雲南農業大學
Yunnan Agricultural University

实践课程报告

课程名称 **Python 数据处理**

题目名称 **爬虫程序设计**

学生学院 **大数据学院**

年级专业 **19 计科专升本**

学　　号 **2019610040**

姓　　名 **杨号星**

完成时间 **2022 年 7 月 8 日**

摘要

随着计算机技术的不断发展，新的编程语言层出不穷，Python 正是其中的佼佼者。相比较早期普及的高级语言（Java,C 语言）等，Python 有着更加实用的模块和库，虽然牺牲了底层性，但却更加方便用于开发小型项目。基于 python 的网络爬虫技术，相比于通用的搜索引擎更具有目的性和灵活性，它能根据选定的抓取目标，有选择的访问万维网上的网页与相关的链接，获取所需要的信息。本文以人民日报新闻爬取和爬取后的保存及查询为研究，实现了一个基于 python 的人民日报新闻文章爬取程序。本论文还阐述了一些网络爬虫实现的常见问题，包括常用的 python 的网络请求、如何解决网页的反爬问题、数据保存写入问题等。

本程序最终可以实现对人民日报（http://paper.people.com.cn/）新闻文章的下载。可以输入要爬取的日期以及结束日期，将这些日期内的文章全部爬取下来，以日期为名自动生成一个主存储目录，爬取到的文章保存写入 txt 文件中,每个文本的存储名字以日期加序号存储。

关键词：网络爬虫；Python；；网络请求；人民日报新闻。

目录

正文

1 绪论

随着科学技术的不断发展，信息流通日益方便，信息数据不断膨胀，充斥在各行各业。由于数据非常庞大，所以即使在搜索引擎存在的情况下，搜索结果的准确率也不高，这使得在网上查找关键有效信息也变为一项极具挑战性的复杂任务。我们可以通过基于 python 的网络爬虫技术很好的解决检索问题，首先我们通常使用的通用搜索引擎的目标是尽可能将网络覆盖率增大，其次在数据形式复杂的情况下，对于具有一定结构且信息含量密集的数据，往往不能被很好的搜索出来。而网络爬虫则更具有目的性，能根据选定的抓取目标，有选择的访问万维网上的网页与相关的链接，获取所需要的信息。

将爬虫技术应用于文章的爬取并筛选有效信息，可以节省科研人员时间，提高资源利用率，将有限的时间发挥更大的价值。

2 相关技术介绍

2.1 网络爬虫技术

2.1.1 网络爬虫技术概述

网络爬虫技术是近些年来成熟并流行起来的一项技术。现阶段研究通常集中在各种不同领域下的运用。其通俗的来说就是通过模拟客户端（各种浏览器）发送网络请求，以获取服务端的响应，并按照规则提取指定数据的程序。

2.1.2 python 的网络请求

python 中常用的 HTTP 网络请求通常有 3 种方式：urllib、urllib3 以及本论文使用的 requests。下面将介绍 requests 模块的使用。

requests 模块是 python 的第三方模块，该模块在实现 HTTP 请求时要比其他 2 种方式简单，在使用前需要先在 cmd 命令行里执行 pip install requests。

requests 模块中使用最多的就是 GET 和 POST 请求方式，2 者的主要区别在于 GET 请求没有请求体，它把数据放在 url 地址中，而 POST 有请求体，常用于登录注册，且它携带的数据量比 GET 请求方式大，所以常用于传输大文本。

本文中使用的是 GET 请求方式，它的请求方式为：

response=requests.get(url,params=params,headers=headers)

1.url 为基准的 url 地址，不包含查询参数。

2.该方法会自动对 params 字典编码，然后和 url 拼接。

2.1.3 如何解决网页的反爬问题

通常我们在请求一个网页时，无论是通过哪种请求，发现如果不带 headers 参数一般会出现 403 错误，这种错误的原因是通常网页为了防止恶意的采集数据信息会设置一些反爬措施，从而拒绝爬虫程序的访问。因而通过携带 headers 参数，可以达到模仿浏览器的头部信息进行访问。具体反爬策略：第一遍，先爬取版面目录，将每一个版面的链接保存下来；第二遍，依次访问每一个版面的链接，将该版面的文章链接保存下来；第三遍，依次访问每一个文章链接，将文章的标题和正文保存到本地。

