

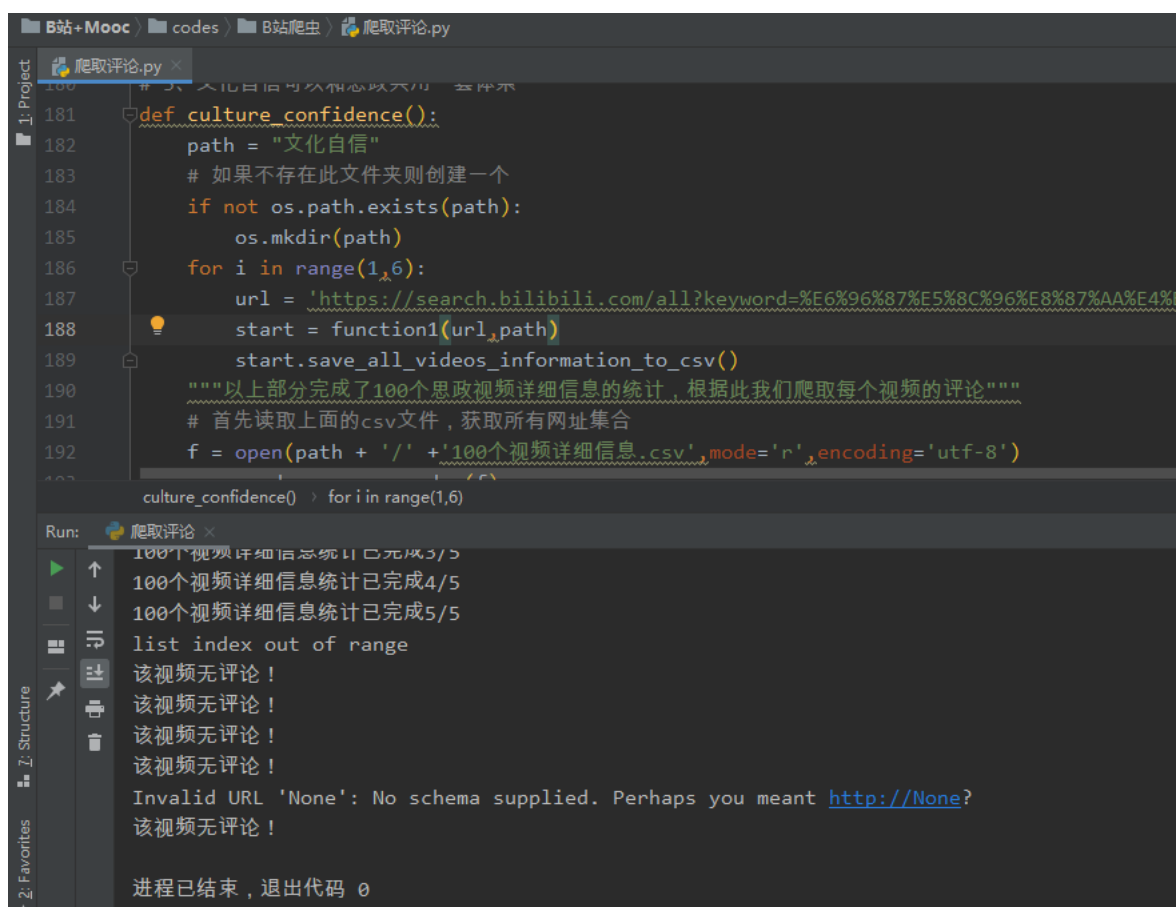
1、B站思政课和文化自信、中国医学史视频评论爬虫

对应的是B站爬虫需求的第一、二点！定义了一个类来减少冗余代码！B站的反爬措施没多少，难点是在与找到评论的接口地址！

大致思路如下：

- B站首页搜索【思政课】，发现一个页面有20个视频，我们将所有视频的链接以及标题等数据存放到csv文件中
- 根据上述的csv文件将得到的视频网址一个个请求，得到av号码
- 根据av号码构造评论链接地址
- 使用json()进行解析，将评论数据写入新的csv文件中

