**网络爬虫**

**实验报告**

姓名： 蔡欣彤

日期： 2023.7.8

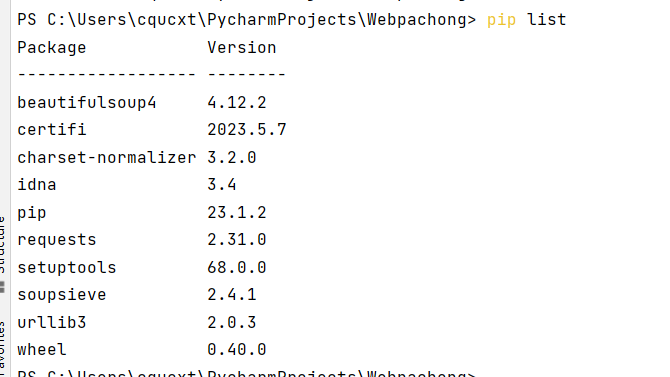
1. **实验目的**
2. 熟悉网络爬虫的工作原理
3. 熟悉网络爬虫的基础语法
4. 能够编写一个简单的网络爬虫
5. **实验内容**

实验要求：

1. 掌握requests模块的安装使用；
2. 掌握BeautifulSoup模块的安装使用；
3. 掌握html的基本知识；
4. **实验结果**

完成实验并截图说明每个步骤

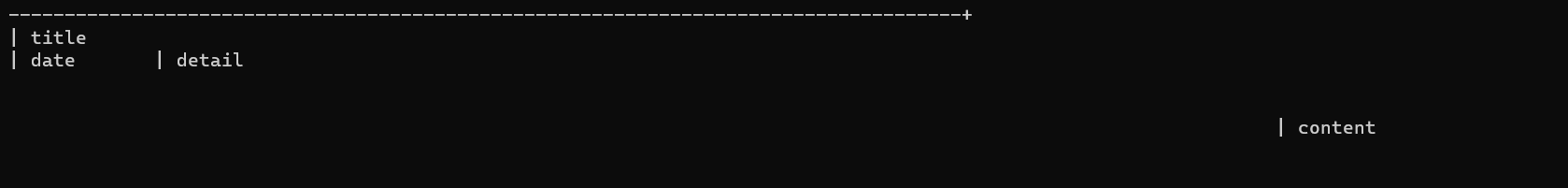
1. **安装requests、**BeautifulSoup4

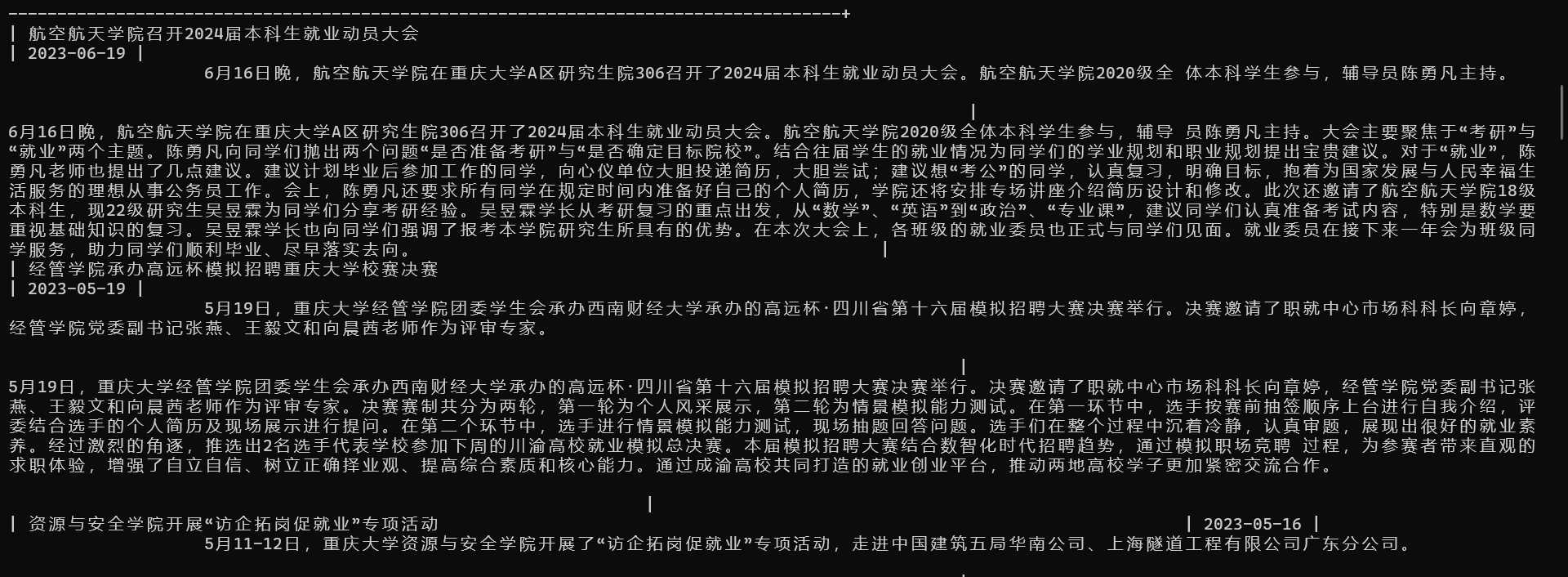
****

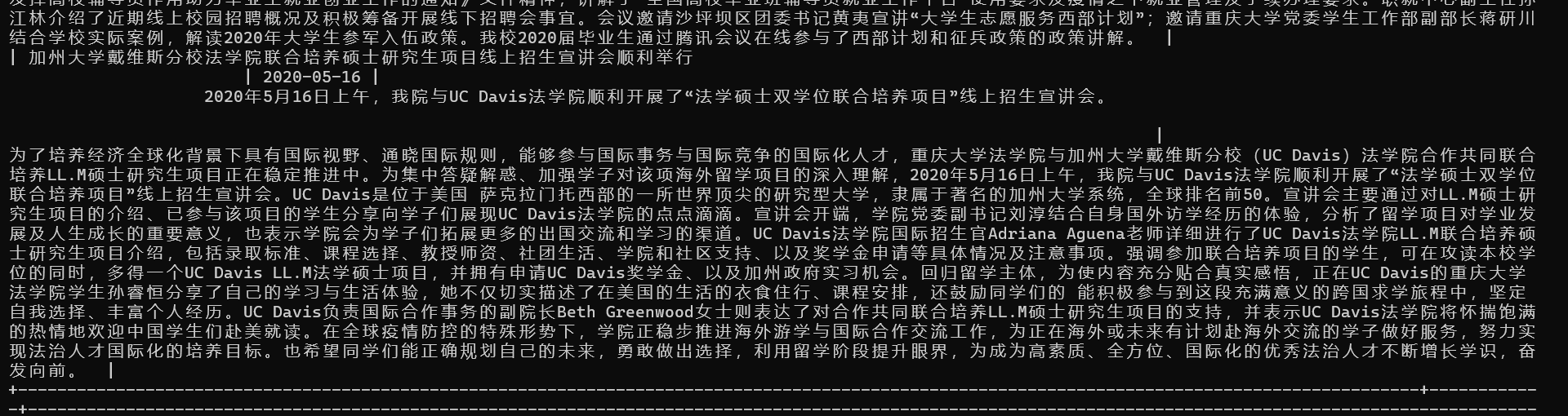
**代码：（使用mysql存储）**

**import** psycopg2 **as** psycopg2  
**import** pymysql **as** pymysql  
**import** requests  
**from** bs4 **import** BeautifulSoup  
**import** sqlite3  
  
**def** wirte\_txt(txt):  
 **with** open(**'news.txt'**,**"w"**,encoding=**'utf-8'**)**as** file:  
 file.write(txt)  
  
**def** wirte\_sqlite(sql):  
 conn = sqlite3.connect(**'news.db'**)  
 cursor=conn.course()  
 cursor.execute(sql)  
 conn.commit()  
 cursor.close()  
  
**def** res(url):  
 headers={  
 **'user-agent'**:**"Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/113.0.0.0 Safari/537.36 Edg/113.0.1774.50"** }  
 res = requests.get(url,headers=headers)  
 res.encoding = res.apparent\_encoding  
 **return** res.text  
  
conn = pymysql.connect(host=**'localhost'**, user=**'root'**, password=**'123456xx'**,db=**'news'**)  
*# conn=psycopg2.connect(database="db\_2020\_01",user="db\_user2020\_271",password="db\_user@123",host="116.205.157.173",port=8000)*b = conn.cursor()  
  
url = **'https://news.cqu.edu.cn/newsv2/news-130.html'**txt = res(url)  
*# print(txt)*soup = BeautifulSoup(txt, **'lxml'**)  
