山东大学 软件工程 学院

2019-2020学年第1学期

Python程序设计与大数据分析 课程实验报告<模板>

**注：实验报告电子版文件名命名方式：学号+班级+姓名+实验\*。**

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：201800301249 | 姓名：王帅 | | 班级： 软件2018.4班 |
| 实验题目：网络爬虫与自动化 | | | |
| 实验学时：4 | | 实验日期： 2020年6月4日 | |
| 一、实验类型：验证型 二、建议学时：4 三、实验目的：掌握网络爬虫 与自动化  四、实验内容： 1、阅读教材第10章：网络爬虫和自动化  2、验证型练习：  2.1、验证10.2 模块10：requests库的使用中P265源代码10-1：获得一个HTML页面的通用代码  2.2、验证10.3 模块11:beautifulsoup4 库的使用中表10.4 BeautifulSoup对象的常用属性  2.3、验证10.3 模块11:beautifulsoup4 库的使用中表10.5 标签对象的常用属性  2.4、验证实例20：中国大学排名爬虫中实例代码20.1 2.5、验证实例21：搜索关键词自动提交实例代码21.1  3、设计型练习：教材P279的10.1,10.3,10.4 | | | |
| 硬件环境：  个人电脑 | | | |
| 软件环境：  IDLE | | | |
| 实验步骤与内容：  实验步骤：   1. 打开IDLE，根据实验要求，熟悉其使用 2. 根据书上的代码，深入自己对python的理解。   实验内容：教材PPT第10章种的范例程序代码 | | | |
| 结论分析与体会：  Python确实是一个比大多数语言都要方便简单的语言，简洁明了，大大减少了编写代码的时间。  虽然语法较为简单，但是自己的特有方法并不少，都是使用起来可以使我们的编程更加简便的方法，应当合理使用。  注意在使用IDLE时控制缩进，否则可能会发生较大的错误。 | | | |

以下分别粘贴实验中验证型题目中通过的程序代码（详细编号）与运行结果；

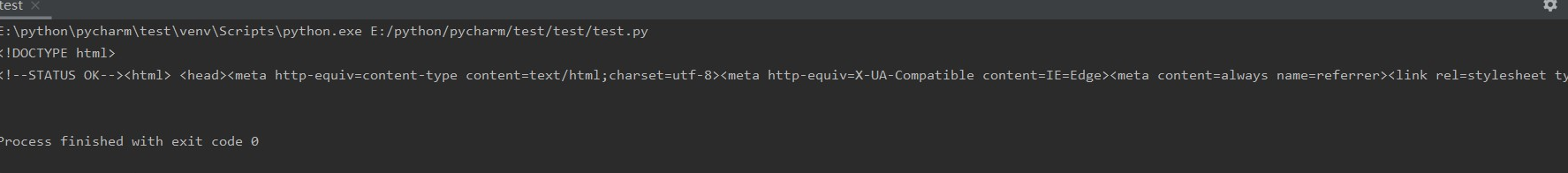
粘贴实验中设计型题目中调试通过的程序代码（详细编号）与运行结果

验证：

10.1、

import requests  
  
  
def getHTMLText(url):  
 try:  
 r = requests.get(url, timeout=30)  
 r.raise\_for\_status() # 如果状态不是200，引发异常  
 r.encoding = 'utf-8' # 无论原来用什么编码，都改成utf-8  
 return r.text  
 except:  
 return ""  
  
  
url = "http://www.baidu.com"  
print(getHTMLText(url))

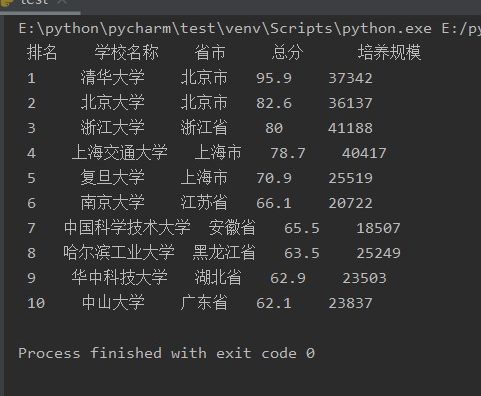
结果：



20.1、

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()

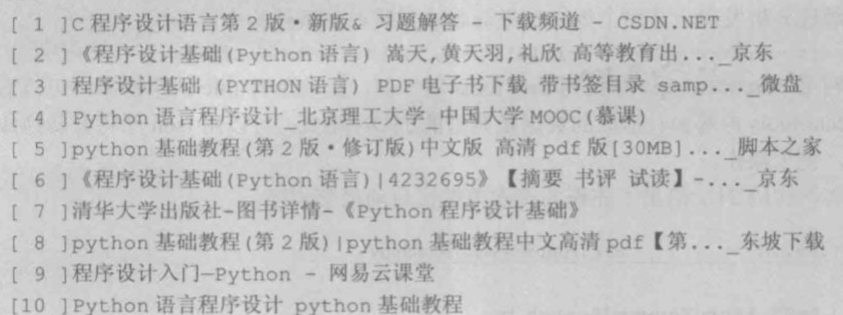
结果：



21.1、

import requests  
from bs4 import BeautifulSoup  
import re  
import json  
  
  
def getKeywordResult(keyword):  
 url = 'http://www.baidu.com/s?wd=' + keyword  
 try:  
 r = requests.get(url, timeout=30)  
 r.raise\_for\_status()  
 r.encoding = 'utf-8'  
 return r.text  
 except:  
 return ""  
  
