**实验报告12**

**学号：**117060400123 **姓名**：黄茜洋 **班级：**应统一班**指导老师：**林卫中**实验名称**：第十章网络爬虫和自动化

**实验要求：**（1）掌握网络爬虫的基本方法

（2）运用beautifulsoup4库解析和处理HTML

**算法实现：**

**1.**

. import requests

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

try:

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

r.raise\_for\_status()

r.encoding=coding

return r.text

except:

return ""

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>

"""

from bs4 import BeautifulSoup

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

print(soup)

2.

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)

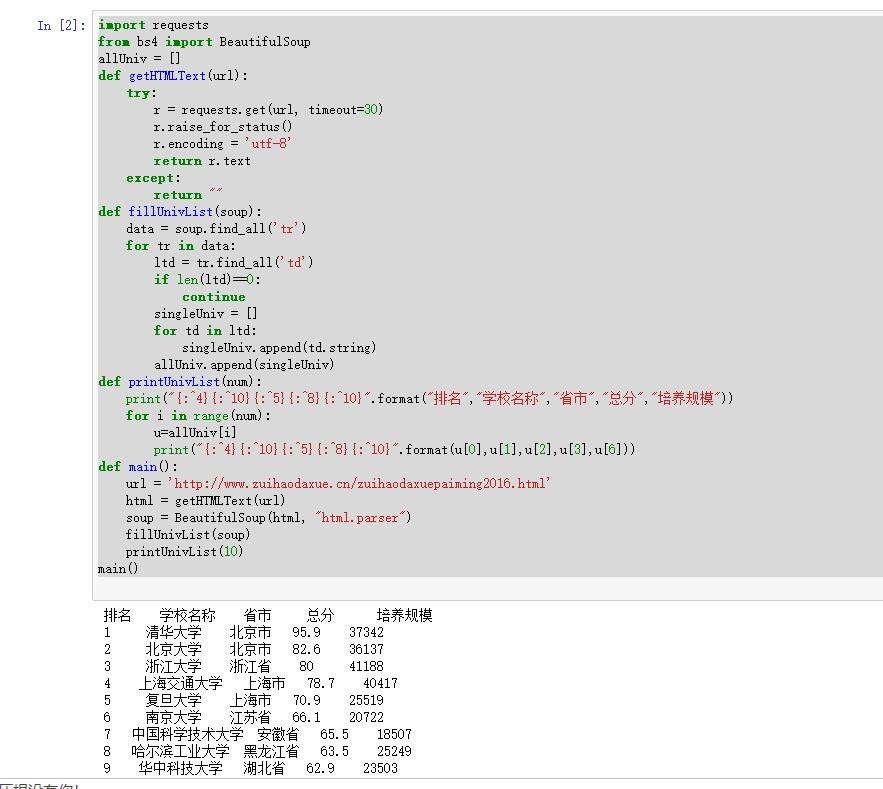
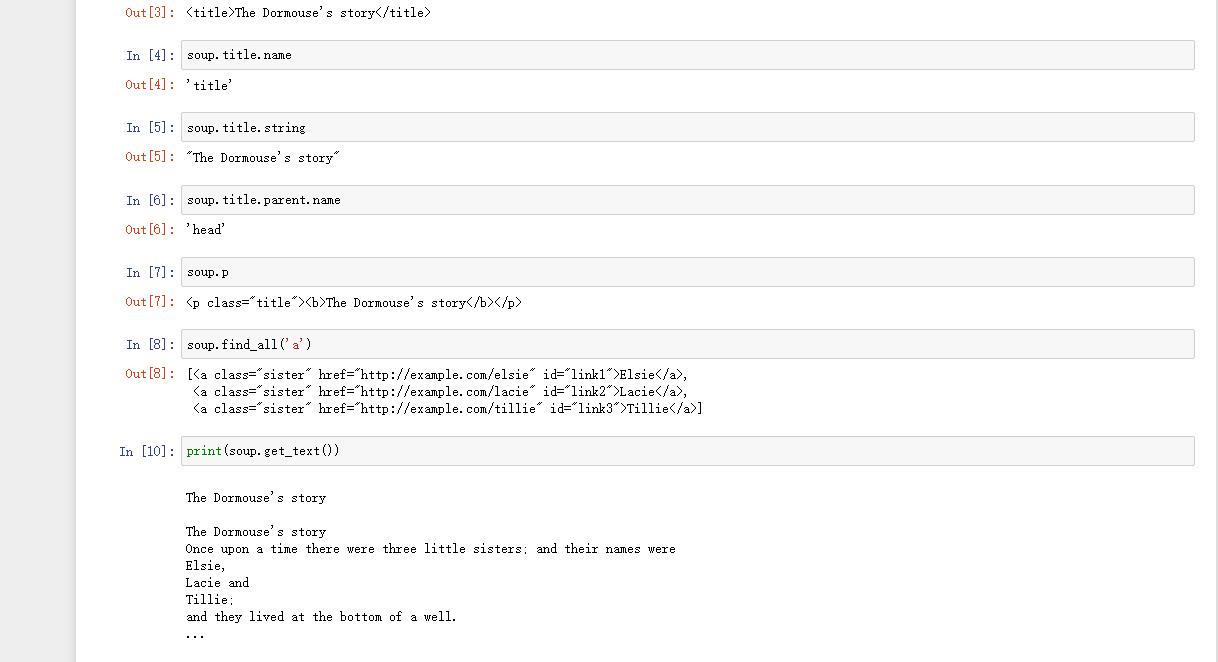
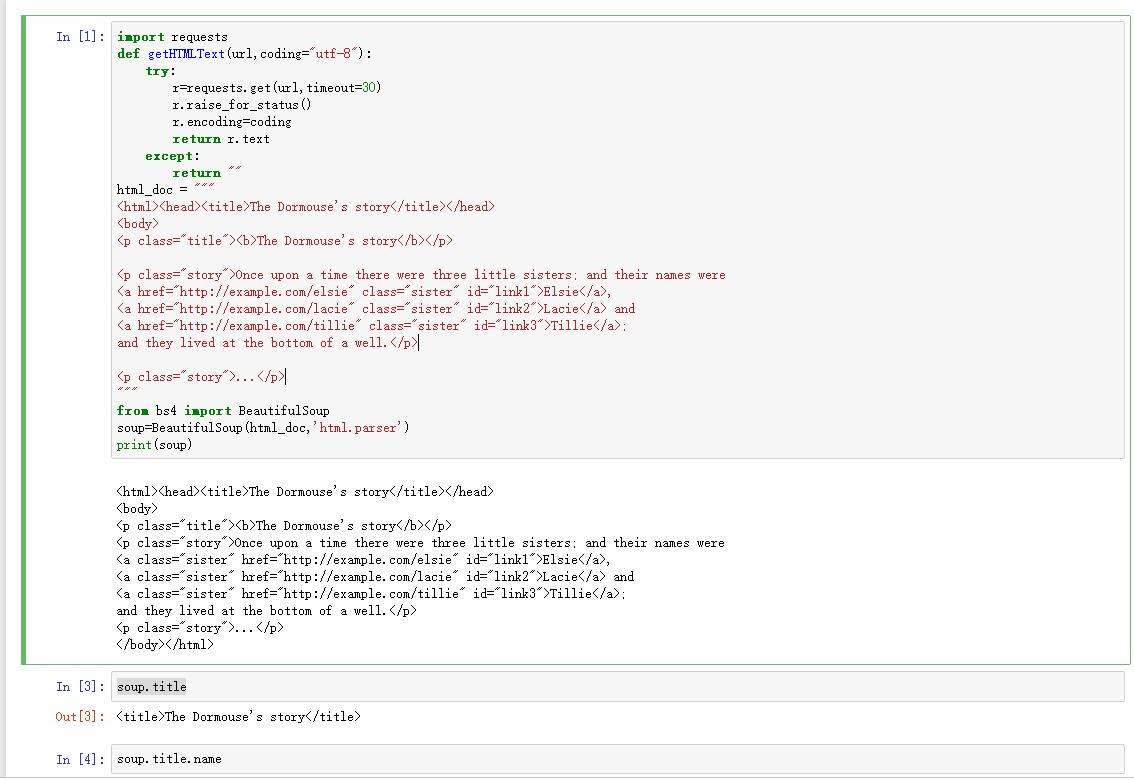
soup = BeautifulSoup(html, "html.parser")

fillUnivList(soup)

printUnivList(10)

main()

实验结果：



实验总结：

经过这次实验，我学会了怎么查看网站源代码，还有怎么看HTML的标题，学会了阅读beautifulsoup文档，学会了如何进行网站爬虫，虽然不是很懂，但是会了点。