第十四周Python实验报告（周三）

**学号**：117060400110 **姓名**：蒙柳双

**班级**：17应用统计学1班 **指导老师**：林卫中老师

**实验目的：**

1. 了解并掌握获取网页的方法（requests库的get）
2. 会运用BeautifulSoup库将获取的网页进行处理、分析
3. 还要会阅读网页的源代码当中代表的意思
4. 学会在源代码当中找到自己需要的资料的标签（获取标签）
5. 了解什么是正则表达式

**实验要求**：

1. 上课不玩手机
2. 要认真编码并且思考

实在不会的就向老师求助

**实验内容步骤：**

**1.程序练习10-1中国大学排名爬虫**

**核心代码：**

import requests

from bs4 import BeautifulSoup

allUniv = []

def getHTMLText(url):

try:

r = requests.get(url, timeout=30)

r.raise\_for\_status()

r.encoding = 'utf-8'

return r.text

except:

return ""

def fillUnivList(soup):

data = soup.find\_all('tr')

for tr in data:

ltd = tr.find\_all('td')

if len(ltd)==0:

continue

singleUniv = []

for td in ltd:

singleUniv.append(td.string)

allUniv.append(singleUniv)

def printUnivList(num):

print("{:^4}{:^10}{:^5}{:^8}{:^10}".format("排名","学校名称","省市","总分","培养规模"))

for i in range(num):

u=allUniv[i]

print("{:^4}{:^10}{:^5}{:^8}{:^10}".format(u[0],u[1],u[2],u[3],u[6]))

def main():

url = 'http://www.zuihaodaxue.cn/zuihaodaxuepaiming2016.html'

html = getHTMLText(url)

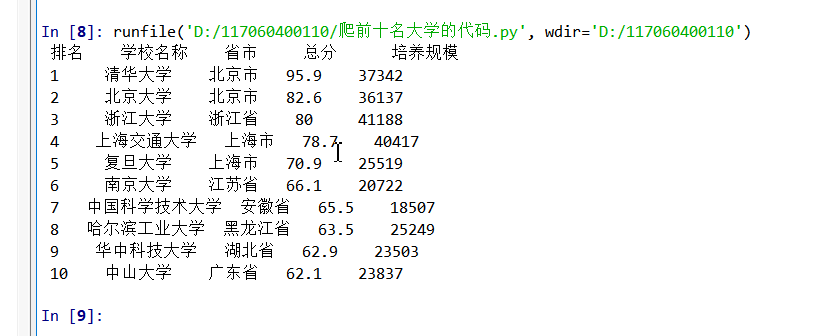
soup = BeautifulSoup(html, "html.parser")

fillUnivList(soup)

printUnivList(10)

main()

**实验结果：**



**2.程序练习10-2按照省份输出中国大学排名。**大学排名网址：

<http://www.zuihaodaxue.cn/zuihaodaxuepaiming2018.html>

请分别输出江西省和山东省的高校排名

**（1）山东省**

**核心代码：**

import requests

from bs4 import BeautifulSoup

allUniv = []

def getHTMLText(url):

try:

r = requests.get(url, timeout=30)

r.raise\_for\_status()

r.encoding = 'utf-8'

return r.text

except:

return ""

def fillUnivList(soup):

data = soup.find\_all('tr')

for tr in data:

ltd = tr.find\_all('td')

if len(ltd)==0:

continue

singleUniv = []

for td in ltd:

singleUniv.append(td.string)

allUniv.append(singleUniv)

def printUnivList(province):

print("{:^4}{:^10}{:^5}{:^8}{:^10}".format("排名","学校名称","省市","总分","培养规模"))

for u in allUniv:

if province in u[2]:

#if float(u[3])>50:

print("{:^4}{:^10}{:^5}{:^8}{:^10}".format(u[0],u[1],u[2],u[3],u[6]))

def main(province):

url = 'http://www.zuihaodaxue.cn/zuihaodaxuepaiming2018.html'

html = getHTMLText(url)

