第十四周Python实验报告（周五）

**学号**：117060400110 **姓名**：蒙柳双

**班级**：17应用统计学1班 **指导老师**：林卫中老师

**实验目的**：

1. 熟练对requests库、beautifulSoup库、re库的使用
2. 考察对以前学习的知识的掌握程度
3. 突破一些网络的限制，从而去爬取更多的内容，下载更多的图片

**实验要求：**

1. 上课不玩手机
2. 要认真编码并且思考

3.实在不会的就向老师求助

**实验内容步骤**：

1. **任务一**：爬取美国大学排名前30名或者更多的大学的学校名称、学费、培养规模 <https://www.usnews.com/>

**核心代码：**

（1）

**核心代码：**

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')})#正则表达式

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(n):

allUniv=[]

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

for i in range(1,n+1):

rl = url + '?\_page=' + str(i)

html = getHTMLText(rl)

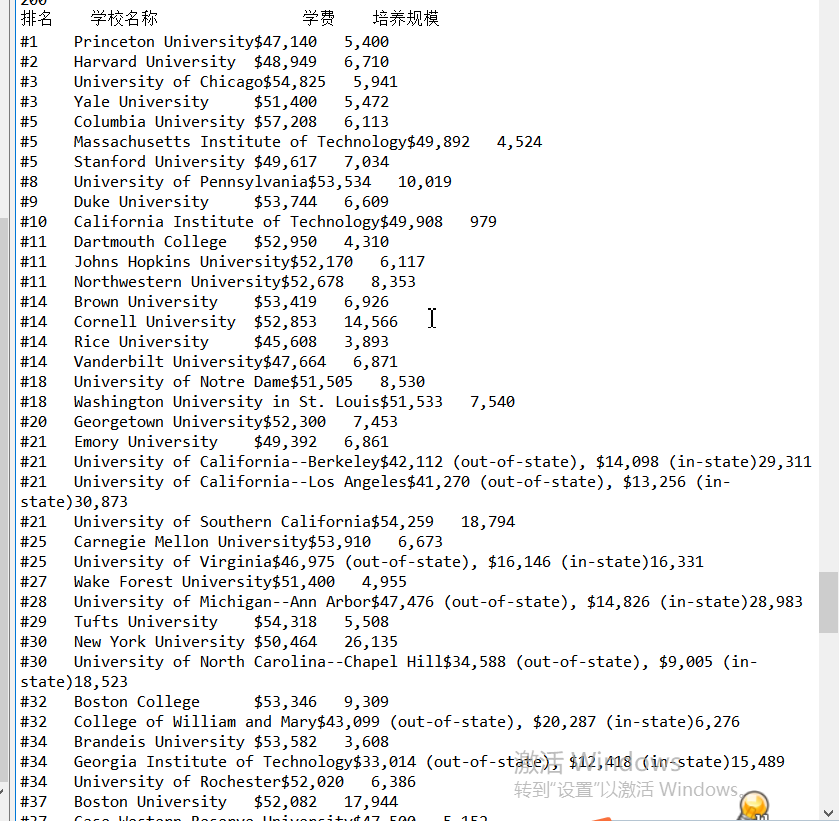
soup = BeautifulSoup(html,'html.parser')

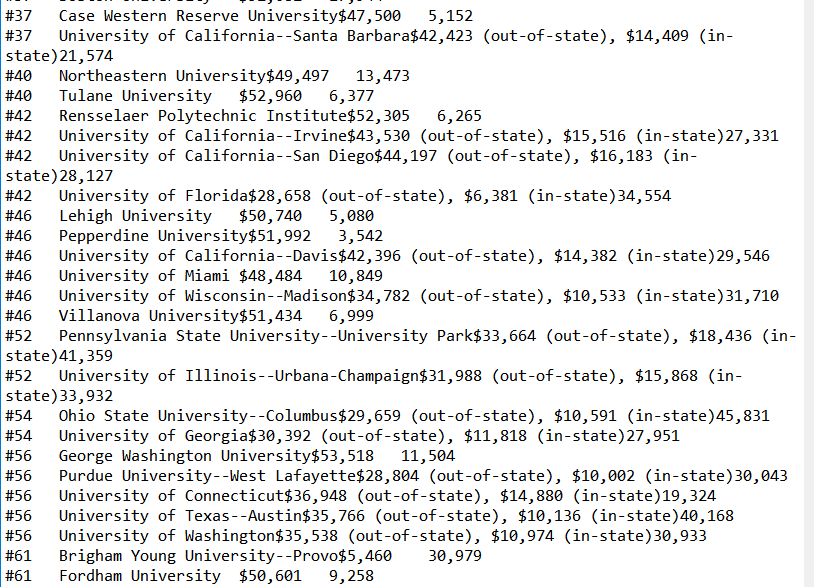
fillUnivList(soup)