3 设计目的与要求

3.1 程序设计的目的与要求

实现对人民日报（http://paper.people.com.cn/）新闻文章的下载。可以输入要爬取的日期以及结束日期，将这些日期内的文章全部爬取下来，以日期为名自动生成一个主存储目录，爬取到的文章保存写入 txt 文件中,每个文本的存储名字以日期加序号存储。

本程序需要在 python 下，并且需要下载程序依赖的包才能运行。本程序需要用到的包主要有：requests、bs4、os、datetime。

4 总体设计

4.1 程序目录结构设计

程序项目结构非常简单，一个主程序（paweb.py）,还有是根据日期分类的资源总目录，总目录下自动根据日期生成存储文章的目录，再下面是是具体文章的 txt 文本，每个 txt 存储一篇文章。

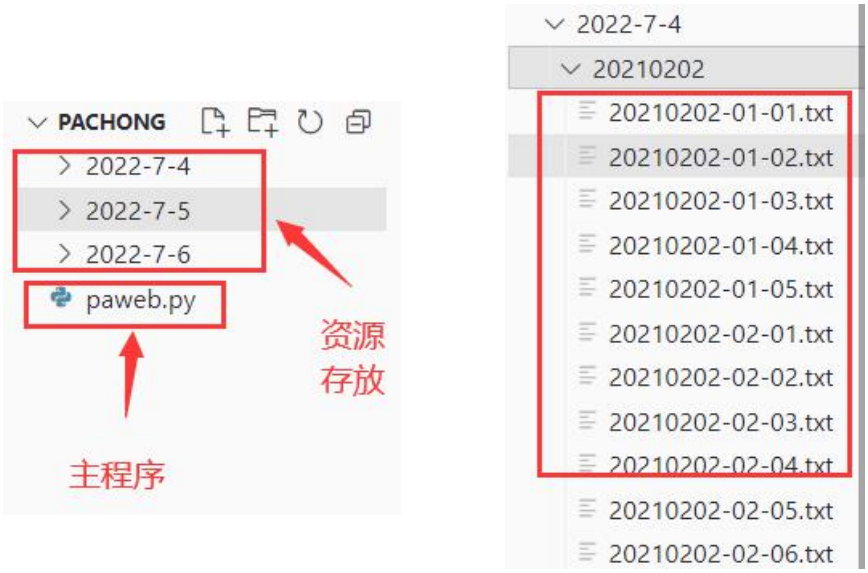

图 4.1 程序结构

4.1 程序总体结构设计

该爬虫程序没有用户界面，基于 python 环境，运行在 Windows PowerShell 窗口中，使用流程为：输入需要爬取的开始日期，结束日期、回车后等待爬取即可，爬取完成后会有提示。工作流程为：根据输入的日期拼接 URL，获取当天报纸的各版面的链接列表，再获取报纸版面的文章链接列表，然后解析 HTML 网页，获取新闻的文章内容，获取到文章标题和正文信息后写入到对用的文件中，最后程序结束运行并提示已经爬取完成。

5 详细设计

5.1 分析目标网站

5.1.1 URL 组成结构

人民日报网站的 URL 的结构比较直观，基本上什么重要的参数，比如日期，版面号，文章编号什么的，都在 URL 中有所体现，构成的规则也很简单，像这样：

版面目录：http://paper.people.com.cn/rmrb/html/2022-07/06/nbs.D110000renmrb_01.htm

文章内容：http://paper.people.com.cn/rmrb/html/2022-07/06/nw.D110000renmrb_20220706_5-01.htm

在版面目录的链接中，“/2022-07/06/”表示日期，后面的 “_01” 表示这是第一版面的链接。在文章内容的链接中，“/2022-07/06/”表示日期，后面的 “_20220706_5_01”表示这是 2019 年 5 月 6 日报纸的第 1 版第 5 篇文章，需要注意的是，在日期的“月”和“日”以及 “版面号”的数字，若小于 10，需在前面补“0”，而文章的篇号则不必。

了解到这个之后，我们可以按照这个规则，构造出任意一天报纸中人一个版面的链接，以及任意一篇文章的链接。

如：2022 年 7 月 3 日第 4 版的目录链接为：

http://paper.people.com.cn/rmrb/html/2022-07/03/nbs.D110000renmrb_01.htm

5.1.2 分析网页 HTML 结构

在 URL 分析中，通过实验发现了网站的页面跳转是通过 URL 的改变完成的，不涉及到 Ajax 这样的动态加载方法。它的所有数据是一开始就加载好的，我们只需要去 html 中提取相应得数据即可。

