第 一 次 实 验 报 告

课程名称	内容安全实验				
学生姓名	陈曦	学号	2020302181081	指导老师	张典
专业	网络安全	班级	2020 级 3 班	实验时间	2023.3.13

一、实验内容

1.爬取武汉大学官方网站，含“樱”的新闻链接标题。
2.爬取百度热搜电影榜单。
3.爬取中国大学排行榜 2021。
4.爬取百度热搜任一榜单，搜索结果按顺序逐行输出。

二、实验原理

使用 Python 中的 BeautifulSoup 库，requests 库，以及 re 库，爬取目的网址数据。

1. BeautifulSoup 库

BeautifulSoup 库提供一些简单的，python 式的函数来处理导航，搜索，修改分析树等工具。它是一个工具箱，通过解析文档为程序员提供需要抓取的数据。它也是一个 HTML/XML 的解析器，主要功能就是解析和提取 HTML/XML 数据。

2. Requests 库

Requests 库是根据 python 内置库 urllib 的重写。它可以请求一个 response 对象，也就是网站的响应，程序员可以查看状态码，内容以及 cookie 等。

3. 基本思路

导入工具包后，获取目的网址的文本内容，并通过循环，函数等方式提取到想要的数据内容，最后再使用循环的方式将数据内容打印到终端。

三、实验步骤

1. 爬取武汉大学官网中含“樱”的新闻链接标题

导入工具库：BeautifulSoup，requests，re。

```
import requests
from bs4 import BeautifulSoup
import re
```

使用 requests 工具库中的 get 获取武汉大学官网。

```
r = requests.get("http://www.whu.edu.cn/")
```

设置字符编码防止乱码。

```
r.encoding = r.apparent_encoding
```

爬取网页内容。

```
demo = r.text
```

查找含“樱”字符的新闻链接标题，并打印。

```
soup = BeautifulSoup(demo, 'html.parser')
for tag in soup.find_all('a', string=re.compile('樱')):
    print(tag.string)
```

爬取的全部代码：

```
import requests
from bs4 import BeautifulSoup
import re

r = requests.get("http://www.whu.edu.cn/")
r.encoding = r.apparent_encoding
demo = r.text
soup = BeautifulSoup(demo, 'html.parser')
for tag in soup.find_all('a', string=re.compile('樱')):
    print(tag.string)
```

2. 爬取百度热搜电影榜单

导入工具包 BeautifulSoup，以及 requests。

```
import requests
from bs4 import BeautifulSoup
```

使用 requests 工具库中的 get 获取百度电影榜单。

```
r = requests.get("https://top.baidu.com/board?tab=movie")
```

设置字符编码防止乱码。

```
r.encoding = r.apparent_encoding
```

爬取网页内容。

```
demo = r.text
```

使用解析器解析该内容。

```
soup = BeautifulSoup(demo, 'html.parser')
```

设置列表，存放电影名。

```
ulist = []
```

打印表格头，并且识别榜单内容。

```
print('序号\t片名')
it = iter(soup.find_all('div', 'c-single-text-ellipsis'))
```

使用 for 循环，打印标号和电影名。

```
for tag in it:
    ulist.append(tag.string)
    print(ulist.index(tag.string) + 1, ulist[ulist.index(tag.string)])
    next(it)
```

故爬取代码为：

```
import requests
from bs4 import BeautifulSoup

r = requests.get("https://top.baidu.com/board?tab=movie")
r.encoding = r.apparent_encoding
demo = r.text
soup = BeautifulSoup(demo, 'html.parser')
ulist = []
print('序号\t片名')
it = iter(soup.find_all('div', 'c-single-text-ellipsis'))
for tag in it:
    ulist.append(tag.string)
    print(ulist.index(tag.string) + 1, ulist[ulist.index(tag.string)])
    next(it)
```

3. 爬取中国大学排行榜 2021

导入工具包 requests，BeautifulSoup，re。

```
import requests
from bs4 import BeautifulSoup as bs
import re
import bs4
```

设置大学排名内容存放的列表 uinfo。设置爬取网址：2021 年大学排名网址。

```
if __name__=='__main__':
    uinfo = []
    url = 'https://www.shanghairanking.cn/rankings/bcur/2021'
```

使用获取网页 HTML 文本函数获取网页内容。

```
html = getHTMLText(url)
```

定义该函数的目的是将我们需要的关键信息提取到一个列表中。即将一个 html 页面放到一个 list 列表中。

```
fillUnivList(uinfo, html)
```

使用打印函数打印排名前二十的大学。

```
printUnivList(uinfo, 20)
```

程序此模块编写完成。

```
if __name__=='__main__':
    uinfo = []
    url = 'https://www.shanghairanking.cn/rankings/bcur/2021'
    html = getHTMLText(url)
    fillUnivList(uinfo, html)
    printUnivList(uinfo, 20)
```

定义函数 getHTMLText 获取网页的 html 文本。储存解析后的文本。设置超时值，设置字符编码，并返回 html 文本。

```
def getHTMLText(url):
    try:
        kv = {
            'user-agent': 'Mozilla/5.0'
        }
        r = requests.get(url, headers=kv, timeout = 30)
        r.raise_for_status()
        r.encoding = r.apparent_encoding
        return r.text
    except:
        return ""
```

定义关键信息提取函数。也使用 for 循环将大学相关信息提取到列表中。

```
def fillUnivList(ulist, html):
    soup = bs(html, "html.parser")
    for tr in soup.find('tbody').children:
        if isinstance(tr, bs4.element.Tag):
            tds = tr('td')
            ulist.append([tds[0].text.replace('\n', '').replace(' ', ''), re.split(' ', tds[1].text)[1], re.search(r'\w{2}', tds[2].text).
                          group(0), re.search(r'\w{2}', tds[3].text).group(0), tds[4].text.replace('\n', '').replace(' ', '')])
```

