十三周星期五实验报告

实验课程：Python语言程序设计 姓名：韩天楠

学号：117060400112 班级：17应用统计学1班

指导老师：林卫中老师 时间：2018.6.3

1. 实验名称：

中国大学排名爬虫

1. 实验目的：
2. 掌握网络爬虫的基本方法；
3. 运用requests库编写基本URL访问过程；
4. 运用beautifulsoup4库解析和处理HTML.
5. 实验源代码

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)

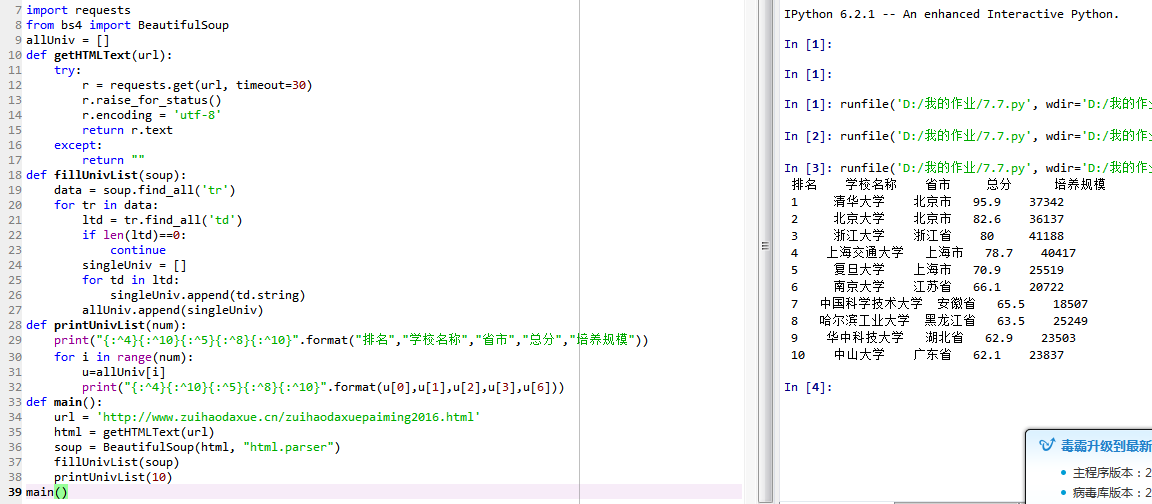
soup = BeautifulSoup(html, "html.parser")

fillUnivList(soup)

printUnivList(10)

main()

1. 实验运行结果



1. 实验小结

本次实验主要介绍了设计并实现网络爬虫的基本方法，结合requests和 beautifulsoup4两个库的使用，通过本次实验学会了一些基本的网络爬虫