**实验报告13**

**学号：**117060400123 **姓名**：黄茜洋 **班级：**应统一班**指导老师：**林卫中**实验名称**：第十章网络爬虫和自动化

**实验要求：**（1）掌握网络爬虫的基本方法

（2）运用beautifulsoup4库解析和处理HTML

**算法实现：**

**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(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'

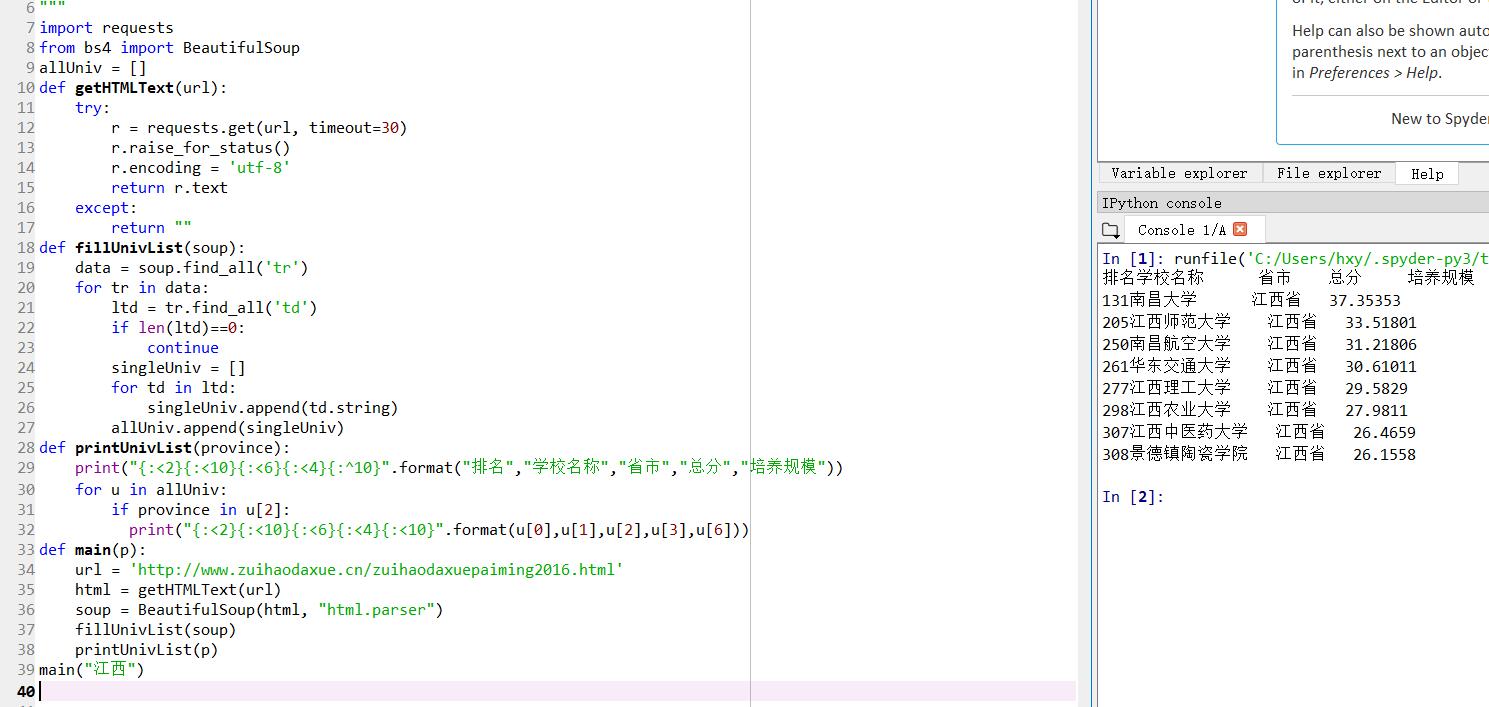
html = getHTMLText(url)

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

fillUnivList(soup)

