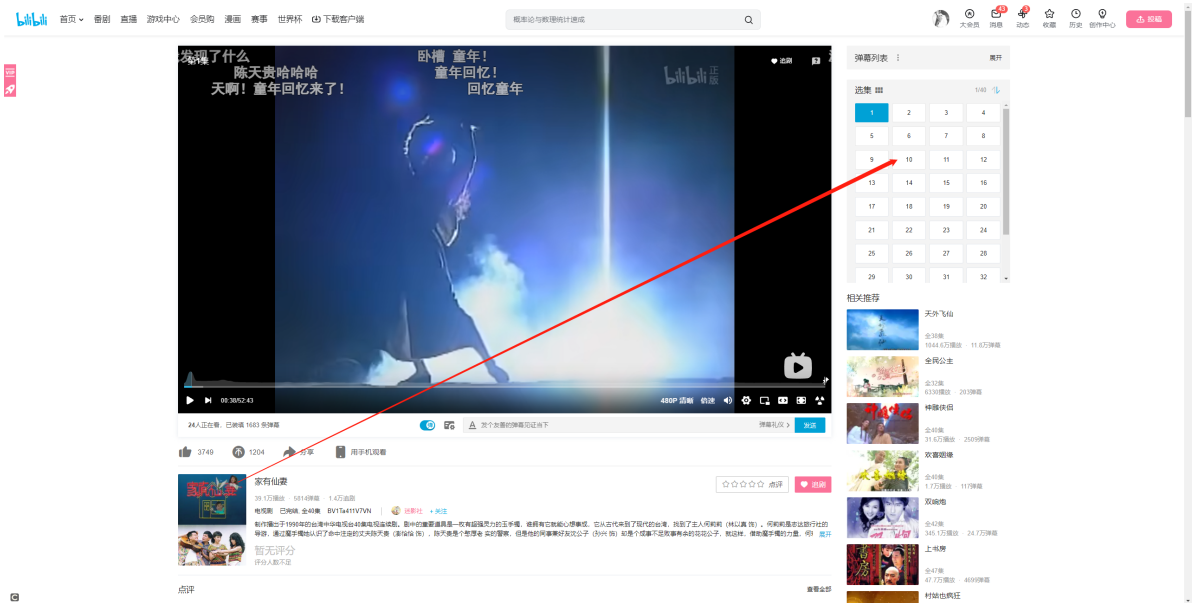


爬取哔哩哔哩弹幕信息案例



本次案例目标是抓取B站电视剧上的家有仙妻前十集的弹幕评论

使用工具:

Google Chrome(108.0.5359.95)浏览器, IDE: PyCharm2022.2.2

网页数据分析:

开始设想:

