湖　北　大　学

计算机与信息工程学院

2021—— 2022 学年度

第 2 学期

学　生　实　验　报　告　册

学生姓名：张林聪

班　　级：软工 2002 班

学　　号：202031119020244

课程名称：大数据分析与应用

任课老师：李洁

学生实验守则

1、学生在规定的时间内进行实验，不得无故缺席或迟到。

2、学生在每次实验前对排定要做的实验应进行预习，并按要求作好预习报告。

3、每次实验前，必须交上次实验报告和本次实验预习报告，并经指导教师提问、检查同意后，才可进行本次实验。

4、学生进入实验室指定位置后，首先根据仪器清单核对自己使用的仪器是否有缺少或损坏，发现问题及时向指导教师报告，严禁擅自动用别组仪器。

5、实验时必须有实事求是、严肃认真的科学态度，严格遵守仪器操作规程和注意事项。

6、实验完毕应将实验数据交给指导教师检查，合格后，整理复原好仪器设备，方可离开实验室。

7、保持实验室肃静和整洁，不得大声喧哗，乱丢垃圾和吃东西。

8、学生在实验过程中，由于不遵守操作规程或未经许可，擅自进行实验而造成事故、损坏仪器设备，应及时报告，并填写损坏清单，按院有关规定进行赔偿。

实 验 报 告 单

实验名称： _____

同组人：　　实验课时：
实验室：　　报告日期：

一、实验目的： 利用网络爬虫进行数据采集

二、实验内容（详细步骤与结果）：

- 导库

```
# -*- coding: utf-8 -*-
    """
    Created on Tue Apr 26 08:21:41 2022
    @author: lenovo
    """
    import requests
    import time
    from bs4 import BeautifulSoup #文本解析库

    print("******\n 爬取中")
    #time.sleep(2)
    print("****\n")
    import requests
```

- 表头和代理 ip

```
#添加 header,伪装成服务器访问
    url='https://www.douban.com'
    headers={'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/
    537.36 (KHTML, like Gecko) Chrome/100.0.4896.127 Safari/537.36',
        }
    get_response = requests.get(url,headers=headers)
```

```
print(get_response.text)
print(get_response.content)

#代理 ip
Proxy_IP = { 'http':'http:\\111.72.126.111:808'}

#后面调佣 requests.get(url,header,proxies) 本实验未用代理 ip
```

- 打开记录文件

```
#打开文件记录
f = open("D:/programming/condaProgram/爬虫/experiment.txt",'a+',encoding = 'utf-8
')
get_time = time.strftime("%Y-%m-%d %H:%M:%S",time.localtime(time.time()))
```

- 设置网址并使用 for 循环写入数据
  1. 分析 url 发现 pn 是帖子数，50 贴为一页
  2. 首先通过连接网址再通过 soup 解析 html
  3. find_all("a",class_="j_th_tit")用来定位指定结点


```
连续爬取的 FOR 循环，以下代码均在该 for 循环中
#设置所爬取的网址
page = [0,50,155600]
for i in page:
    time.sleep(1)
    f.write("[时间： "+get_time+"]\n[标题]湖北大学贴吧第"+str((i+1)/50)+"
页内容"+'\n')
    url = "https://tieba.baidu.com/f?kw=湖北大学&ie=utf-8&tp=0&pn="+ s
tr(i)


    html = requests.get(url)#获取网页信息
```

```
#网页解析
    soup = BeautifulSoup(html.content, "lxml")
    print(soup.prettify())#打印格式

#从网页中提取想要的数据所在的节点
  #这里要注意，find_all 返回的是一个列表
    all = soup.find_all("a",class_="j_th_tit")#获取标题查找标签中的 class_
="j_th_tit"，

    print(all)

 #将</a>作为分隔符去掉
  #str.split 返回的仍然是一个列表，之后是一个字符串列表，内层中括号
此时变成字符了
    ALL=str(all).split('</a>')
    ALL
```

将过滤后

```
 ', <a class="j_th_tit" href="/p/7415586984" rel="noopener" target="_blank" title="夏日炎炎是考研学习的最佳时期～">夏日炎炎
习的最佳时期～',
 ', <a class="j_th_tit" href="/p/7818005246" rel="noopener" target="_blank" title="考研生物与医药">考研生物与医药',
 ', <a class="j_th_tit" href="/p/7823574781" rel="noopener" target="_blank" title="有没有视传的学长学姐啊（我是大一的）">
的学长学姐啊（我是大一的）',
 ', <a class="j_th_tit" href="/p/7823566881" rel="noopener" target="_blank" title="想问个问题">想问个问题',
 ', <a class="j_th_tit" href="/p/7814507554" rel="noopener" target="_blank" title="😭😭😭如何评价">😭😭😭如何评价',
 ', <a class="j_th_tit" href="/p/7823504418" rel="noopener" target="_blank" title="湖大有没有各系学生来出个避雷和推荐老师
没有各系学生来出个避雷和推荐老师啊',
 ', <a class="j_th_tit" href="/p/7822125957" rel="noopener" target="_blank" title="🐨🐨们会做饭吗">🐨🐨们会做饭吗',
 ', <a class="j_th_tit" href="/p/7823406348" rel="noopener" target="_blank" title="悬赏">悬赏',
 ', <a class="j_th_tit" href="/p/7813719744" rel="noopener" target="_blank" title="千言万语一个字:()!">千言万语一个字:()
 ', <a class="j_th_tit" href="/p/7823376786" rel="noopener" target="_blank" title="啦啦啦啦啦啦">啦啦啦啦啦啦',
 ', <a class="j_th_tit" href="/p/7820167367" rel="noopener" target="_blank" title="希望各位8u能严肃思考">希望各位8u能严肃
 ', <a class="j_th_tit" href="/p/7823204478" rel="noopener" target="_blank" title="To讨厌任婧婷的人">To讨厌任婧婷的人',
 ', <a class="j_th_tit" href="/p/7822654493" rel="noopener" target="_blank" title="这是我表弟的学习笔记，能上🐼带吗？">这
的学习笔记，能上🐼带吗？',
 ', <a class="j_th_tit" href="/p/6114342924" rel="noopener" target="_blank" title="计信答辩好像没通过……东西不是自己做的，
题答不上来">计信答辩好像没通过……东西不是自己做的，老师问的问题答不上来',
 ', <a class="j_th_tit" href="/p/7821266171" rel="noopener" target="_blank" title="好的学习氛围多么重要">好的学习氛围多么
 ', <a class="j_th_tit" href="/p/7815140478" rel="noopener" target="_blank" title="我想出去，不想坐牢了">我想出去，不想坐
```

```
#要注意将最后一个中括号 pop 出去，由于最后一个</a>在'】'之前，这时最后一个
中括号被单独当成字符串。
    ALL.pop()
    l=[]
    for s in ALL:
        #这里体现了为什么之前要 pop，不然访问 title 后面的标题会出现 outOfI
```

```
    ndex,也可以通过限制循环次数
            l.append(s.split('title="')[0])
            #print(s.split('title="')[0])
        print(l)

        i=0
        for s in l:
            #再次分片，由于分成两部分，用q接收第一部分，w接收第二部分
            q,w = s.split('">')
            i+=1
            f.write('【标题'+str(i)+'】: '+ q +'\n')

    f.close()
```

