第十三周Python实验报告（周五）

**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. 今天是对requests库的了解、熟悉与使用
2. 对beautifulsoup库的理解与使用
3. 对网络爬虫的一些小应用

**实验要求：**

1. 上课不玩手机
2. 要认真编码并且思考
3. 实在不会的就向老师求助

**实验内容步骤：**

1. **通过访问http://docs.python-requests.org对request库有更深的了解**

**（1）核心代码**：

import requests

def getHTMLText(url,coding = 'utf-8'):

try:

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

print(r)

r.raise\_for\_status() #如果状态不是200，引发异常

r.encoding = 'utf-8' #无论原来用什么编码，都改成utf-8

return r.text

except:

return ""

html\_text = "http://www.sina.com"

print(getHTMLText(html\_text))

**实验结果**：



1. 这是以爱丽丝梦游仙境为例（https://www.crummy.com/software/BeautifulSoup/bs4/doc/index.zh.html）在这学习

（**1）核心代码：**

html\_doc = """

<html><head><title>The Dormouse's story</title></head>

<body>

<p class="title"><b>The Dormouse's story</b></p>

<p class="story">Once upon a time there were three little sisters; and their names were

<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,

<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and

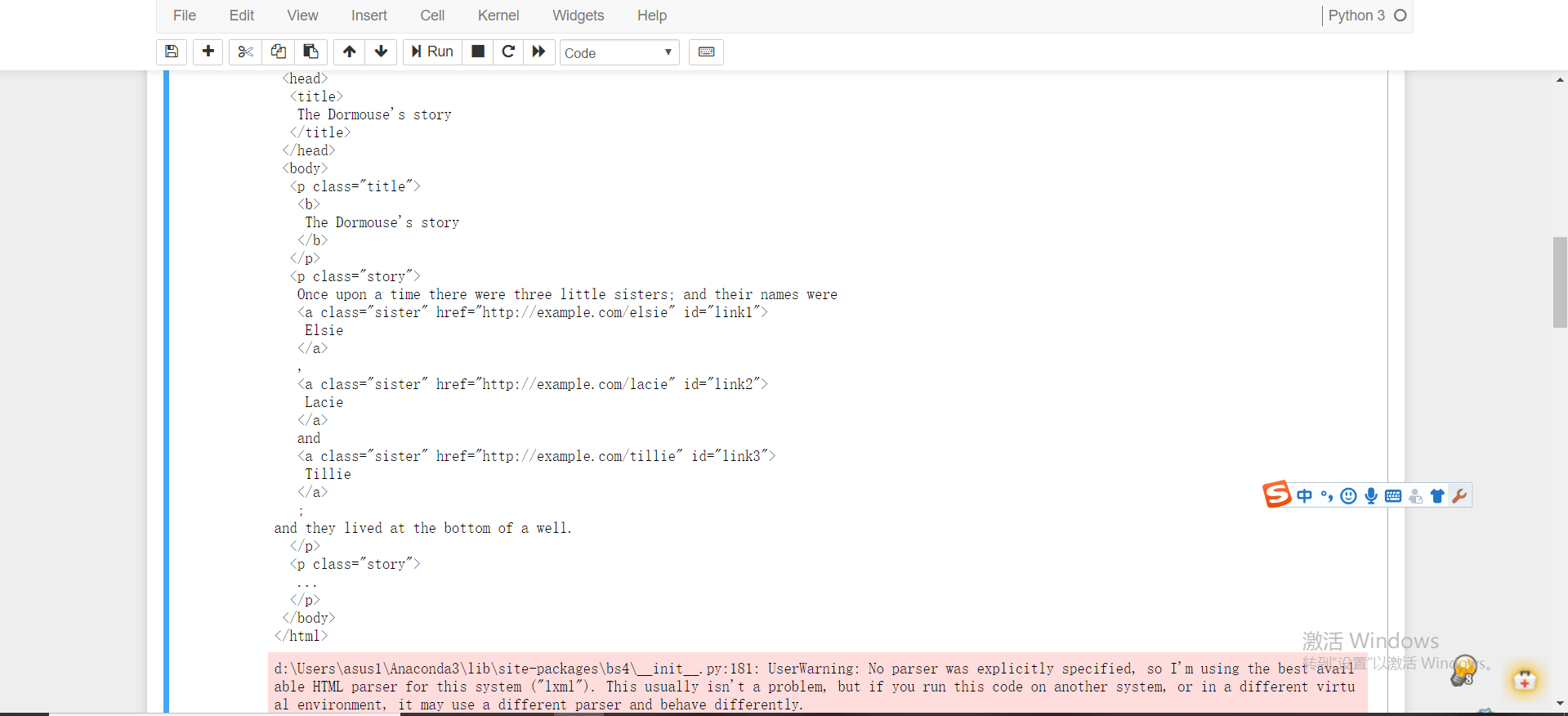
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;

