Python实验报告13

班级：17应用统计学1班

姓名：王贵珍

学号：117060400124

指导老师：林卫中

实验题目：

按照省份输出中国大学排名http://www.zuihaodaxue.cn/zuihaodaxuepaiming2018.html

请分别输出江西省和山东省的高校排名

实验代码：

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(province):

print("{:^4}{:^10}{:^5}{:^8}{:^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("{:^4}{:^10}{:^5}{:^8}{:^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)

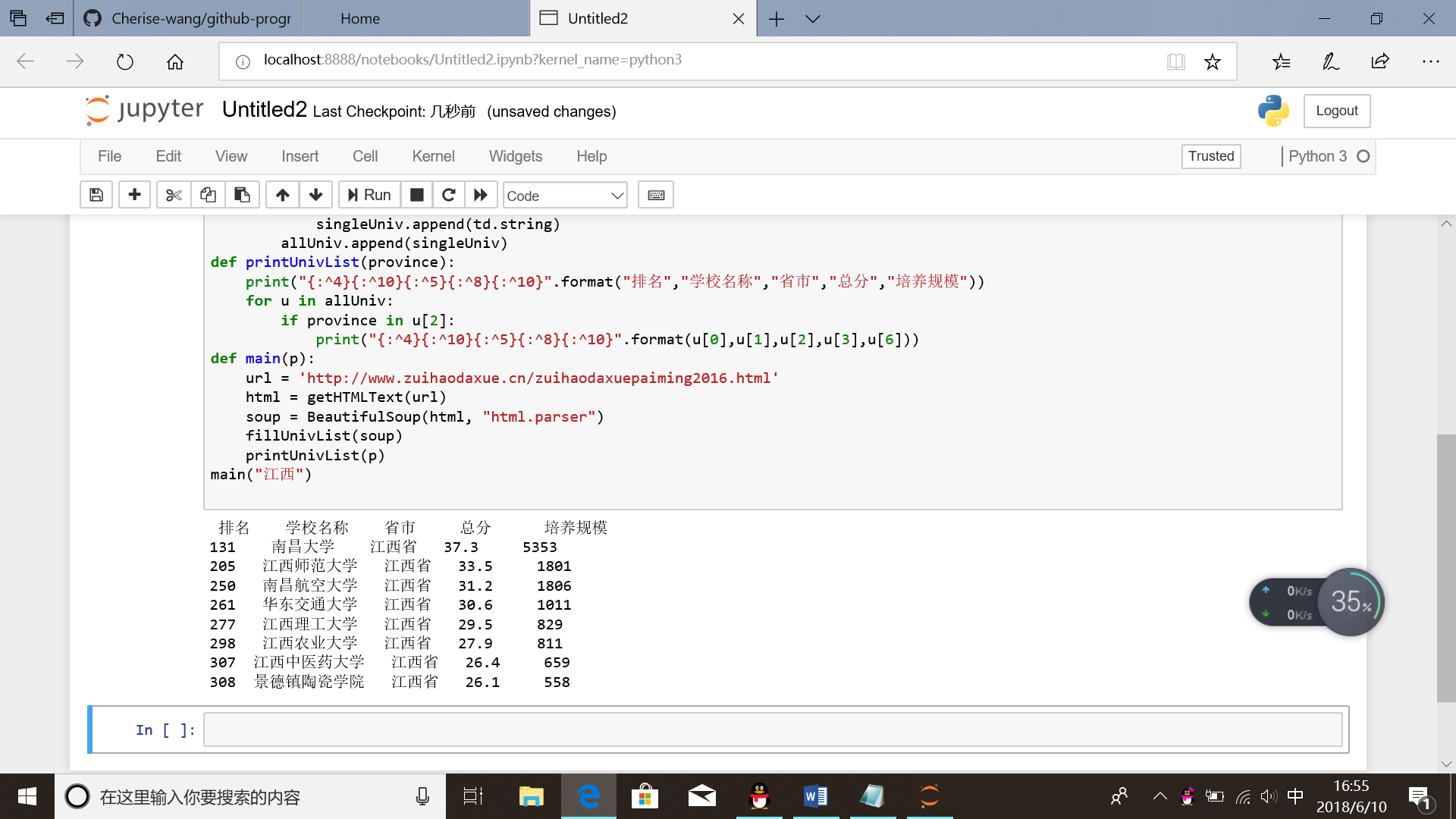
soup = BeautifulSoup(html, "html.parser")

fillUnivList(soup)

printUnivList(p)

main("江西")

实验结果：



实验题目：USNEWS美国大学排名爬虫。美国大学排名网址如下：

https://www.usnews.com/best-colleges/rankings/national-universities

注意：由于该网站有反爬虫机制，需要模拟浏览器行为进行爬虫，参考代码如下：

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 ""

实验代码：

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 = []

#排名

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

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

#学校名称

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

div2 = div.find('h3')

#City

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)#学费

singleUniv.append(lstrong[1].string)#培养规模

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()

实验结果：