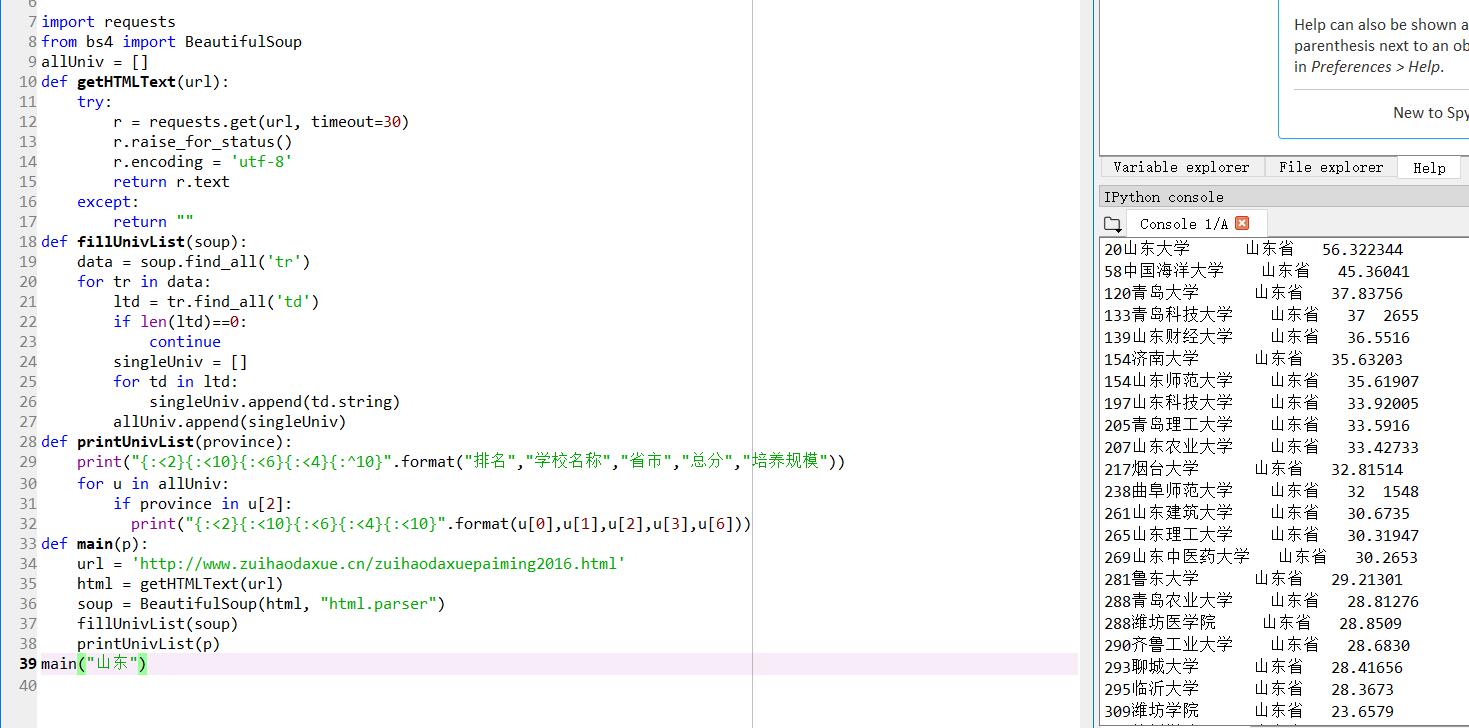
printUnivList(p)

main("江西")



