云南大学旅游文化学院信息学院

实验报告

课程名称：数据分析程序语言设计　实验名称：实验 4 爬虫入门

序号 10　学号 20201200737　专业班级：计算机科学与技术一班

第 4 次实验

姓名 平德祥　成绩　代码行统计：共 27 行代码

一、实验目的

（1）熟练掌握简单爬虫的步骤
（2）爬取简单的网页数据

二、实验环境

python3

三、实验原理

综合实验中用到的函数和点方法，复习填写在此部分

1. urllib.request 模块定义了有助于在复杂环境中打开 URL 的函数和类-基本身份验证和摘要身份验证，重定向，Cookie 等。(.urlopen（url）.read)
2. find_all()
   查找标签 soup.find_all('tag')
   查找文本 soup.find_all(text='text')
   根据 id 查找 soup.find_all(id='tag id')
   使用正则 soup.find_all(text=re.compile('your re')), soup.find_all(id=re.compile('your re'))
   指定属性查找标签 soup.find_all('tag', {'id':'tag id','class':'tag class'})
3. Pd.DataFrame,类似于 json 的创建格式

四、实验步骤

1、爬取百度热搜榜相关数据。

网站地址为：https://top.baidu.com/board?tab=novel

读取网页的相关数据，爬取小说的名称、作者、类型、热搜指数。

```
import pandas as pd
from bs4 import BeautifulSoup
import urllib
```

```
NovelUrl = r"https://top.baidu.com/board?tab=novel"
htmlNovel = urllib.request.urlopen(NovelUrl).read()
soup = BeautifulSoup(htmlNovel, "html.parser")
hotNovelData1 = soup.find_all(class_="c-single-text-ellipsis")
hotNovelData2 = soup.find_all(class_="intro_1l0wp")
hotNovelData3 = soup.find_all(class_="hot-index_1Bl1a")
NovelName = [i.text.strip() for i in hotNovelData1[::2]]
NovelAuthor = [i.text.strip().replace("作者：", "") for i in hotNovelData2[::2]]
NovelType = [i.text.strip().replace("类型：", "") for i in hotNovelData2[1::2]]
NovelIndex = [i.text.strip().strip() for i in hotNovelData3]

data = pd.DataFrame({
    "名称": NovelName,
    "作者": NovelAuthor,
    "类型": NovelType,
    "热搜指数": NovelIndex
})
data.to_excel(r"E:\python\file\baidu_novel.xlsx", index=False)
```

2、将小说的名称、作者、类型、热搜指数整理成 dataframe 格式，保存为 baidu_novel.xlsx。
（保存 excel 的结果截图显示）

	名称	作者	类型	热搜指数
2	少年歌行	周木楠	历史	463883
3	万古神帝	飞天鱼	玄幻	373974
4	深空彼岸	辰东	都市	257676
5	剑来	烽火戏诸	玄幻	251794
6	传奇	墨舞碧歌	古代言情	246912
7	万相之王	天蚕土豆	玄幻	243534
8	我有一剑	青鸾峰上	玄幻	164207
9	将军	香无	历史	142179
10	宇宙职业	我吃西红	科幻	135741
11	斗罗大陆	唐家三少	玄幻	117733
12	九星霸体	平凡魔术	玄幻	113285
13	执掌风云	笔龙胆	都市	111380
14	修罗武神	善良的蜜	玄幻	103034
15	雪中悍刀	烽火戏诸	玄幻	102489
16	皇子	清风闲人	历史	96744
17	王妃	妖谈	古代言情	87468
18	种田	囚鸽	古代言情	82684
19	唐门	卧枕江山	武侠	81038
20	庆余年	猫腻	历史	71210
21	斗破苍穹	天蚕土豆	玄幻	68793
22	飞天	跃千愁	玄幻	66691
23	学霸	星光依旧	青春	66001
24	混沌剑神	心星逍遥	玄幻	64354
25	归来	dp	都市	61756
26	人道大圣	莫默	玄幻	58318
27	女帝	安步	幻想言情	57586
28	上门龙婿	叶公子	都市	57414
29	宠妻	暗恋彼岸	现代言情	57362
30	诛仙	萧鼎	玄幻	56904
31	十年	乔艺	青春	55745

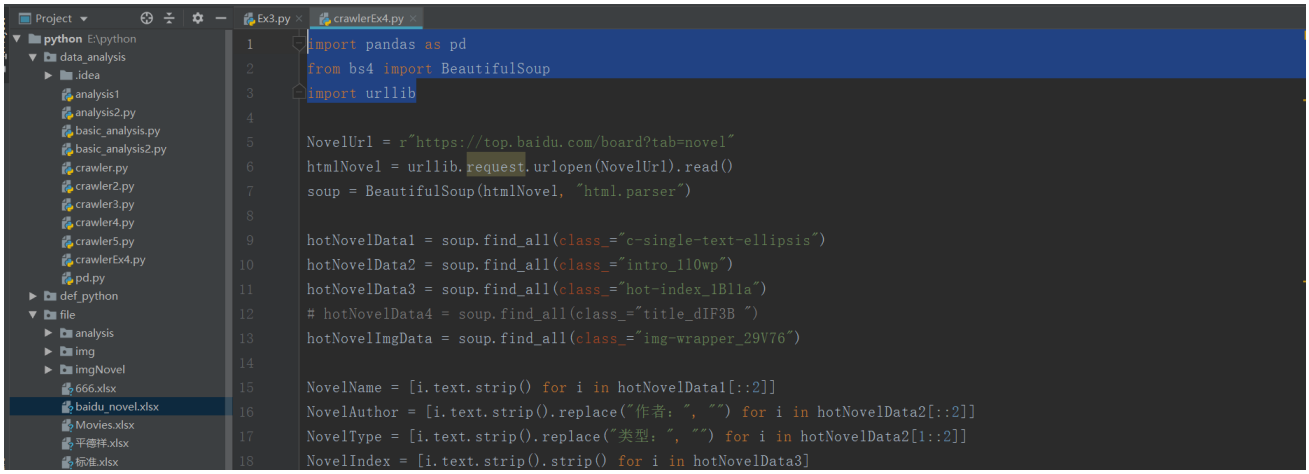
3、保存小说所有的封面图片，保存完成后，截图显示代码结果。

```
hotNovelImgData = soup.find_all(class_="img-wrapper_29V76")
NovelImg = [i.img.attrs['src'] for i in hotNovelImgData]

for i in range(len(NovelImg)):
    ImgHtml = urllib.request.urlopen(NovelImg[i]).read()
    with open(r'E:\\python\\file\\imgNovel' + f'\\{NovelName[i]}' + ".jpg", "wb+")
as f:
        f.write(ImgHtml)
    print(NovelName[i])
```

4、将该流程模块化。

```
import pandas as pd
from bs4 import BeautifulSoup
import urllib
```



五、结果分析

了解了基本的爬虫，也是熟悉了百度热搜的爬虫，按 class 标签即可爬取，适合入门

对于模块化有 spring 的基础所以比较容易实现，主要难点在于字符串的处理和图片的存储，字符串处理得随机应变，正则，切割，替代等等都可以实现