图 5.1 标题示例

图 5.2 新闻正文示例

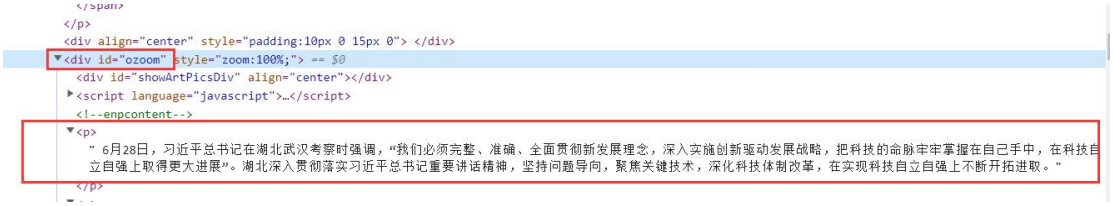

图 5.3 文章具体存放位置

由上图分析后发现，进入文章内容页面之后，文章标题存放在 h1，h2，h3 标签中（有的文章标题只用到了 h1 标签，而有的文章有副标题可能会用到 h2 或 h3 标签），正文部分存放在 id = “ozoom”的 div 标签下的 p 标签里。至此，目标网站的 HTML 页面分析完成。具体爬虫程序分析流程图如下：

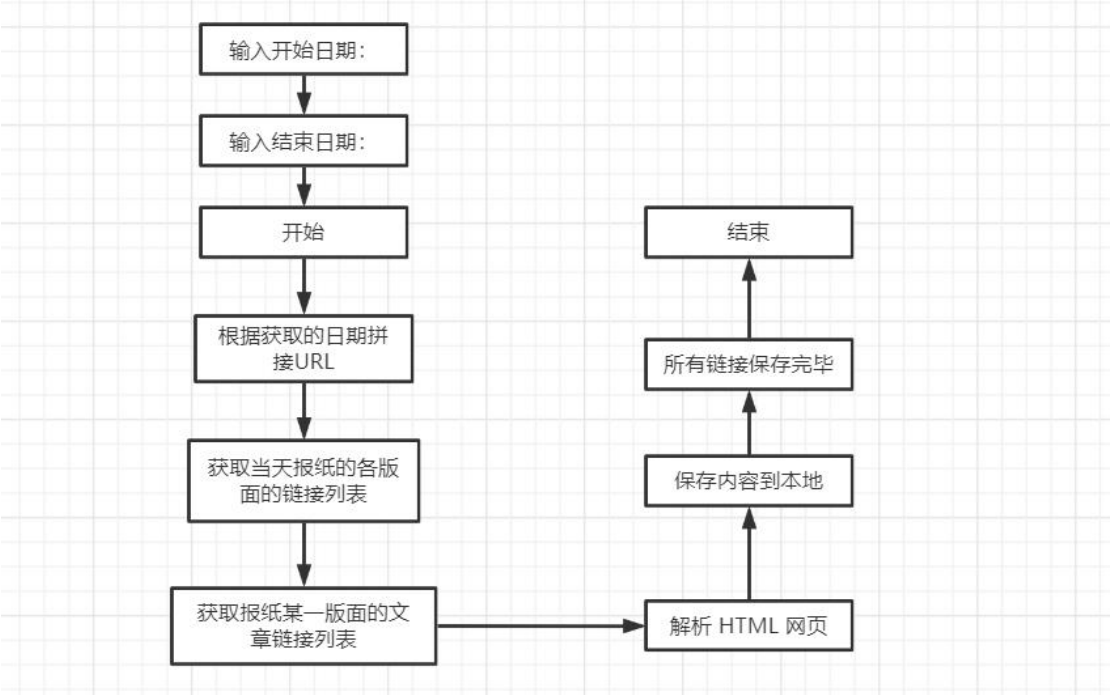

图 5.4 爬虫程序分析流程图

6 数据结构设计描述，各模块（函数）的功能介绍

6.1 数据结构设计描述

程序由输入的开始时间和结束时间来获取要爬取的新闻，由于时间拼接 URL，获取到文章列表后解析 HTML，然后获取到需要的文章标题、文章正文等信息后写入本地。流图如下：