- 结果，pn = 50 意味着这是第二页，由于截图大小限制，网页图和 txt 文件都只截取开头一部分

查看 txt 文件

[时间：2022-05-05 21:14:54]
[标题]湖北大学贴吧第2页内容
【标题1】：这都能刷到，输麻了！
【标题2】：校外租房的拿通行证就进吗？还是要在网上申请才能刷卡呀
【标题3】：考研上岸了
【标题4】：物理院电子信息考研
【标题5】：输麻了🐱带
【标题6】：今日分享一首歌
【标题7】：🐱带郭♀魅力时刻!
【标题8】：各位辅导员有没有说让暑假不让留校的问题?
【标题9】：考研湖北大学
【标题10】：完了，这下输了

正确

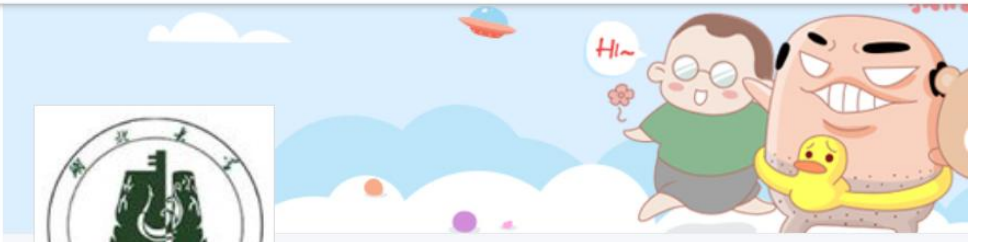
以下为第三页和最后一页

[时间：2022-05-05 21:14:54]
[标题]湖北大学贴吧第3页内容
【标题1】：远离渣👦👌
【标题2】：如何评价发表白墙酸别人谈恋爱的这位同学
【标题3】：湖北大学哲学考研资料
【标题4】：湖带就没有这种瓜吗
【标题5】：如何评价表白墙上的这位据理力争的香水狂魔大佬
【标题6】：湖带的家人们引以为戒！

最后一页

https://tieba.baidu.com/f?kw=湖北大学&ie=utf-8&tp=0&pn=155600

Baidu 贴吧 湖北大学



湖北大学吧 + 关注 关注：151,337 贴子：4,316,702

沙湖之滨，长江之畔，一环中最闪亮那枚星。 目录： 华南地区高等

看贴 图片 吧主推荐 视频

5	置顶 精 五、六月汇总贴（推-广/兼-职/租-房/发布贴）	香草酚蓝
9	学校外面有地方能做核酸吗，有没有兄弟知道的可否吱一声	qqqqq
		qqqqq
13	输😭😭😭😭	跟着二崽...
	🐰🐰们我们怎么没有 🦊带 我的🦊带😭 ✋🈚你坏事做尽😭😭	跟着二崽...

时间：2022-05-05 21:20:55]
标题]湖北大学贴吧第3113页内容
【标题1】：五、六月汇总贴（推-广/兼-职/租-房/发布贴）
【标题2】：学校外面有地方能做核酸吗，有没有兄弟知道的可否吱一声
【标题3】：输😭😭😭😭
【标题4】：🐭🐭们会做饭吗
【标题5】：好的学习氛围多么重要
【标题6】：大声说出那个字！
【标题7】：友友们湖大校内有电动车充电桩吗？
【标题8】：湖北大学公共管理上岸啦，出自用复习资料
【标题9】：新传专硕拟录取啦，便宜出自用复习资料

三、心得体会：

本次实验体会了 爬虫的基本流程，获取 url,解析网页，Beautiful Soup 的使用，正则表达式的使用，split 去除多余信息，最后写入文件；必要时使用代理 ip 和添加 Header 伪装成浏览器。

一开始一步一步跟着 ppt 上的例子做，去体会每一步的这样做的理由，过程中发现有些功能搞不懂或不知道怎么实现，通过百度最后弄懂，最终程序成型。

成绩：

批阅教师：

日 期：