**实验报告**

姓名：周蓓 学号：117060400109

实验名称：网络爬虫

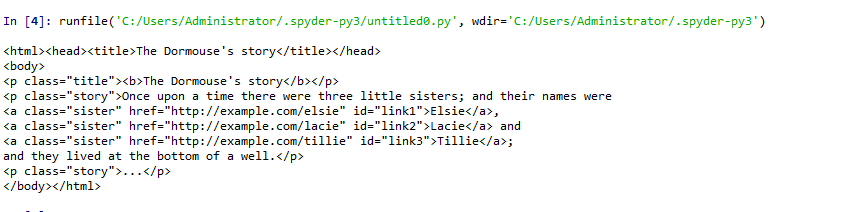
实验目的：（1）掌握网络爬虫的基本方法。

1. 学会运用requests库编写基本URL访问过程。
2. 学会运用beautifulsoup4库分析网页。

实验内容：（1）网页获取。requests库支持非常丰富的链接访问功能，最大的优点是程序编写过程更接近正常URL访问过程。requests库会产生几种常用异常，所以需要对获取网页的代码进行异常处理。代码如下：

import requests  
def getHTMLText(url,coding='urtf-8'):  
    try:  
        r = requests.get(url,timeout=30)  
        print(r)  
        r.raise\_fof\_status()  
        r.encoding = coding  
        return r.text  
    except:  
        return""  
html\_text = getHTMLText('http://www.sina.com.cn')  
print(html\_text)

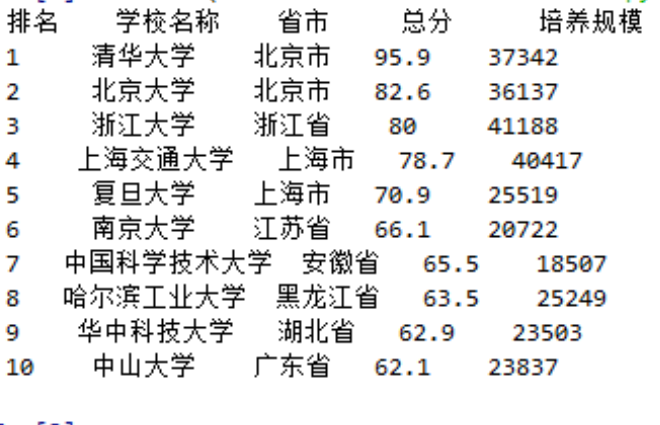
运行结果：



1. 中国大学排名爬虫。采用requests库爬取网页内容，使用beaufulsoup4库分析网页中的数据，提取排名前十的学校的相关数据。大学排名爬虫的构建需要三个步骤：第一，从网络上获取网页内容，第二，分析网页内容并提取有用数据到恰当的数据结构中，第三，利用数据结构展示或进一步处理数据。代码如下：

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 = 'IMG_256http://www.zuihaodaxue.cn/zuihaodaxuepaiming2016.html'  
    html = getHTMLText(url)  
    soup = BeautifulSoup(html, "html.parser")  
    fillUnivList(soup)  
    printUnivList(10)  
main()

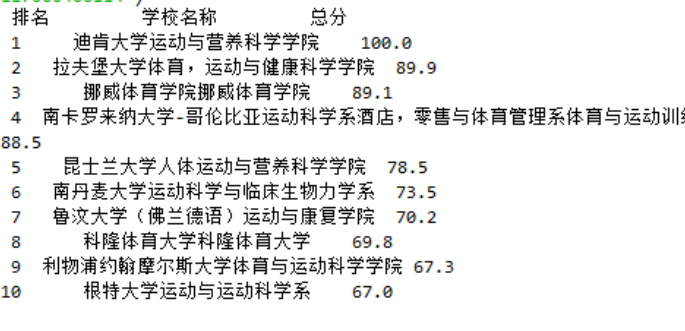
运行结果：



（3）世界大学排名爬虫。与（2）同样的方法，其代码如下：

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 = []  
        singleUniv.append(ltd[0].string)  
        singleUniv.append(ltd[1].get\_text())  
        singleUniv.append(ltd[3].string)  
        allUniv.append(singleUniv)  
def printUnivList(num):  
    print("{:^4}{:^20}{:^5}".format("排名","学校名称","总分"))  
    for i in range(num):  
        u=allUniv[i]  
        print("{:^4}{:^20}{:^5}".format(u[0],u[1],u[2]))  
def main():  
    url = 'IMG_256http://www.zuihaodaxue.cn/Sport-Science-Schools-and-Departments-2017.html'  
    html = getHTMLText(url)  
    soup = BeautifulSoup(html, "html.parser")  
    fillUnivList(soup)  
    printUnivList(10)  
main()

运行结果：



实验总结：通过本次实验课，网络爬虫有了初步的认识。网络爬虫可以更高效地获取指定信息需要定向抓取并分析网页资源，有效地提取并利用信息很大程度上提高解决问题的效率。当然，学习网络爬虫不能肆意地爬取网络数据，这是不文明的。