希望通过抓包工具(Network)对页面弹幕评论加载方式进行分析



接下来我们来尝试点击查看历史弹幕，来请求获取返回的数据包。

点击原本展开的位置是返回了，如此多通过通过ajax交互返回的数据包。

正当我开始吐槽的时候才发现事情远远没有这么简单！

```
1 |
2 \rs a a o dLEdLE jCANSOH EH ( d d d d d BEL2 BS d135f15b: s2 哈哈哈哈哈@ i o c kHearBoc1178855
3 T rs 莹 o dLEdLE jCANSOH EH ( d d d d d BEL2 BS ca906788: c kHHeelBoc11789003954
4 k rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS ca906788: esc女主业务能力可以的@ o c kHvtBoc1178902607
5 T rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS ca906788: c kHHeelBoc1178902607
6 b rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 1ee3547f: n哈哈哈哈哈哈哈@ o c kHvtBoc1178902607
7 c rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS eecb47df: n我愛他爹第一嘛@ o c kHvtBoc1178902607
8 j rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 21311a25: es:这个套路现在还在用@ o c kHvtBoc1178902607
9 V rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS a6feb51e: 哈哈哈哈哈@ o c kHvtBoc1181227
10 n rs u o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 44ade4d5: es 恶名昭彰的台湾客人...@ o c kHvtBoc1181227
11 j rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS a8d30338: s1 哇, 抽烟啦@ o c kHvtBoc1181227
12 s o h BS i o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 780fa70: e受何丽丽的影响, 后来我也做
13 W rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 4420a0b2: 梅有财@ o c kHvtBoc1181227
14 e rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 63c354b5: c歐阳娜娜的爸爸: @ o c kHvtBoc1181227
15 V rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 41729b24: 有一腿@ o c kHvtBoc1181227
16 P rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 41729b24: etu哈哈@ o c kHearBoc1181727422
17 S rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS e22b913b: !真正的主角登场了, 哈哈@ o c kHvtBoc1181727422
18 Z rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 368a5452: ff 三吉彩花@ o c kHvtBoc1181727422
19 \rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 110687f7: s2 哈哈哈哈哈哈哈@ o c kHvtBoc118205975
20 p rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 88364d2: !我記得梅总才是搞笑担当@ o c kHvtBoc118205975
21 x rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 92129d7e: 大家记住: 凡尔赛最早的出处@ o c kHvtBoc118205975
22 V rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 56cd2ed4: c kHvtBoc1182580699811
23 W rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 28eb1cde: 见鬼啦@ o c kHvtBoc1182580699811
24 p rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS ec188b49: !哈哈, 手镯子很爱女主人啊@ o c kHvtBoc1182580699811
25 Y rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS ec188b49: s哈哈哈哈哈哈哈@ o c kHvtBoc1182580699811
26 Y rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS ec188b49: ff 笑死我了@ o c kHvtBoc1182580699811
27 l rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 12623423: s 见莉莉好像就是本名啊@ o c kHvtBoc1182580699811
28 V rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 12623423: "恶名昭彰的台湾客人? 有故事啊"
29 Z rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS f6fea556: ff 特效不错@ o c kHvtBoc1182580699811
30 l rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS f6fea556: es 现在抽烟镜头都没有了@ o c kHvtBoc1182580699811
31 Z rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS fc2bb340: "哈哈哈哈哈, 看那个表情已经笑
32 m rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS d8e8789d: s居然来这部剧! 太牛啦@ o c kHvtBoc1182580699811
33 * rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 268336: s歐阳娜娜的爸爸是欧阳龙好吧, 瞎子
34 b rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS c274001b: n有点像慧慧姐耶@ o c kHvtBoc1182580699811
35 k rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 22f16d1: s这嘴巴子哈哈哈哈哈哈哈@ o c kHvtBoc1182580699811
36 \rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS db2d5d87: s哈哈哈哈哈哈哈@ o c kHvtBoc1182580699811
37 a rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 697295b: n早期铃木吉妮尼@ o c kHvtBoc1182580699811
38 s o h BS i o dLEdLE jCANSOH EH ( d d d d d BEL2 BS 98cba2b1: U那时候的演员真的是演什么像什
39 Z rs e o dLEdLE jCANSOH EH ( d d d d d BEL2 BS c3138ff7: ff 香奈儿啊@ o c kHvtBoc1182580699811
```

