**数据结构实验报告12**

**学号：** 117060400211 **姓名**：王婕冉 **班级：** 应用统计学2班 **指导老师：林卫中**

**实验名称**： 网络爬虫和自动化

**实验要求：** 掌握网络爬虫的基本方法

**实验题目：中国大学排名爬虫**

**算法实现：**

**一、中国大学排名**

**#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()**

1. **世界体育大学排名**

**#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(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","学习名称","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()**

**三、**

import requests

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 ""

四、

def downloadImageFile(imgUrl, destUrl):

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

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

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

    with open(destUrl + "/" + local\_filename, '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

五、

def downloadImageFile(imgUrl, destUrl, fname):

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

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

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

    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

**实验结果：**