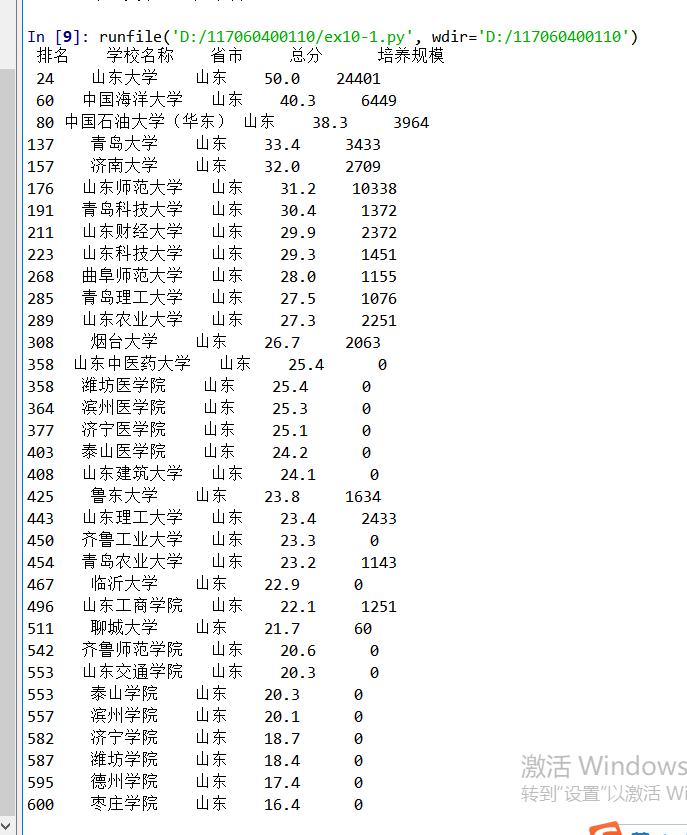
soup = BeautifulSoup(html, "html.parser")

fillUnivList(soup)

printUnivList(province)

main('山东')

**实验结果：**



1. **江西省**

**核心代码：**

import requests

from bs4 import BeautifulSoup

allUniv = []

def getHTMLText(url):

try:

r = requests.get(url, timeout=30)

r.raise\_for\_status()

r.encoding = 'utf-8'

return r.text

except:

return ""

def fillUnivList(soup):

data = soup.find\_all('tr')

for tr in data:

ltd = tr.find\_all('td')

if len(ltd)==0:

continue

singleUniv = []

for td in ltd:

singleUniv.append(td.string)

allUniv.append(singleUniv)

def printUnivList(province):

print("{:^4}{:^10}{:^5}{:^8}{:^10}".format("排名","学校名称","省市","总分","培养规模"))

for u in allUniv:

if province in u[2]:

#if float(u[3])>50:

print("{:^4}{:^10}{:^5}{:^8}{:^10}".format(u[0],u[1],u[2],u[3],u[6]))

def main(province):

url = 'http://www.zuihaodaxue.cn/zuihaodaxuepaiming2018.html'

html = getHTMLText(url)

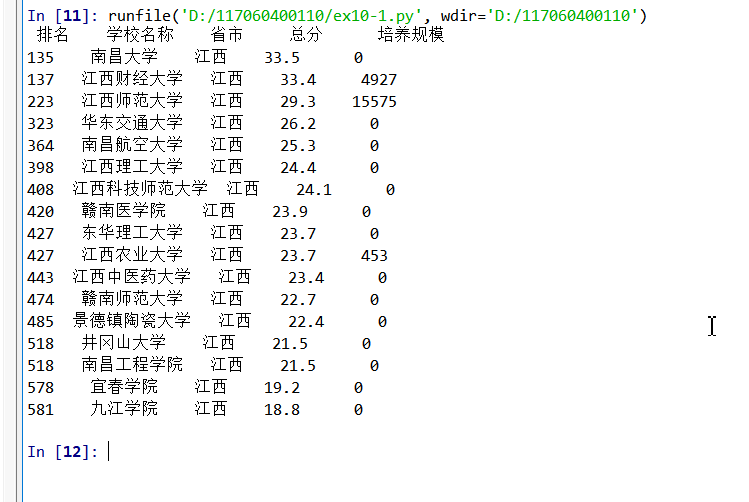
soup = BeautifulSoup(html, "html.parser")

fillUnivList(soup)

printUnivList(province)

main('江西')