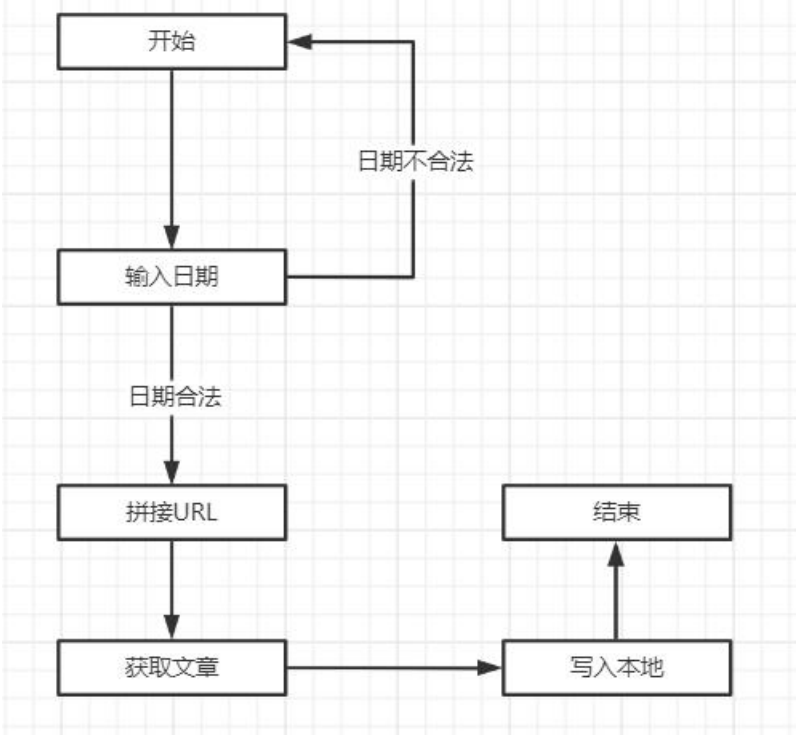

图 6.1 数据结构流图

6.2 主要函数的功能介绍

```
1.def fetchUrl(url):
    headers = {
        'accept':
'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=
0.8',
        'user-agent': 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/68.0.3440.106 Safari/537.36',
    }
    r = requests.get(url,headers=headers)
    r.raise_for_status()
    r.encoding = r.apparent_encoding
    return r.text
```

功能：访问 url 的网页，获取网页内容并返回

参数：目标网页的 url

返回：目标网页的 html 内容

```
2.def getPageList(year, month, day):
```

功能：获取当天报纸的各版面的链接列表

参数：年，月，日，改变年月日拼接成需要爬取的 url

返回当天报纸的各版面的链接列表

3.def getTitleList(year, month, day, pageUrl):

功能：获取报纸某一版面的文章链接列表

参数：年，月，日，该版面的链接

返回文章链接列表

4.def getContent(html):

功能：解析 HTML 网页，获取新闻的文章内容

参数：html 网页内容

返回结果 标题+内容

5.def saveFile(content, path, filename):

功能：将文章内容 content 保存到本地文件中

参数：要保存的内容，路径，文件名

如果没有该文件夹，则自动生成

6.def download_rmrb(year, month, day, destdir):

功能：爬取《人民日报》网站的新闻内容，并保存在指定目录下

参数：年，月，日，文件保存的根目录

然后保存文件到本地

7.def get_date_list(beginDate, endDate):

获取日期列表

param start: 开始日期

param end: 结束日期

```
8.if __name__ == '__main__':
    # 输入起止日期，爬取之间的新闻
    print('---文章爬取系统---')
    beginDate = input('请输入开始日期(格式如 20220706):')
    endDate = input('请输入结束日期(格式如 20220706):')
    data = get_date_list(beginDate, endDate)
    for d in data:
        year = str(d.year)
```

```
        month = str(d.month) if d.month >=10 else '0' + str(d.month)
        day = str(d.day) if d.day >=10 else '0' + str(d.day)
        #爬取后文章统一存到这个文件夹下,没有会自动创建
        destdir = "./2022-7-6"
```

```
        download_rmrb(year, month, day, destdir)
        print('---文章爬取系统---')
        print("爬取文章完成！")
        print('爬取文章时间为：' + year + '/' + month + '/' + day + '的文章已成功写入文件！')
        print('---文章爬取系统---')
```