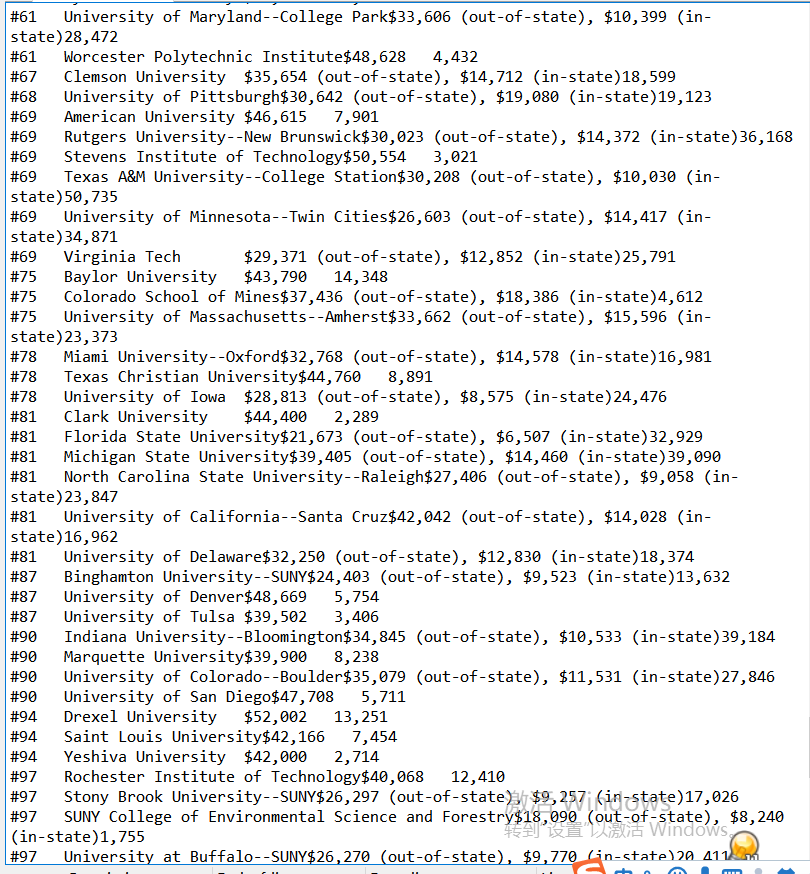
printUnivList()

main(10)

**实验结果：**







1. **学费少于50000$的国外大学**

**核心代码**

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 = []

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

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

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

univName = div.find('h3')

singleUniv.append(univName.get\_text().strip())

ldiv = div.find\_all('div',{'style':'padding-right: 0.5rem;'})

singleUniv.append(ldiv[0].strong.string)

singleUniv.append(ldiv[1].strong.string)

allUniv.append(singleUniv)

def printUnivList():

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

for u in allUniv:

s = u[2].split(' ')

f=s[0].replace(',','')

f=f.replace('$','')

#if s[0] in ['1','2','3','4','5']:

if int(f)<50000:

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

def main(n):

#allUniv=[]

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

for i in range(1,n+1):

rl = url + '?\_page=' + str(i)

html = getHTMLText(rl)

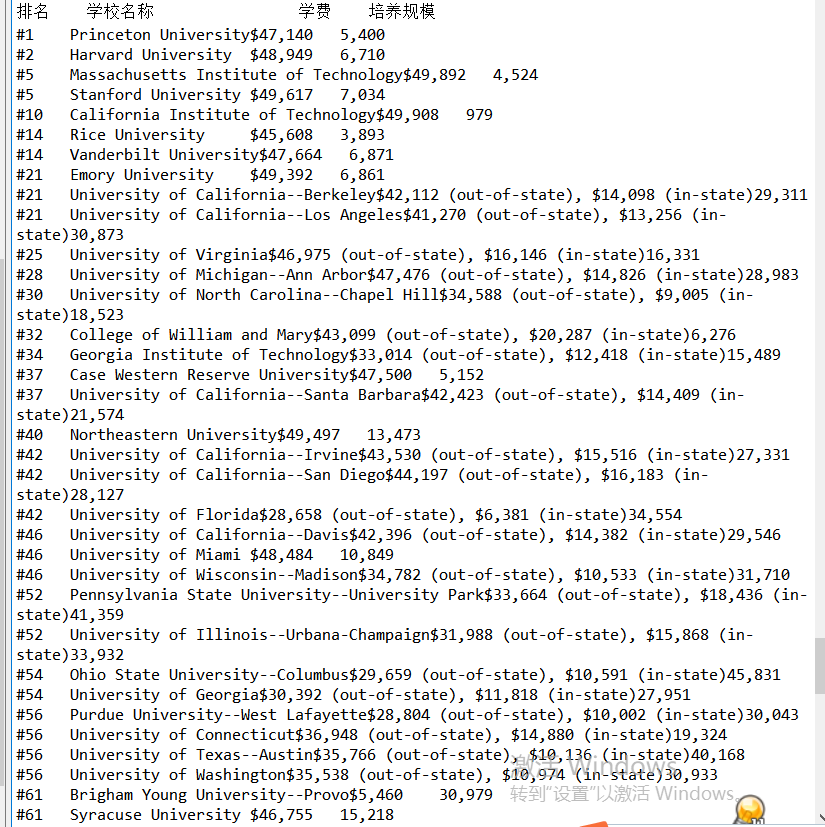
soup = BeautifulSoup(html,'html.parser')