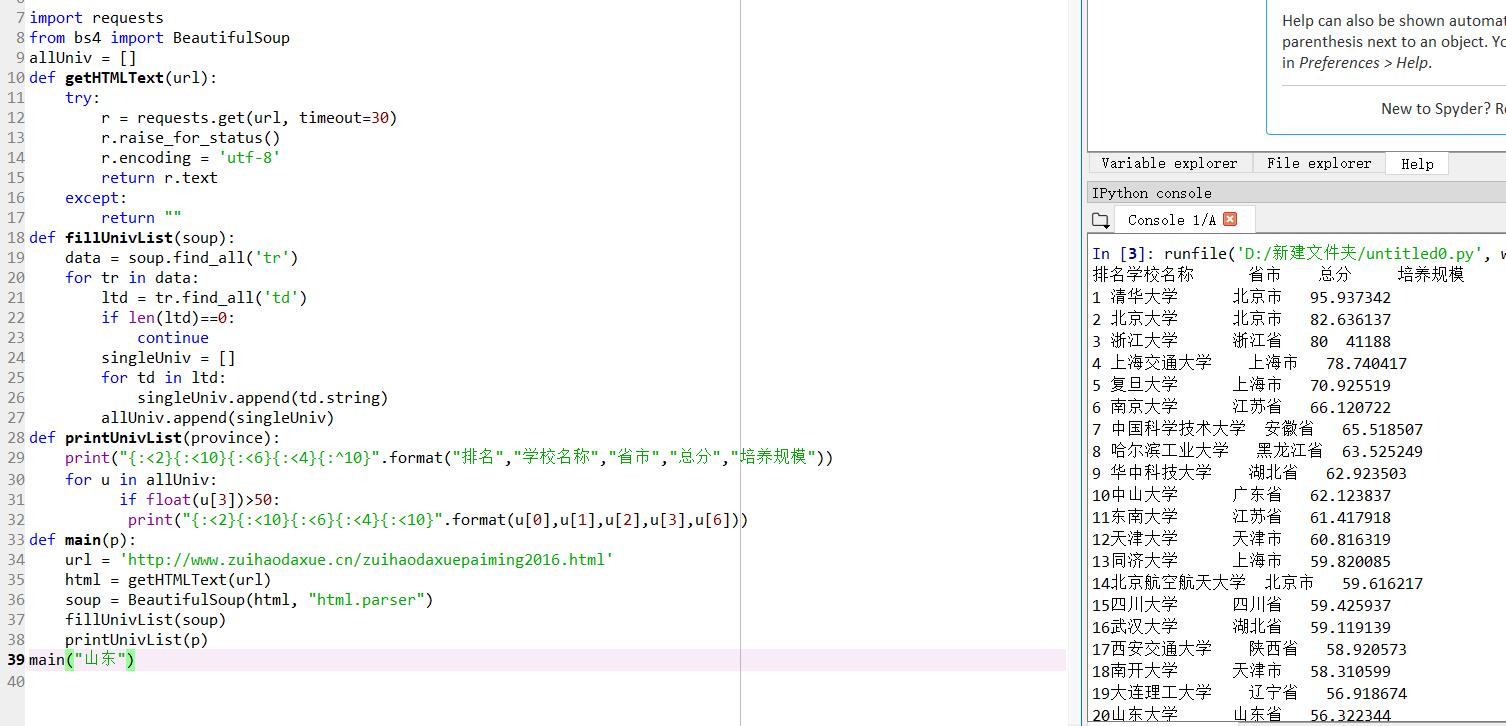
2.

import requestsfrom bs4 import BeautifulSoupallUniv = []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("山东")