定义打印函数。使用 for 循环打印大学的排名号，学校名称，省市，类型和总分。

```
def printUnivList(ulist, num):
    print("{:^10}\t{:^10}\t{:^6}\t{:^6}\t{:^10}".format("排名", "学校名称", "省市", "类型", "总分"))
    for i in range(num):
        u = ulist[i]
        print("{:^10}\t{:^10}\t{:^6}\t{:^6}\t{:^10}".format(u[0], u[1], u[2], u[3], u[4]))
```

4. 爬取百度热搜游戏榜单

导入工具包 BeautifulSoup，以及 requests。

```
import requests
from bs4 import BeautifulSoup
```

使用 requests 工具库中的 get 获取百度游戏榜单。

```
r = requests.get("https://top.baidu.com/board?tab=game")
```

设置字符编码防止乱码。

```
r.encoding = r.apparent_encoding
```

爬取网页内容。

```
demo = r.text
```

使用解析器解析该内容。

```
soup = BeautifulSoup(demo, 'html.parser')
```

设置列表，存放游戏名。

```
ulist = []
```

打印表格头，并且识别榜单内容。

```
print('序号\t游戏')
it = iter(soup.find_all('div', 'c-single-text-ellipsis'))
```

使用 for 循环，打印标号和电影名。

```
for tag in it:
    ulist.append(tag.string)
    print(ulist.index(tag.string) + 1, ulist[ulist.index(tag.string)])
    next(it)
```

故爬取代码为：

```python
import requests
from bs4 import BeautifulSoup

r = requests.get("https://top.baidu.com/board?tab=game")
r.encoding = r.apparent_encoding
demo = r.text
soup = BeautifulSoup(demo, 'html.parser')
ulist = []
print('序号\t游戏')
it = iter(soup.find_all('div', 'c-single-text-ellipsis'))
for tag in it:
    ulist.append(tag.string)
    print(ulist.index(tag.string) + 1, ulist[ulist.index(tag.string)])
    next(it)
```

四、实验结果

1. 武汉大学官网含“樱”标题链接爬取结果

```
D:\chenxi\try\ide\anaconda\python.exe D:/chenxi/code/wormtest/main2.py
武汉大学关于加强2023年樱花开放期间校园管理的通告
2023年樱花开放期间校园管理问答录
关于加强樱花开放期间校园安全工作的通知
2023年樱花开放期间校园交通管理的通知
2023赏樱预约通道

进程已结束,退出代码0
```

2. 百度热搜电影榜单爬取结果

```
D:\chenxi\try\ide\anaconda\python.exe D:/chenxi/code/wormtest/main3.py
序号 片名
1    流浪地球2
2    满江红
3    正义回廊
4    黑豹2
5    万里归途
6    白雪公主
7    交换人生
8    中国乒乓
9    东北告别天团
10    天龙八部之乔峰传
```

```
11    回廊亭
12    疾速追杀4
13    奥黛丽
14    断网
15    哥，你好
16    魔女2
17    平凡英雄
18    诛烬枭亡
19    人生大事
20    独行月球
```

```
21    非凡营救
22    触不可及
23    b级文件
24    速度与激情10
25    扫黑行动
26    掮客
27    新·奥特曼
28    寒战3
29    追击
30    想见你

进程已结束,退出代码0
```

3.中国大学排行榜爬取结果

```
D:\chenxi\try\ide\anaconda\python.exe D:/chenxi/code/wormtest/main.py
    排名          学校名称          省市      类型          总分
    1           清华大学          北京      综合       969.2
    2           北京大学          北京      综合       855.3
    3           浙江大学          浙江      综合       768.7
    4          上海交通大学         上海      综合       723.4
    5           南京大学          江苏      综合       654.8
    6           复旦大学          上海      综合       649.7
    7         中国科学技术大学        安徽      理工       577.0
    8          华中科技大学         湖北      综合       574.3
    9           武汉大学          湖北      综合       567.9
    10         西安交通大学         陕西      综合       537.9
    11        哈尔滨工业大学         黑龙      理工       522.6
    12          中山大学          广东      综合       519.3
    13         北京师范大学         北京      师范       518.3
    14          四川大学          四川      综合       516.6
    15        北京航空航天大学        北京      理工       513.8
    16          同济大学          上海      理工       508.3
    17          东南大学          江苏      综合       488.1
    18         中国人民大学         北京      综合       487.8
    19         北京理工大学         北京      理工       474.0
    20          南开大学          天津      综合       465.3

进程已结束,退出代码0
```

4.爬取百度热搜任一榜单，搜索结果按顺序逐行输出。

```
D:\chenxi\try\ide\anaconda\python.exe D:/chenxi/code/wormtest/main3.py
序号 游戏
1    原神
2    王者荣耀
3    和平精英
4    蛋仔派对
5    明日方舟
6    植物大战僵尸
7    我的世界
8    迷你世界
9    香肠派对
10    英雄联盟
11    地铁跑酷
12    暗区突围
13    球球大作战
14    元气骑士
15    复古传奇
16    光遇
17    三国志战略版
18    梦幻西游
19    第五人格
20    坦克世界
21    永劫无间
22    三国杀
23    金铲铲之战
24    绝地求生
25    崩坏3
26    艾尔登法环
27    荒野大镖客2
28    阴阳师
29    部落冲突
30    使命召唤

进程已结束,退出代码0
```

五．实验心得

通过这个实验我学习了爬取网站的技能，了解了如何使用 requests 库，BeautifulSoup 库的基本使用方法，并且爬取了想要爬到的数据。初步学习了爬取 HTML 页面后的数据处理方法，并打印在终端上。