主函数：程序入口

7 结果分析

7.1 运行结果及分析

1. 开始运行程序，输入爬取文章的开始日期，如图：

图 7.1 输入开始日期

2. 输入爬取文章的结束日期如图

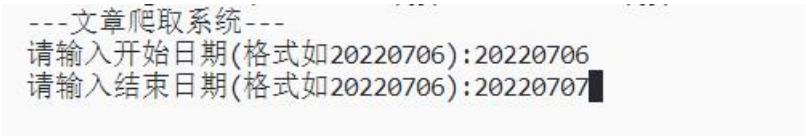

图 7.2 输入结束日期

3. 回车后开始运行程序，如图：

图 7.3 运行图

4. 爬取文章完成后，写入本地，然后会自动结束程序：

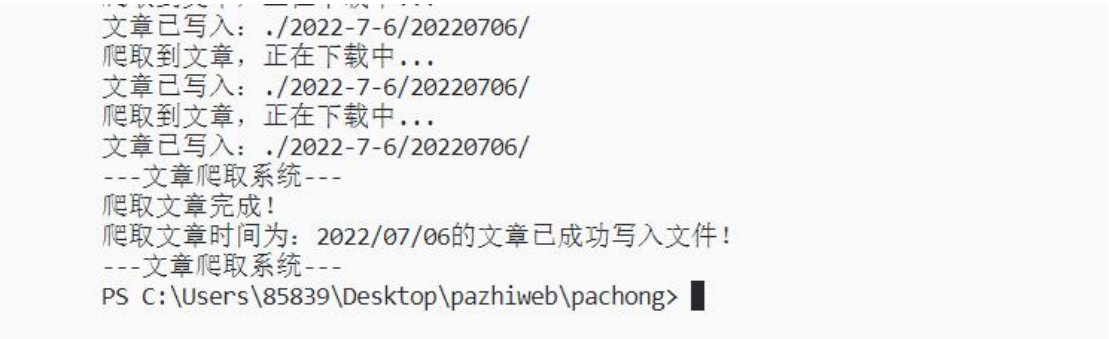

图 7.4 运行结束图

5. 爬取完成后成功写入本地中，每个文章一个 txt：

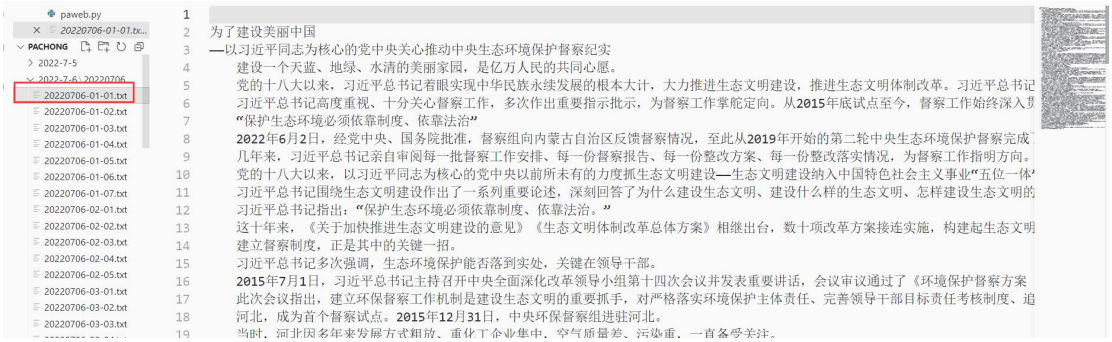

图 7.5 本地 txt

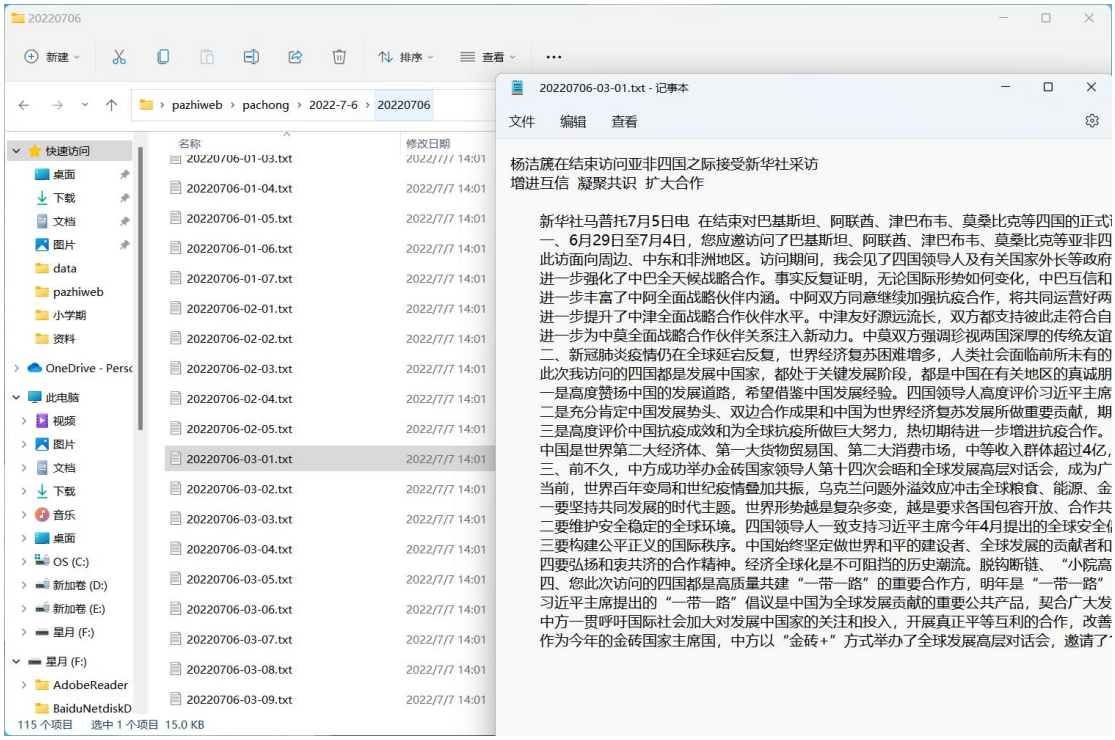

图 7.6 本地 txt 查看

8 总结

从开始这次小学期课程设计到完成这次课程设计一共经历了一周时间。在这一周时间里，一边写程序一边学习，晚上基本都学到半夜，但是当程序功能一点一点距离自己的预期时，就会觉得所有的付出都值了，学到了很多东西，我从对网络爬虫技术一无所知，到能成功的编写出简单的网络爬虫程序，并对网络爬虫中的一些基本技术有了一定的理解。

这次小学期课程设计我完成的是一个基于 python 的爬虫程序，能根据输入的日期来爬取对于日期的《人民日报》的新闻文章，并存到本地，每个文章一个 txt，文件名以日期加序号来命名，文章爬下来结构清晰。

