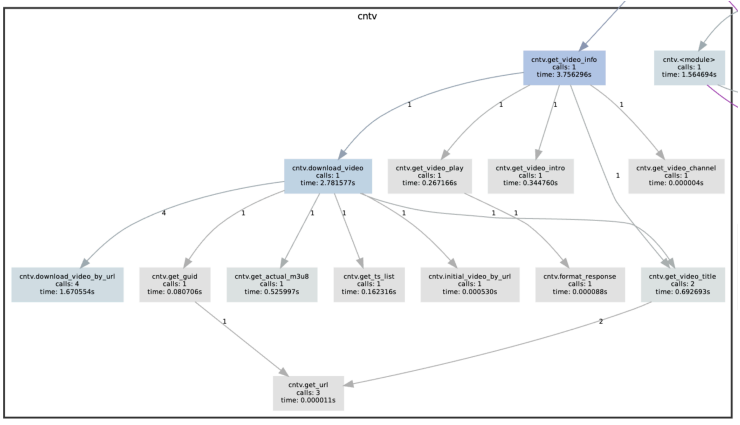

图: 例子： cntv 模块的流程图分析

目录

3 讨论与展望

亮点与可改进之处

3 讨论与展望

本项目的一大特点在于

- **全程使用 `requests` 库构建与网站的交互**，相较于使用 selenium 或 playwright，效率上有所优势。

显而易见的，本次项目只是一个初步的尝试，在性能、鲁棒性等角度都还有很大的提升空间，具体来说有

1. 实践过程中，偶尔会因为个别请求超时或中断而影响整个程序运行，可以加入 `try-except` 提升整体的鲁棒性。

未解之谜

好看视频 cookie 易过期。假如一段时间爬取过于频繁，即便设置 user-agent 与 cookie，也会被拒绝爬取。放着一段时间又好了！

亮点与可改进之处

3 讨论与展望

本项目的一大特点在于

- **全程使用 `requests` 库构建与网站的交互**，相较于使用 selenium 或 playwright，效率上有所优势。

显而易见的，本次项目只是一个初步的尝试，在性能、鲁棒性等角度都还有很大的提升空间，具体来说有

1. 实践过程中，偶尔会因为个别请求超时或中断而影响整个程序运行，可以加入 `try-except` 提升整体的鲁棒性。
2. 可以使用异步、高并发的方式提升下载性能。

未解之谜

好看视频 cookie 易过期。假如一段时间爬取过于频繁，即便设置 user-agent 与 cookie，也会被拒绝爬取。放着一段时间又好了！

思考：从爬虫对抗一窥攻防博弈

3 讨论与展望

对各个网站的爬取过程是对攻防实践非常生动的诠释，这是一个螺旋上升的过程。

- 攻击：从静态网页硬编码的 mp4 URL 中提取。
- 防御：动态网页加载 mp4URL；静态网页显示 blob:xxx。
- 攻击：selenium（本次实验未采用）提取渲染后的 HTML 文件；或者观察对 mp4 文件的 js 请求。
- 防御：分段传输视频文件。
- 攻击：通过全局表示 guid 寻找视频片段（凤凰网）。
- 攻击：将 User-Agent 切换到手机查看（B 站）。
- 防御：检查时间戳 timestamp（好看视频）。

思考：屡禁不绝 v.s. 有意为之?

在这些攻防对抗中，看到了很多防御/混淆措施，但是爬虫之所以不能够根治的原因是什么呢?

个人猜测有技术和非技术两方面的因素：

- 技术上：对于访问媒体文件的真实用户，用户本地的浏览器必须拿到媒体资源的 URL，才能够播放。无论传输过程如何加密，由于视频文件的播放这一步骤必须在本地进行，因此总会暴露出媒体文件的地址。

思考：屡禁不绝 v.s. 有意为之?

在这些攻防对抗中，看到了很多防御/混淆措施，但是爬虫之所以不能够根治的原因是什么呢?

个人猜测有技术和非技术两方面的因素：

- 技术上：对于访问媒体文件的真实用户，用户本地的浏览器必须拿到媒体资源的 URL，才能够播放。无论传输过程如何加密，由于视频文件的播放这一步骤必须在本地进行，因此总会暴露出媒体文件的地址。
- 非技术：视频网站需要被搜索引擎收录、被互联网用户使用，倘若反爬虫做到极致，也就与内网无异了。

致谢

感谢杨杼鑫、秦子茗同学的交流、讨论与指导！

感谢朱老师不那么 push 的 DDL

白雪爬虫和七个网站的故事

大作业完结撒花！

任何问题?