**实验结果：**



**2. USNEWS美国大学排名爬虫。**美国大学排名网址如下：

<https://www.usnews.com/best-colleges/rankings/national-universities>

**核心代码：**

import requests

import re

from bs4 import BeautifulSoup

allUniv=[]

def getHTMLText(url):

send\_headers = {

"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36",

"Connection": "keep-alive",

"Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,\*/\*;q=0.8",

"Accept-Language": "zh-CN,zh;q=0.8"}

try:

r = requests.get(url, headers=send\_headers)

r.raise\_for\_status()

print(r.status\_code)

r.encoding = 'utf-8'

return r.text

except:

return ""

def fillUnivList(soup):

data = soup.find\_all('div',{'class':re.compile('shadow-dark')})

for div in data:

singleUniv = []

#排名

div1 = div.find('div',{'style':'margin-left: 2.5rem;'})

rank = div1.get\_text().strip()

#学校名称

singleUniv.append(rank.split(' ')[0])

div2 = div.find('h3')

#City

singleUniv.append(div2.get\_text().strip())

address = div.find('div',{'class':re.compile('block-normal')})

singleUniv.append(address.string)

lstrong = div.find\_all('strong')

singleUniv.append(lstrong[0].string)#学费

singleUniv.append(lstrong[1].string)#培养规模

allUniv.append(singleUniv)

def printUnivList():

print("{:<6}{:<20}{:<10}{:<10}{:<10}".format("排名","学校名称","City","学费","培养规模"))

for u in allUniv:

print("{:<6}{:<20}{:<10}{:<10}{:<10}".format(u[0],u[1],u[2],u[3],u[4]))

def main():

url = 'https://www.usnews.com/best-colleges/rankings/national-universities'

html = getHTMLText(url)

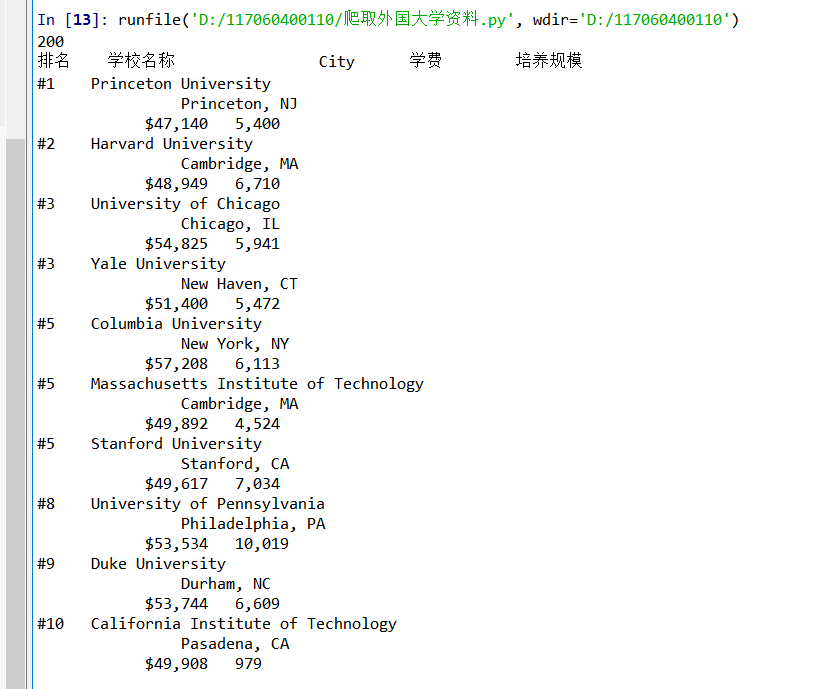
soup = BeautifulSoup(html,'html.parser')

fillUnivList(soup)

printUnivList()

main()

**实验结果：**



1. **程序练习10-6分析百度图片搜索返回结果的HTML代码，编写爬虫抓取图片并下载形成专题图片库。**

**核心代码：**

import requests

import re#引用正则表达式的库

def getHTMLText(url,coding='gbk'):

try:

r = requests.get(url,timeout=30)

print(r)

r.raise\_for\_status()

r.encoding = coding

return r.text

except:

return ""

def downloadImageFile(imgUrl, destUrl, fname=''):

local\_filename = imgUrl.split('/')[-1]

print('Download Image File={}'.format(local\_filename))

try:

r = requests.get(imgUrl, stream=True)

r.raise\_for\_status()

if len(fname) == 0:

fname = local\_filename

print('fname={}'.format(fname))

with open(destUrl + "/" + fname, 'wb') as f:

for chunk in r.iter\_content(chunk\_size=1024):

if chunk:

f.write(chunk)

f.flush()

f.close()

return r.status\_code

except:

return r.status\_code

def getImg(html):

imgre = re.compile('"objURL":"(.\*?)"')