lst = soup.find(**'div'**,class\_=**"lists"**)  
itm = lst.find\_all(**'div'**,class\_=**"item"**)  
*# page=soup.find('div',class\_="page").find\_all('a')[-1]['href']  
# print(page[-1]['href'])  
# i=itm[0]*num=5  
**while** num!=0:  
 *# print(0)* num=num-1  
 **for** i **in** itm:  
 url1 = i.find(**'a'**)[**'href'**]  
 url2=**"https://news.cqu.edu.cn"**+format(url1)  
 *# print(url)* txt1=res(url2)  
 *# print(txt1)* soup1 = BeautifulSoup(txt1, **'lxml'**)  
 title=soup1.find(**'h1'**,class\_=**"dtitle"**).text  
 date=soup1.find(**'div'**,class\_=**"ibox"**).find\_all(**'span'**)[-1].text  
 detail=soup1.find(**'div'**,class\_=**"adetail"**).text  
 content=soup1.find(**'div'**,class\_=**"acontent"**).text  
 *# sql = 'INSERT INTO new (title,date,detail,content) values (%s,%s,%s,%s);'  
 # param = (title, date, detail, content)  
 # print("test1")* b.execute(**"INSERT INTO new (title,date,detail,content) values (%s,%s,%s,%s)"**,(format(title), format(date), format(detail), format(content)))  
 *# print(content)  
 # print("test2")* page = soup.find(**'div'**, class\_=**"page"**).find\_all(**'a'**)[-1][**'href'**]  
 url3=**"https://news.cqu.edu.cn"**+format(page)  
 txt = res(url3)  
 *# print(txt)* soup = BeautifulSoup(txt, **'lxml'**)  
 lst = soup.find(**'div'**, class\_=**"lists"**)  
 itm = lst.find\_all(**'div'**, class\_=**"item"**)  
  
  
 *# print(url3)*conn.commit()  
b.close()  
print(**"ok"**)

**select \* from new**

****

****

****

1. **实验总结**

（在实验中遇到的问题及解决方法、收获是什么）

安装BeautifulSoup遇到问题，改为安装BeautifulSoup4