and they lived at the bottom of a well.</p>

<p class="story">...</p>

"""

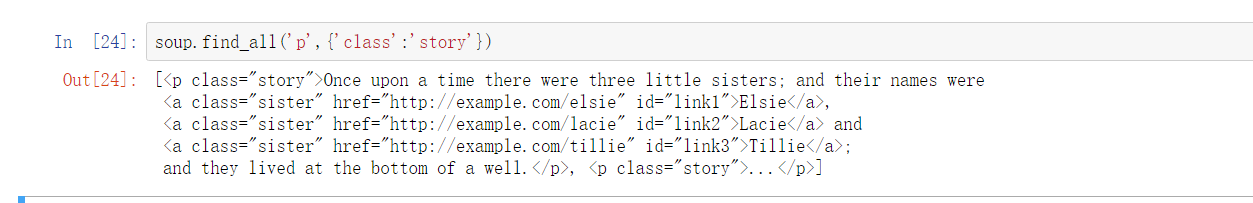
**实验结果：**



（2）







1. **对一些网络进行爬虫练习。**
2. **对国内大学进行爬虫**

**核心代码：**

#e23.1CrawUnivRanking.py

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：**



1. **对国外大学**

**核心代码：**

#e23.1CrawUnivRanking.py

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')

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

ldiv = tr.find\_all('div')

if len(ltd) == 0:

continue

singleUniv = []

for td in ltd:

singleUniv.append(td.string)

allUniv.append(singleUniv)

for strong in lstrong:

singleUniv.append(strong.string)

allUniv.append(singleUniv)

for div in ldiv:

singleUniv.append(div.string)

allUniv.append(singleUniv)

#a = [td,strong,div]

#b = [ltd,lstrong,ldiv]

#for a in b:

#singleUniv.append(a.string)

#allUniv.append(singleUniv)

def printUnivList(num):

print("{:^20}{:^20}{:^20}{:^20}".format("排名","大学","院系","国家","总分"))

for i in range(num):

u=allUniv[i]

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

def main():

url = 'http://www.zuihaodaxue.cn/Sport-Science-Schools-and-Departments-2017.html'

html = getHTMLText(url)