运行截图：



```
def culture_confidence():
    path = "文化自信"
    # 如果不存在此文件夹则创建一个
    if not os.path.exists(path):
        os.mkdir(path)
    for i in range(1,6):
        url = 'https://search.bilibili.com/all?keyword=%E6%96%87%E5%8C%96%E8%87%AA%E4%B'
        start = function1(url,path)
        start.save_all_videos_information_to_csv()
    """ 以上部分完成了100个思政视频详细信息的统计，根据此我们爬取每个视频的评论 """
    # 首先读取上面的csv文件，获取所有网址集合
    f = open(path + '/' + '100个视频详细信息.csv',mode='r',encoding='utf-8')
    culture_confidence()
    for i in range(1,6)
```

Run: 爬取评论 x

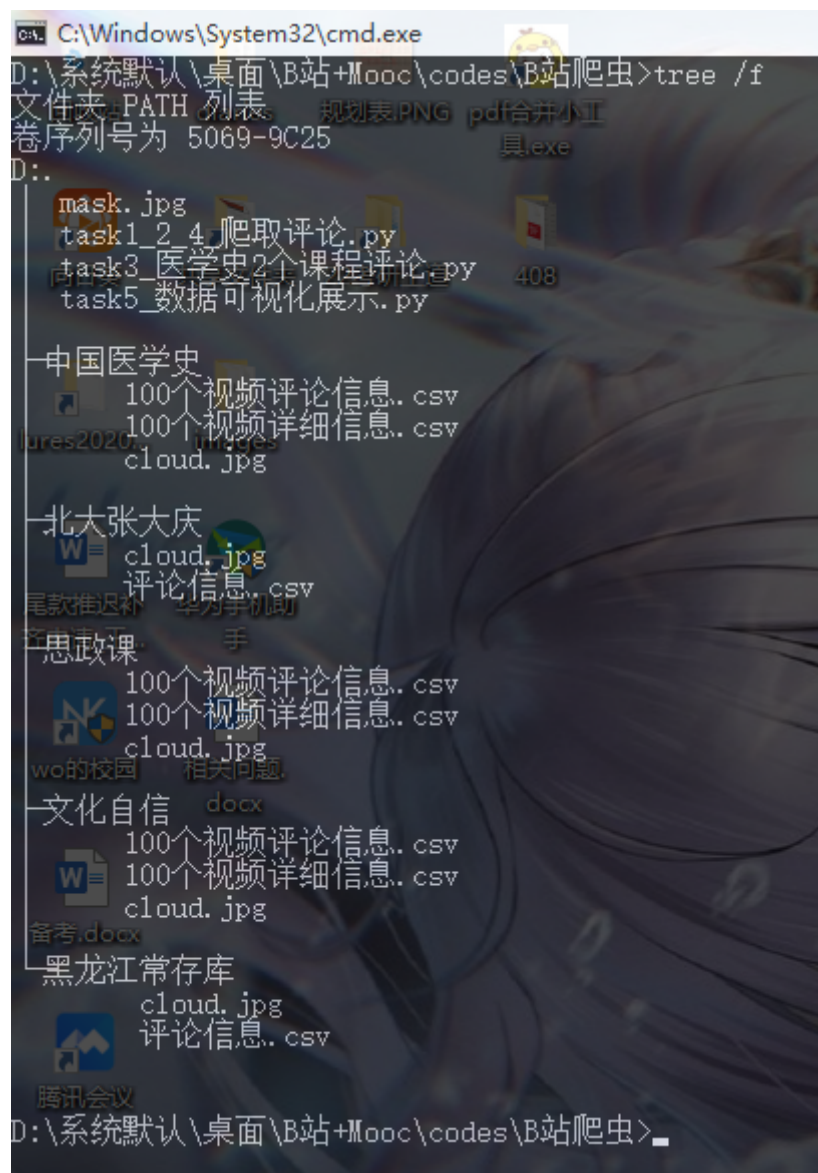
100个视频详细信息统计已完成3/5
100个视频详细信息统计已完成4/5
100个视频详细信息统计已完成5/5
list index out of range
该视频无评论！
该视频无评论！
该视频无评论！
该视频无评论！
Invalid URL 'None': No schema supplied. Perhaps you meant http://None?
该视频无评论！
进程已结束，退出代码 0

