**数据结构实验报告9**

**学号：** 117060400225 **姓名**： 池艳 **班级：应用统计学二班**

**指导老师：林卫中**

**实验名称**：网络爬虫

**实验要求：（1）网络爬虫的基本方法**

（2）request库的使用

（3）**获取一个网页内容的函数**

（4）了解BeautifulSoup的find()和find\_all()方法

**实验题目：网络爬虫**

**算法实现：**

**1.软科中国大学排名：**

**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()**

**2.世界排名：**

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

**html = getHTMLText(url)**

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

**fillUnivList(soup)**

**printUnivList(10)**

**main()**

**3.获取一个网页内容的函数**

**import requests**

**def getHTMLText(url):**

**try:**

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

**r.raise\_for\_status()**

**r.encoding = 'utf-8'**

**return r.text**

**except:**

**return ""**

**text = getHTMLText('http://www.sina.com.cn')**

**print(text)**

**实验结果：**