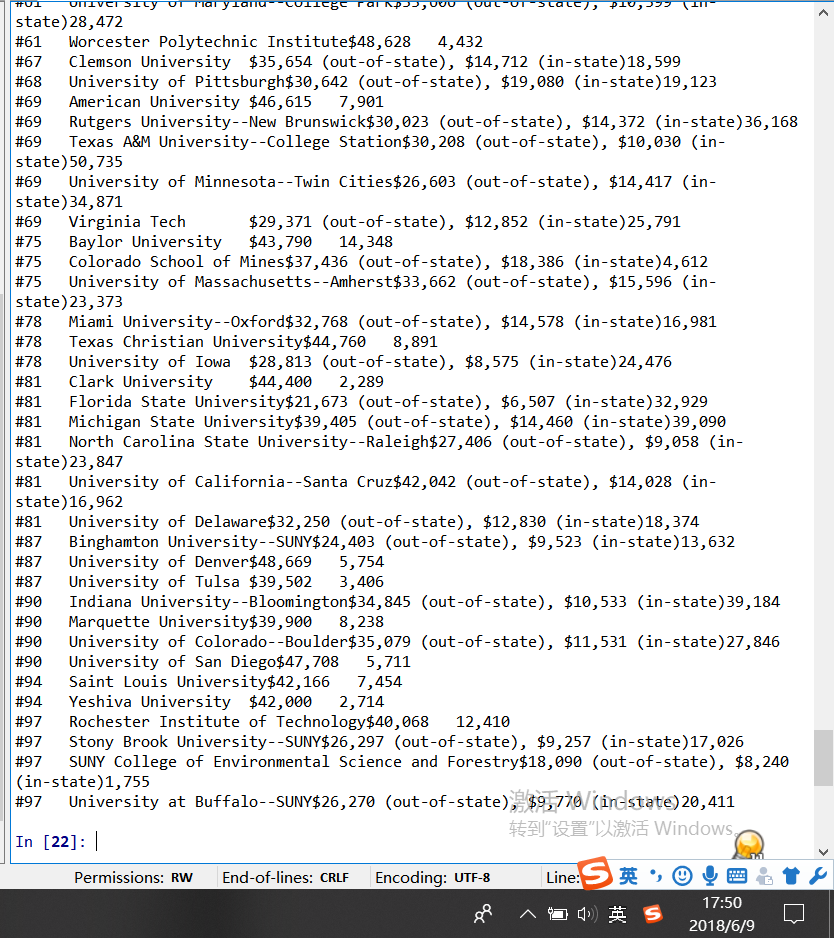
fillUnivList(soup)

printUnivList()

main(10)

**实验结果：**





1. **任务二：从百度图片上爬取90张范冰冰（或其他明星）的照片**

**核心代码：**

from bs4 import BeautifulSoup

import re

import requests

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 getMorePages(kw, pages):

params = []

for i in range(30, 30\*pages+30, 30):

params.append({

'ipn': 'rj',

'ct': 201326592,

'is': '',

'fp': 'result',

'queryWord': kw,

'cl': 2,

'lm': -1,

'ie': 'utf-8',

'oe': 'utf-8',

'adpicid': '',

'st': -1,

'z': '',

'ic': 0,

'word': kw,

's': '',

'se': '',

'tab': '',

'width': '',

'height': '',

'face': 0,

'istype': 2,

'qc': '',

'nc': 1,

'fr': '',

'pn': i,

'rn': 30,

'gsm': '1e',

'1528253616462': ''