当我点开返回包预览的时候发现这里的数据都被加密了(有点类似与字符加密),也有可能是经过压缩gzip压缩的目前现在传输中基本上都是用压缩模式进行传输的。

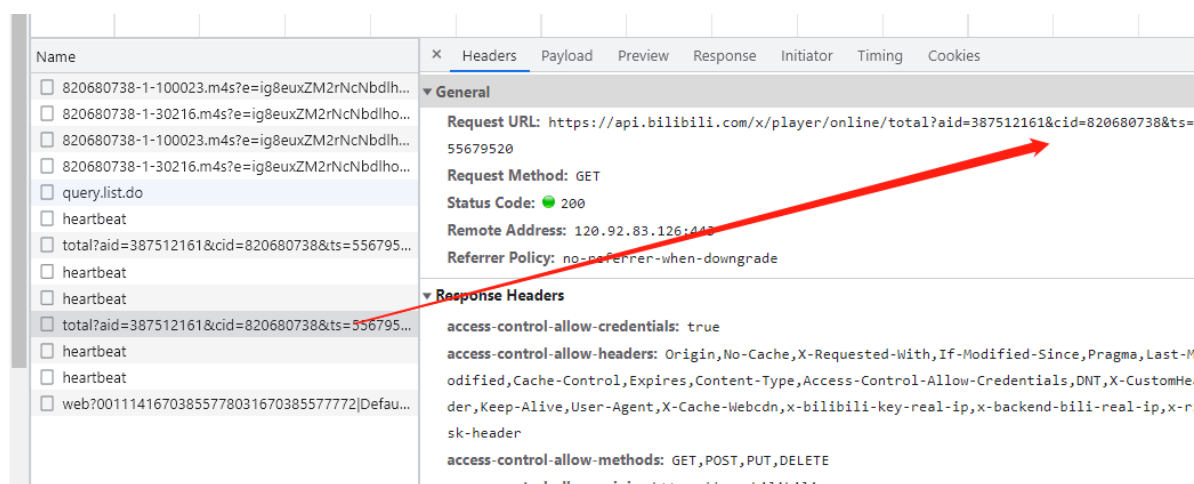
到这里貌似也没有特别友好的解决方案。

那么我没有没有一种办法可以绕过字体反爬直接获取b站特定评论的api接口呢?

答案是有的

确定方向:

留意这个包



这是b站一个接口给我们返回的数据链接

平常我们在看视频时,弹幕是出现在视频上的。实际上在网页中,弹幕是被隐藏在源代码中,以XML的数据格式进行加载的:

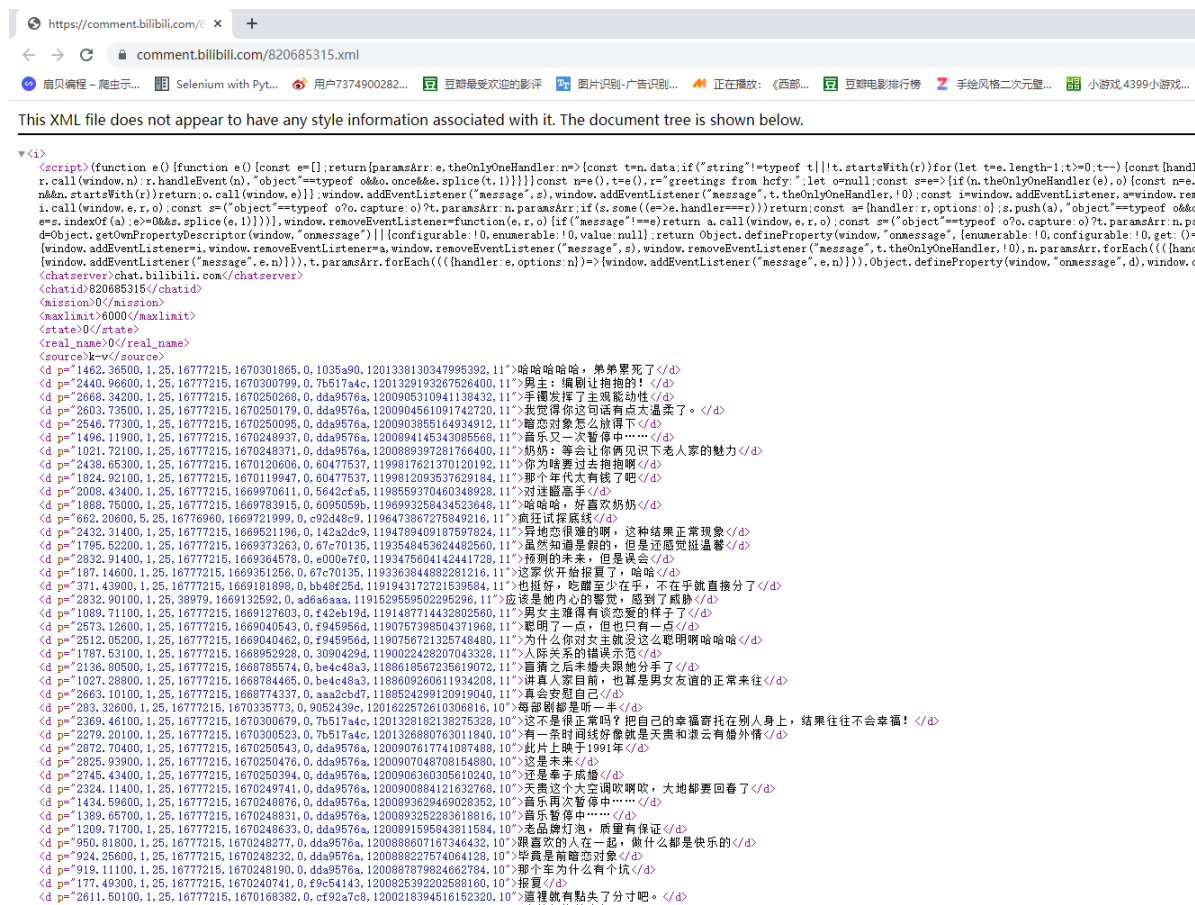
XML和JSON、YAML一样是一种通用的标记信息表达方式,可以简单的理解为一种记录数据的格式。