imglist = re.findall(imgre,html)

return imglist

def download(urls,path):

index = 1

for url in urls:

print("Download Image from page:{}".format(url))

status = downloadImageFile(url,path,str(index)+".jpg")

try:

if str(status)[0] == '4':

print("未下载成功{}".format(url))

continue

except Exception as e:

print("未下载成功{}".format(url))

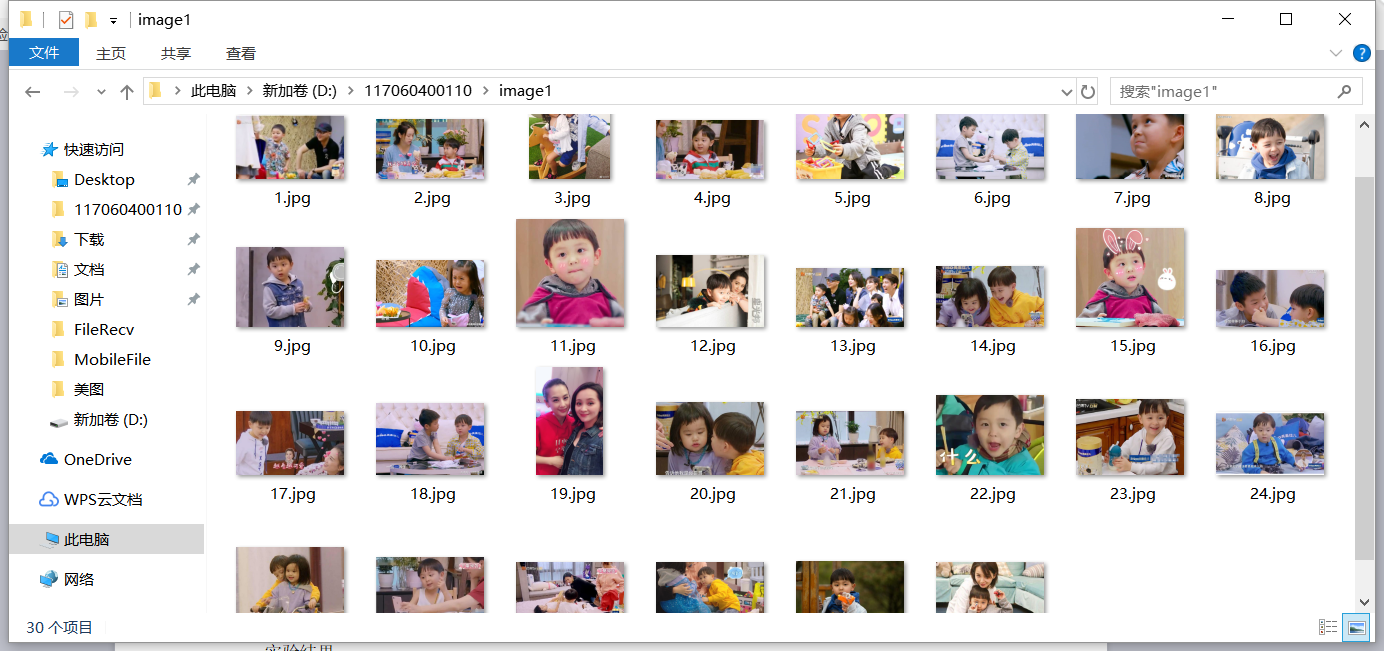
index += 1

page = 'http://image.baidu.com/search/index?tn=baiduimage&word=大麟子'

html= getHTMLText(page,'utf-8')

download(getImg(html),'d:\\117060400110\\image1')

**实验结果：**



**实验总结：**

1. 看着这些代码，发现以自己的能力是写不出这样的代码，而且自己的动手写代码写得少吧，发现自己能把它看懂就好了
2. 通过这些实验，自己也更加的可以通过源代码去了解一些网页的东西以及如何去获取，找到自己想要的东西在那些标签里面。