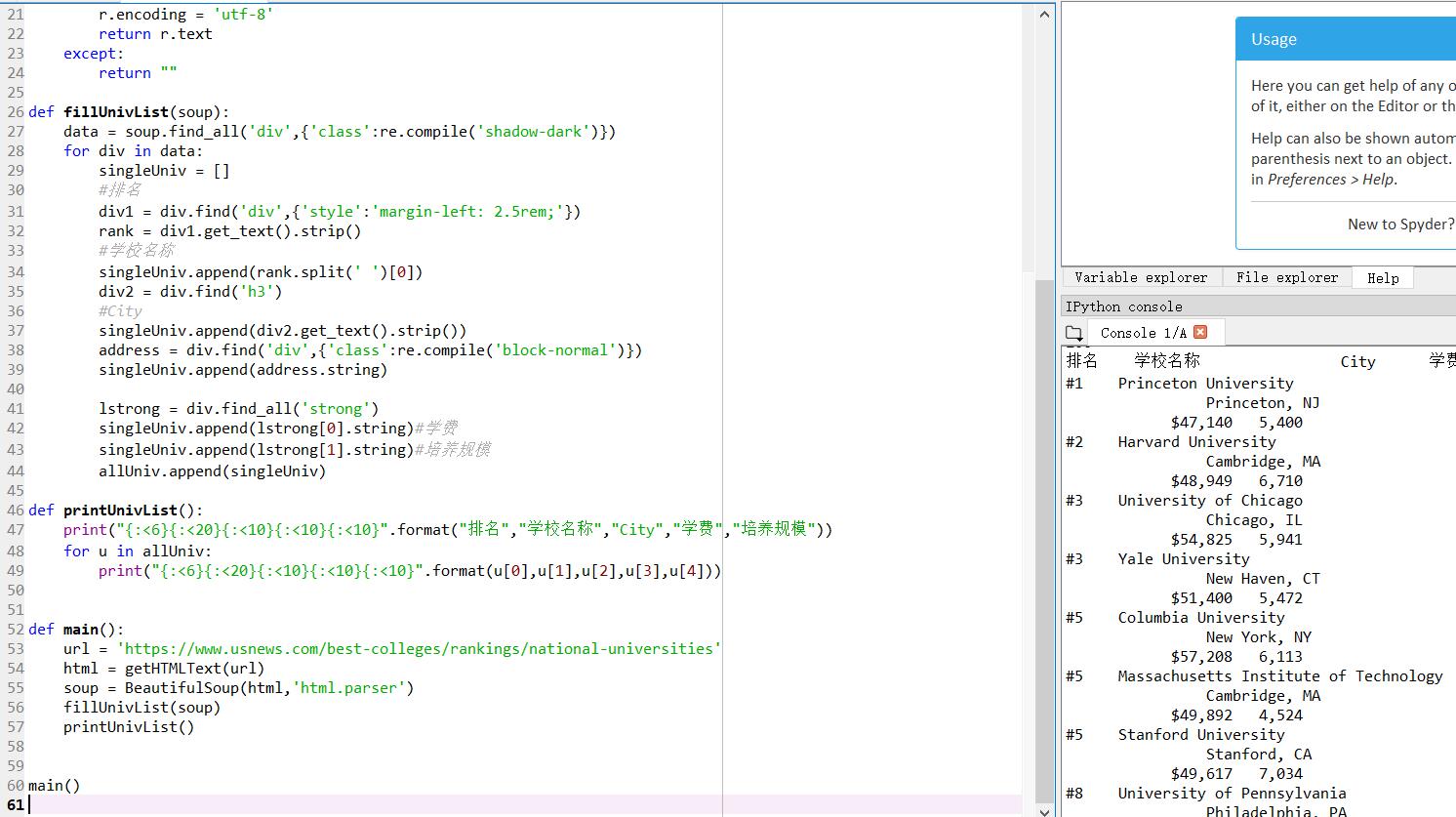


3.

import requestsfrom bs4 import BeautifulSoupallUniv = []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 float(u[3])>50: 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("山东")