那么这个弹幕文件的url是什么呢?

它以一个固定的url地址+视频的cid+.xml组成。只要找到你想要的视频cid,替换这个url就可以爬取所有弹幕了(b站大部分网页给出的字幕限制是1000条)

```
1 | url = f'http://comment.bilibili.com/{cid}.xml'
```

XML和描述网页的语言HTML非常像，所以你会在截图中看到这样的标签。



解决方案:

好了 通过对xml文件的访问居然可以直接获取对应评论的所有信息，那么我们可以直接根据每个视频特定的cid代码就可以获取到对应的视频弹幕url了

我们把前十集的cid抓取出来放到一个列表里对每个特定评论的url链接进行拼接
最终一次全部抓取到弹幕数据

```
1 # 家有仙妻 前十集的cid值
2 cids = [820680426, 820680738, 820681510, 820682146,
820682647, 820682989, 820683722, 820684233, 820684795,
820685315]
```

然后通过循环对每个cid进行拼接:

```
1 for cid in cids:
2     url = f'http://comment.bilibili.com/{cid}.xml'
```

从而获得每个评论仓库的特定地址

树对象

```

1 def getTree(url):
2     res = requests.get(url, headers=headers)
3     res.encoding='utf-8'
4     tree = etree.HTML(res.text.encode('utf-8'))
5     return tree

```

抓取函数

```

1 def fetch(cid):
2
3
4     url = f'http://comment.bilibili.com/{cid}.xml'
5     print(f'{url}准备被抓取')
6     # 获取树解析对象
7     tree = getTree(url)
8     # 解析方法返回一组数据列表
9     rows = parse_data(tree)

```

解析数据

需要用到的python库

```

1 from lxml import etree

```

解析函数

```

1 def parse_data(tree):
2
3     element_list = tree.xpath('//d')
4
5     rows = []
6     for attrs_ele in element_list:
7         #
8         2695.34400,1,25,16707842,1670310086,0,355b8c5,120140709472
9         9993216,11
10        attr_ele = attrs_ele.xpath('./@p')[0]
11
12        # 弹幕出现的时间

```



```

11         emerge_time = attr_ele.split('.')[0]
12         m_emerge_time = time.strftime("%H:%M:%S",
time.gmtime(int(emerge_time)))
13         # 时间戳
14         timestamp = time.strftime('%Y-%m-%d %H:%M:%S',
time.localtime(int(attr_ele.split(',')[4])))
15         # 内容
16         texts = attr_ele.xpath('./text()')[0]
17         rows.append([m_emerge_time, timestamp, texts])
18
19     return rows