  
def parserLinks(html):  
 soup = BeautifulSoup(html, "html.parser")  
 links = []  
 for div in soup.find\_all('div', {'data-tools': re.compile('title')}):  
 data = div.attrs['data-tools'] # 获得属性值  
 d = json.loads(data) # 将属性值转换成字典  
 links.append(d['title']) # 将返回链接的题目返回  
 return links  
  
  
def main():  
 html = getKeywordResult('Python语言程序设计基础(第2版)')  
 ls = parserLinks(html)  
 count = 1  
 for i in ls:  
 print("[{:^3}]{}".format(count, i))  
 count += 1  
  
  
main()

结果：



设计：

10.1、

import requests  
from bs4 import BeautifulSoup  
  
url = 'http://www.zuihaodaxue.com/zuihaodaxuepaiming2018.html'  
r = requests.get(url)  
r.encoding = "utf-8"  
soup = BeautifulSoup(r.text, 'html.parser')  
zpm = soup.find('div', {'class': 'news-blk'}).find('tbody', {'class': 'hidden\_zhpm'}).find\_all('tr', {'class': 'alt'})  
# 以上为爬取大学排名数据  
Alluniv = [] # 用于创建一个包含一条条大学排名的列表  
SF = [] # 创建一个省份列表，后期用来set()一个省份集合（集合的不重复性）  
for pm in zpm:  
 singleuniv = [] # 单个大学的排名信息  
 ltd = pm.find\_all('td') # 解析pm中所有的td标签的文本  
 for td in ltd:  
 singleuniv.append(td.string)  
 Alluniv.append(singleuniv)  
  
with open('E:\\大学排名.csv', 'a', encoding='ANSI')as f:  
 for i in range(len(Alluniv)):  
 u = Alluniv[i] # 单独一条大学排名信息  
 SF.append(u[2])  
 sf = set(SF) # 创建一个省份集合（集合的无重复性）  
 f.write("{},{},{},{},{}\n".format(u[0], u[1], u[2], u[3], u[4])) # 将结果写入文件  
 for ssff in sf: # 遍历省份集合  
 for i in range(len(Alluniv)):  
 u = Alluniv[i]  
 if ssff == u[2]: # 遍历每一条大学排名信息，如果u[2]==此ssff省份，则进行这条信息的显示输出  
 print("{:^4}{:^10}{:^4}{:^4}{:^4}".format(u[0], u[1], u[2], u[3], u[4]))

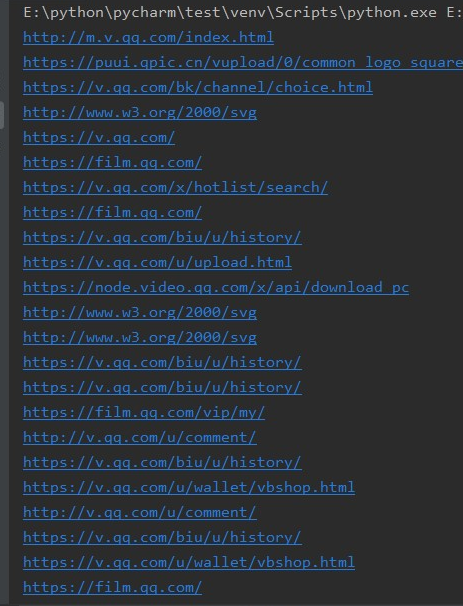
结果：



10.3、

import re  
import urllib.request  
  
  
def getlink(url):  
 headers = ("User-Agent",  
 "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/62.0.3202.94 Safari/537.36")  
 opener = urllib.request.build\_opener()  
 opener.addheaders = [headers]  
 urllib.request.install\_opener(opener)  
 file = urllib.request.urlopen(url).read()  
 file = file.decode('utf-8')  
 pattern = '(https?://[^\s)";]+(\.(\w|/)\*))'  
 link = re.compile(pattern).findall(file)  
 # 去重  
 # link = list(set(link))  
 return link  
  
  
#爬取腾讯视频  
url = "https://v.qq.com"  
linklist = getlink(url)  
for link in linklist:  
 print(link[0])  
print(len(linklist))

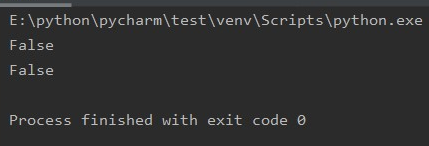
结果：



10.4、

from urllib.robotparser import RobotFileParser  
  
rp = RobotFileParser()  
rp.set\_url('http://www.jianshu.com/robots.txt')  
rp.read()  
print(rp.can\_fetch('\*', 'http://www.jianshu.com/p/b67554025d7d'))  
print(rp.can\_fetch('\*', "http://www.jianshu.com/search?q=python&page=1&type=collections"))

结果：



说明网页不可被抓取。