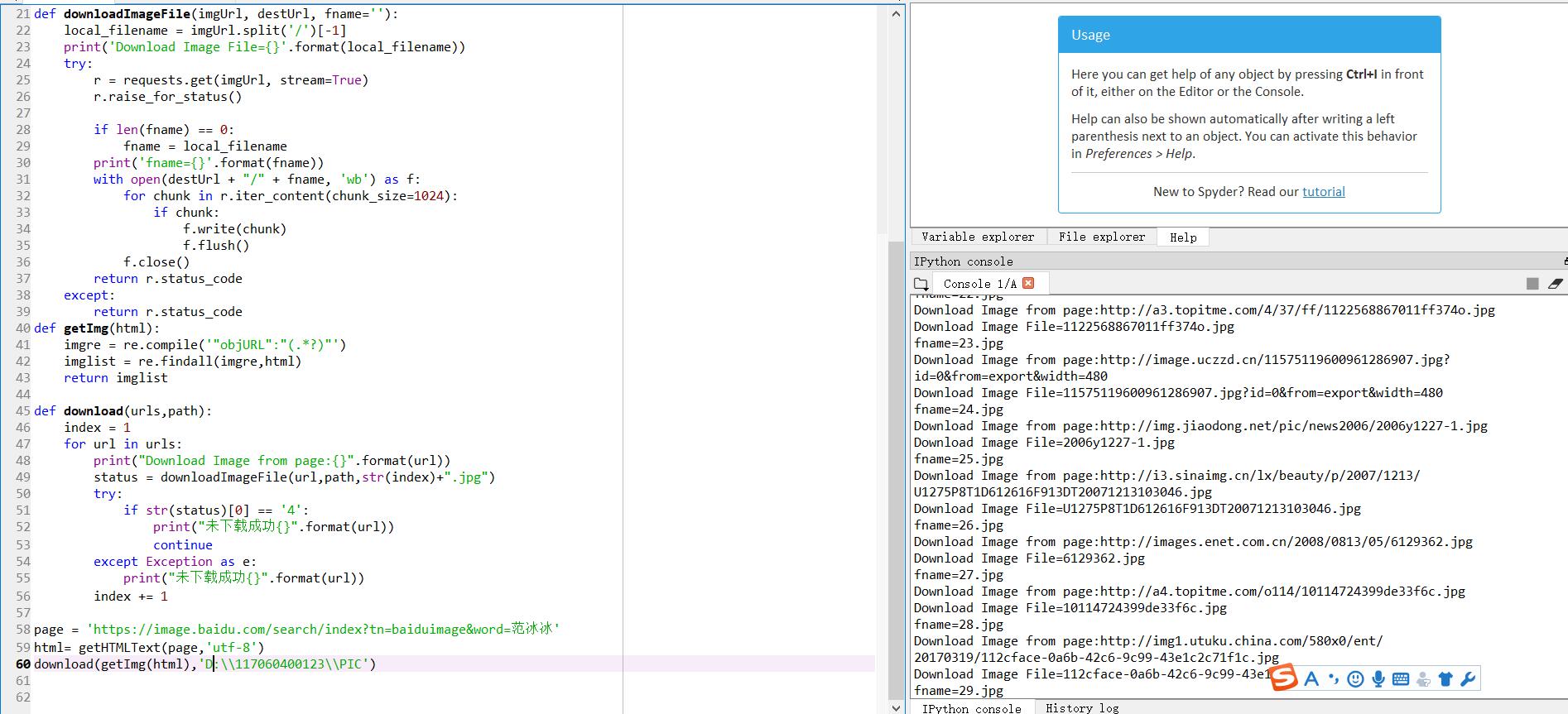


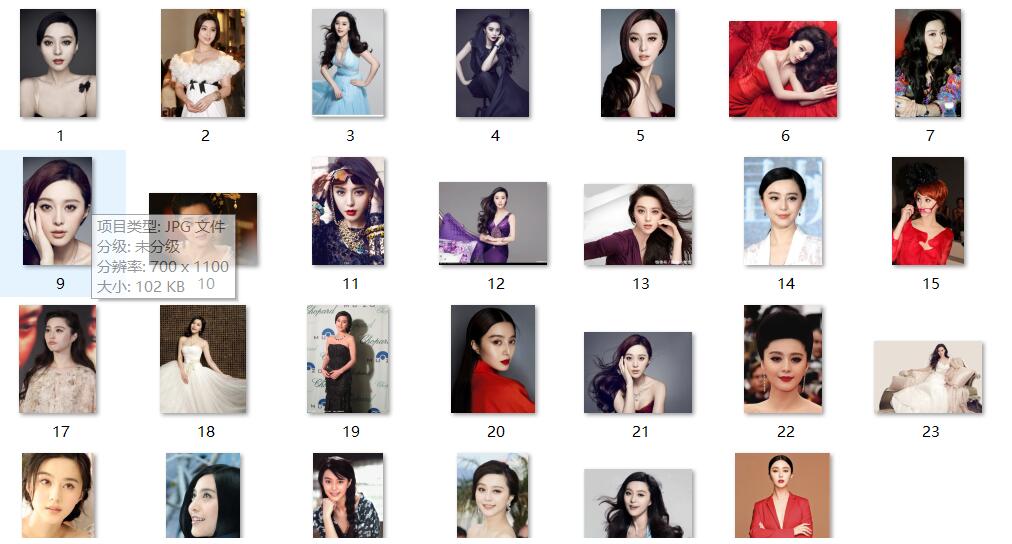
4.

import requestsimport refrom bs4 import BeautifulSoupallUniv=[]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()

5.

import requestsimport redef 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\_codedef getImg(html): imgre = re.compile('"objURL":"(.\*?)"') imglist = re.findall(imgre,html) return imglistdef 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 += 1page = 'https://image.baidu.com/search/index?tn=baiduimage&word=范冰冰'html= getHTMLText(page,'utf-8')download(getImg(html),'D:\\117060400123\\PIC')



实验总结：经过这次实验，我大致地学会了如何查看代码。然后进行网上爬虫，想深层次的学下去，感觉挺好玩的。突然对python有了点兴趣