**Python实验报告9**

**学号：** 117060400130 **姓名**： 罗建燕 **班级：** 17应用统计（1）班  **指导老师：** 林卫中

**实验名称1**： 大学排名

**算法实现：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(province):**

**print("{:^2}{:^10}{:^6}{:^4}{:^10}".format("排名","学校名称","省市","总分","培养规模"))**

**for u in allUniv:**

**if province in u[2]:**

**print("{:^2}{:^10}{:^6}{:^4}{:^10}".format(u[0],u[1],u[2],u[3],u[6]))**

**def main(p):**

**url = 'http://www.zuihaodaxue.cn/zuihaodaxuepaiming2016.html'**

**html = getHTMLText(url)**

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

**fillUnivList(soup)**

**printUnivList(p)**

**main('江西')**

**实验名称2：反爬虫机制**

**算法实现：import requests**

**from bs4 import BeautifulSoup**

**allUniv = []**

**def getHTMLText(url):**

**try:**

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

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

**print(r.status\_code)**

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

**return r.text**

**except:**

**print(r.status\_code)**

**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(province):**

**print("{:^2}{:^10}{:^6}{:^4}{:^10}".format("排名","学校名称","省市","总分","培养规模"))**

**for u in allUniv:**

**if province in u[2]:**

**print("{:^2}{:^10}{:^6}{:^4}{:^10}".format(u[0],u[1],u[2],u[3],u[6]))**

**def main(p):**

**url = 'http://www.zuihaodaxue.cn/zuihaodaxuepaiming2016.html'**

**html = getHTMLText(url)**

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

**fillUnivList(soup)**

**printUnivList(p)**

**getHTMLText("https://www.usnews.com/best-colleges/rankings/national-universities")**

**实验名称3：美国大学排名**

**算法实现：import requests**

**import re**

**from bs4 import BeautifulSoup**

**allUniv=[]**

**def getHTMLText(url):**

**send\_headers = {**

**"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.100 Safari/537.36",**

**"Connection": "keep-alive",**

**"Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,\*/\*;q=0.8",**

**"Accept-Language": "zh-CN,zh;q=0.8"}**

**try:**

**r = requests.get(url, headers=send\_headers)**

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

**print(r.status\_code)**

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

**return r.text**

**except:**

**return ""**

**def fillUnivList(soup):**

**data = soup.find\_all('div',{'class':re.compile('shadow-dark')})**

**for div in data:**

**singleUniv = []**

**div1 = div.find('div',{'style':'margin-left: 2.5rem;'})**

**rank = div1.get\_text().strip()**

**singleUniv.append(rank.split(' ')[0])**

**div2 = div.find('h3')**

**singleUniv.append(div2.get\_text().strip())**

**address = div.find('div',{'class':re.compile('block-normal')})**

**singleUniv.append(address.string)**

**lstrong = div.find\_all('strong')**

**singleUniv.append(lstrong[0].string)**

**singleUniv.append(lstrong[1].string)**

**allUniv.append(singleUniv)**

**def printUnivList():**

**print("{:<6}{:<20}{:<10}{:<10}{:<10}".format("排名","学校名称","City","学费","培养规模"))**

**for u in allUniv:**

**print("{:<6}{:<20}{:<10}{:<10}{:<10}".format(u[0],u[1],u[2],u[3],u[4]))**

**def main():**

**url = 'https://www.usnews.com/best-colleges/rankings/national-universities'**

**html = getHTMLText(url)**

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

**fillUnivList(soup)**

**printUnivList()**

**main()**

**实验名称4：编写爬虫抓取图片**

**算法实现：import requests**

**import re**

**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 ""**

**def downloadImageFile(imgUrl, destUrl, fname=''):**

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

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

**try:**

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

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

**if len(fname) == 0:**

**fname = local\_filename**

**print('fname={}'.format(fname))**

**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**

**except:**

**return r.status\_code**

**def getImg(html):**

**imgre = re.compile('"objURL":"(.\*?)"')**

**imglist = re.findall(imgre,html)**

**return imglist**

**def download(urls,path):**

**index = 1**

**for url in urls:**

**print("Download Image from page:{}".format(url))**

**status = downloadImageFile(url,path,str(index)+".jpg")**

**try:**

**if str(status)[0] == '4':**

**print("未下载成功{}".format(url))**

**continue**

**except Exception as e:**

**print("未下载成功{}".format(url))**

**index += 1**

**page = 'https://image.baidu.com/search/index?tn=baiduimage&word=范冰冰'**

**html= getHTMLText(page,'utf-8')**

**download(getImg(html),'e:\\129\\f')**