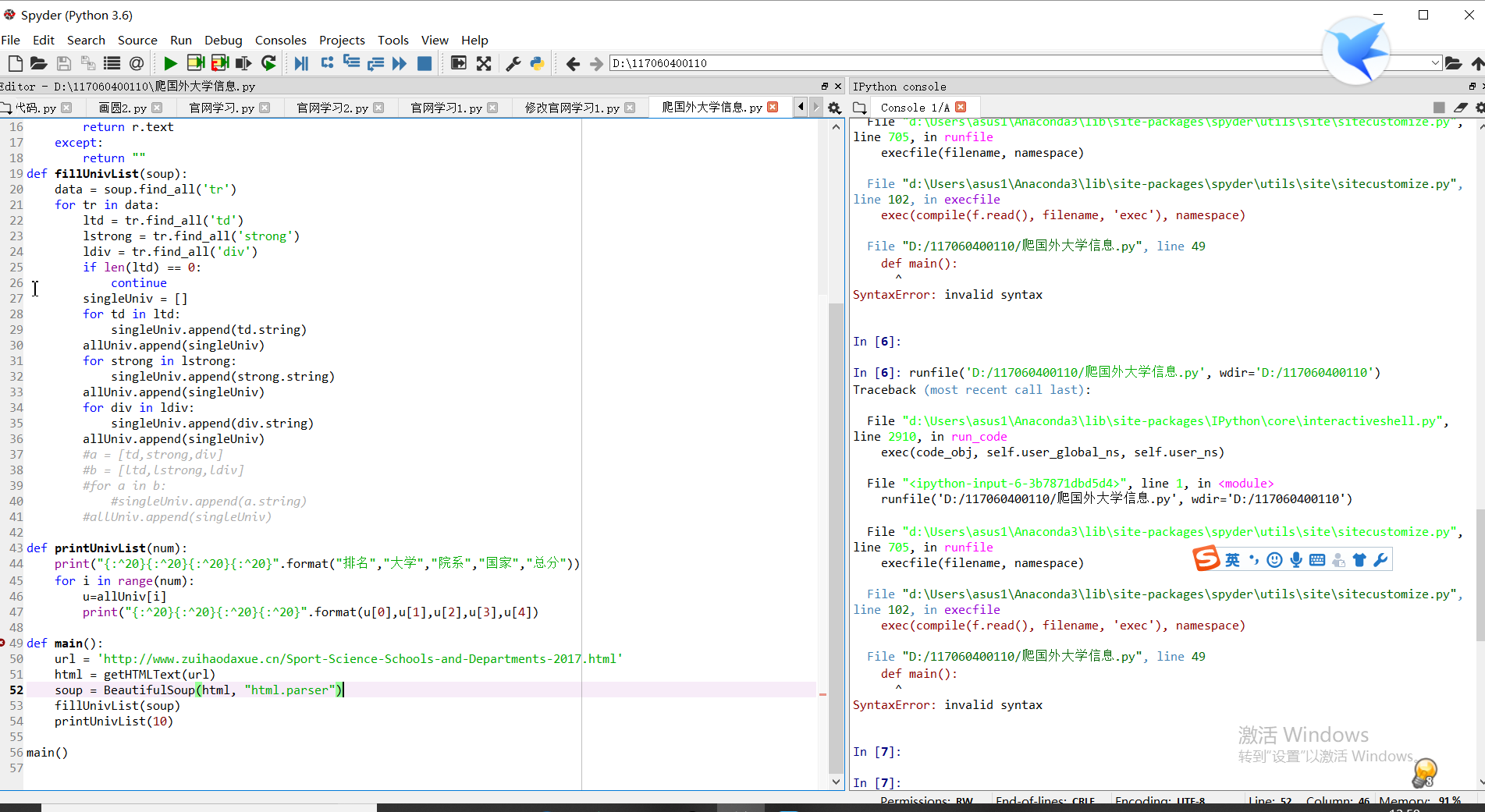
soup = BeautifulSoup(html, "html.parser")

fillUnivList(soup)

printUnivList(10)

main()

**实验结果**：



**注：**这是根据自己对中国大学排名爬虫的代码的理解，然后看源代码，之后写出的代码，我知道提示错了，但是所有代码都是在英文状态下的代码，但是还是错了。

**实验总结：**

1. 通过实验了解了requests库是对数据的获取，beautifulsoup库是对获取的数据进行分析。
2. 对beatifulsoup库是通过每个小的代码的输出的结果去理解beatifulsoup库
3. 大学网络爬虫的构建需要3个步骤：（1）从网络上获取网页内容（2）分析网页内容并提取有用数据到恰当的数据结构中（3）利用结构展示或进一步处理数据}