```

数据存储(CSV)

需要用到的python库:

```

1 import csv
2 import os

```

创建总的文件夹、再分别存在每一集的CSV

```

1 def save_data(episode_name, rows, index):
2
3     # 创建csv文件夹来对每一集的弹幕进行存储
4     if not os.path.exists(episode_name):
5         os.mkdir(episode_name)
6     csv_name = f'家有仙妻第{index}集'
7     path = os.path.join(episode_name, csv_name + '.csv')
8     with open(path, 'a', encoding='utf-8', newline='') as
f:
9         csv_writer = csv.writer(f)
10        header = ['弹幕信息出现分钟', '弹幕出现日期', '弹幕内
容']
11        csv_writer.writerow(header)
12        # 接下来循环遍历这个数组列表
13        for row in rows:
14            csv_writer.writerow(row)
15
16    print(csv_name, '已存储完成')

```

到此，我们完全就可以通过修改cid值来获取所有我们想要获取的视频评论了。

最终效果展示:

名称	修改
家有仙妻第1集.csv	20%
家有仙妻第2集.csv	20%
家有仙妻第3集.csv	20%
家有仙妻第4集.csv	20%
家有仙妻第5集.csv	20%
家有仙妻第6集.csv	20%
家有仙妻第7集.csv	20%
家有仙妻第8集.csv	20%
家有仙妻第9集.csv	20%
家有仙妻第10集.csv	20%

Project	代理池.py	干微博.py	抓课表.py	独行月球短评.csv	电影评论api——exl.py
家有仙妻弹幕汇总	1	弹幕信息出现分钟, 弹幕出现日期, 弹幕内容			
家有仙妻第1集.csv	2	00:48:28, 2022-12-06 16:40:46, 哈哈哈哈哈			
家有仙妻第2集.csv	3	00:48:04, 2022-12-06 16:40:17, 笑死哈哈哈			
家有仙妻第3集.csv	4	00:45:09, 2022-12-06 16:37:16, 这发型和许半夏的一样诶			
家有仙妻第4集.csv	5	00:38:02, 2022-12-05 20:33:38, 女主怎么突然对男主很好了			
家有仙妻第5集.csv	6	00:41:58, 2022-12-04 20:10:56, 我想看到他, 每次他出现就很欢乐			
家有仙妻第6集.csv	7	00:48:12, 2022-12-03 21:47:37, 哈哈			
家有仙妻第7集.csv	8	00:10:14, 2022-12-03 00:16:53, 角度新奇			
家有仙妻第8集.csv	9	00:13:32, 2022-12-02 21:25:27, 小扫把多可爱啊			
家有仙妻第9集.csv	10	00:04:10, 2022-12-02 21:14:01, 天贵命中带花			
家有仙妻第10集.csv	11	00:01:52, 2022-11-29 19:25:56, 沈公子这头发应该可以缓冲下			
	12	00:10:19, 2022-11-29 11:09:34, 哈哈那会吸引苍蝇吧			
	13	00:02:42, 2022-11-29 11:01:44, 好家伙最危险的地方最安全			
	14	00:16:12, 2022-11-26 14:47:10, 这是要恩将仇报啊哈哈			
	15	00:30:24, 2022-11-25 22:20:09, 这女儿真好			
	16	00:02:48, 2022-11-21 18:26:52, 嘿! 我还没上车呢!			

!!!

其实，对于file_name是可以做一个抓取函数的来抓取特定页面的电视剧名称，进而不需要手写file_name了(这个感兴趣的小伙伴可以自己实现，在这里只做案例，不负责封装类)，这样以后就可以直接修改cid来打包每一个想要的视频评论了，当然想要获取更多的评论，需要具体问题具体分析，根据特定电视剧的集数来获取对于的cid值再存入cid列表里就好啦！