在编写程序过程中，因为之前学过一点 python 基础，但是对于爬虫没有接触，遇到了很多问题，比如怎么爬、爬取后怎么解析、怎么存储数据等。通过百度、csdn、b 站、问同学等，一边学习一边编写，经过 4 天时间终于简单完成了爬取人民日报的文章并存储功能。这些问题的解决都让我受益匪浅。

本次程序的局限性在于没有可视化的操作以及没有用户操作界面，还没有打包让用户能使用，对于这方面的知识还有待学习，本次爬虫选取《人民日报》的一个原因是因为这个网站数据返回是直接到 HTML 页面中，容易解析和获取数据，其他稍微困难的还有点学习，以后往这方面学习研究。最后是在数据存储端方面，我只将数据进行了轻量提取后存了 txt 的文本文件中，并未将数据存入数据库。在数据的处理中，也没有涉及到一些没有用文章的去除以及多次存储后的去重工作等。

参考文献

1. csdn（liu 志军），《python 爬虫的基本原理》
2. Csdn（何极光）https://blog.csdn.net/qq_44034384/article/details/107854112
3. Csdn：https://www.liaoxuefeng.com/wiki/1016959663602400/1017648783851616

2.崔庆才，《python3 网络爬虫开发实战》

3.卢杨，《python 从入门到精通》

4.哔哩哔哩，《黑马程序员爬虫视频》