})

url = 'https://image.baidu.com/search/acjson?tn=resultjson\_com'

datalist = []

for param in params:

dj = requests.get(url, params=param).json()

data = dj['data']

if data is not None and len(data) > 0:

datalist.append(data)

return datalist

def main(kw, pages, desurl):

datalist = getMorePages(kw, pages)

index = 1

for data in datalist:

for i in data:

if i.get('thumbURL') is not None:

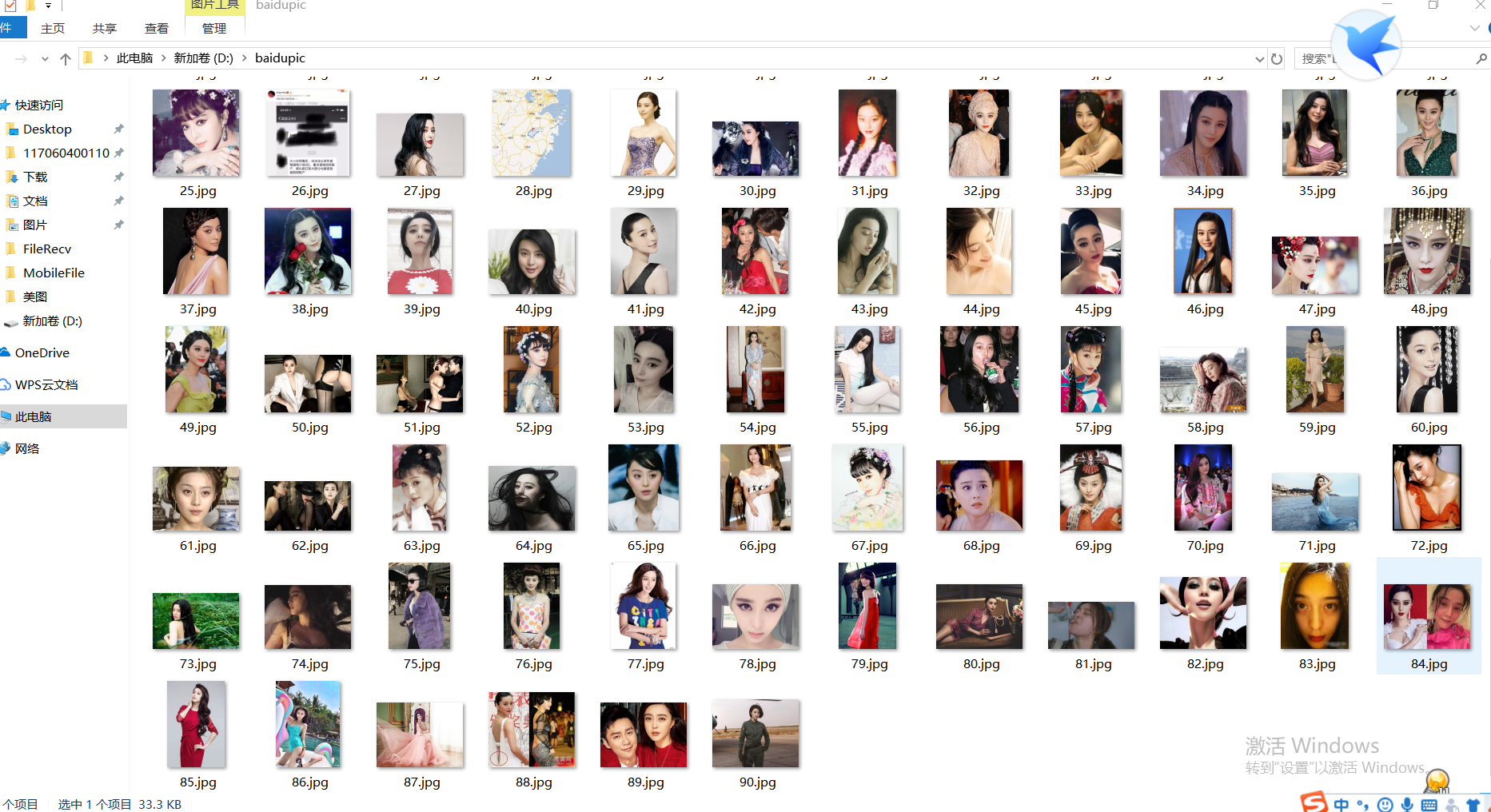
ir = i.get('thumbURL')

downloadImageFile(ir, desurl, str(index)+'.jpg')

index = index + 1

main('范冰冰',3, 'd:/baidupic')

**实验结果：**



实验总结：

1. 更加熟练的对requests库、beautifulSoup库、re库的使用
2. 发信自己对于学过的知识掌握的不是很好,就是对一些综合运用的题目不会解题（需自我反省）
3. 当然了，通过这次实验，我可以突破一些网络的限制去爬取自己需要的内容，需要的图片。