对应的代码相对路径为：B站+Mooc/codes/B站爬虫/task1_2_爬取评论.py

2、B站两门关于【医学史】的课程评论爬取

和第一问的一部分类似，因为观察到两个课程评论的页数都是2页，所以可以构造for循环以及上一例子用到的评论接口，所以可以快速爬取两个课程的评论信息！

运行截图：



4、Mooc爬取4门课程的评价信息

首先，一定要下载对应浏览器版本的 webdriver，具体操作过程可参考文章：

https://blog.csdn.net/weixin_44335092/article/details/109054128（不负责教环境安装）

注意，要和当前项目的解释器 python.exe 放到同一文件夹中

i > 本地磁盘 (D:) > python

名称	修改日期	类型	大小
DLLs	2019-04-23 1:04	文件夹	
Doc	2019-04-23 1:03	文件夹	
docx-template	2019-11-20 22:18	文件夹	
etc	2019-09-07 11:03	文件夹	
ffmpeg	2020-03-08 11:59	文件夹	
include	2019-05-21 8:58	文件夹	
Lib	2020-11-06 18:37	文件夹	
libs	2019-04-23 1:04	文件夹	
pdfkit_wkhtmltox	2020-05-16 21:14	文件夹	
pyinstaller-develop	2020-11-01 12:59	文件夹	
Scripts	2021-01-03 15:43	文件夹	
share	2020-10-23 7:27	文件夹	
tcl	2019-04-23 1:04	文件夹	
Tools	2019-05-18 1:17	文件夹	
venv	2020-06-14 17:07	文件夹	
chromedriver.exe	2020-09-02 5:09	应用程序	9,494 KB
geopy-1.20.0-py2.py3-none-any.whl	2019-11-02 16:14	WHL 文件	95 KB
LICENSE.txt	2018-06-27 5:03	文本文档	30 KB
NEWS.txt	2018-06-27 5:03	文本文档	595 KB
opencv_videoio_ffmpeg411_64.dll	2019-09-07 0:37	应用程序扩展	21,325 KB
python.exe	2018-06-27 5:01	应用程序	98 KB
python3.dll	2018-06-27 5:00	应用程序扩展	58 KB
python37.dll	2018-06-27 5:00	应用程序扩展	3,755 KB
pythonw.exe	2018-06-27 5:01	应用程序	97 KB

最后，指定 webdriver 的绝对路径：

```

1 from selenium import webdriver
2 from time import sleep
3 import os
4 import re
5 from bs4 import BeautifulSoup #executable_path为chromedriver.exe的解压安装目录，需要与chrome浏览器同一文件夹下
6 driver=webdriver.Chrome(executable_path="D:/python/chromedriver.exe")
7 url='https://www.icourse163.org/course/BIT-268001' #爬取Python语言程序设计为例
8 driver.get(url)
9 cont=driver.page_source #获得初始页面代码，接下来进行简单的解析

```

正确安装后，运行截图：

task3_医学史2个课程评论.py

task1_2_4_爬取评论.py

task5_数据可视化展示.py

demo.py

```

1 from selenium import webdriver
2 from time import sleep
3 import os
4 import re
5 from bs4 import BeautifulSoup
6 driver=webdriver.Chrome(executable_path="D:/python/chromedriver.exe")
7 url='https://www.icourse163.org/course/BIT-268001'
8 driver.get(url)
9 cont=driver.page_source

```

Run: demo

D:\python\python.exe

中国医学史课程评价爬取

思想道德修养与法律基础

走近中华优秀传统文化_南京大学

https://www.icourse163.org/course/NJU-1002190001

Cent Browser 正受到自动测试软件的控制。

新乡铁一中夏雪 ★★★★★

受益匪浅，感受到中华传统文化的魅力。

发表于 2020-05-22 第6次开课

新乡市铁路高级中学李铭伟 ★★★★★

很有收获，继续学习

发表于 2020-05-22 第6次开课

上一页

1

...

16

17

18

19

20

...

98

下一页

数据分析部分和前面B站爬虫类似，都是绘制词云图！

