**实验报告**

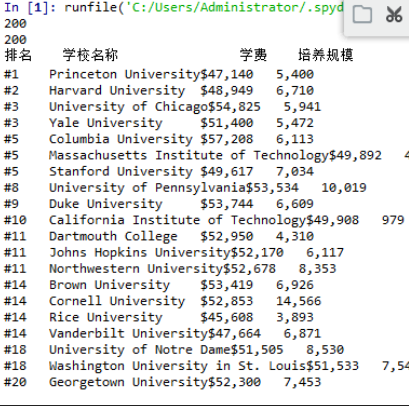
姓名：周蓓 学号：117060400109 指导老师：林卫中

实验名称：爬虫实验指导

实验要求：熟练运用requests和beautifulsoup4函数库爬取数据，利用百度搜索提供的这个链接接口，可以通过requests的get（）函数提交查询，响应结果为百度搜索结果。

实验内容：1、爬取美国大学排名前30名的学校名称、学费、培养规模。代码如下：

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:  
        print("{:<6}{:<20}{:<10}{:<10}".format(u[0],u[1],u[2],u[3]))  
def main(n):  
    url = 'IMG_256https://www.usnews.com/best-colleges/rankings/national-universities?\_page='  
    for i in range(1,n):  
         ri = url + str(i)  
         html = getHTMLText(ri)  
         soup = BeautifulSoup(html,'html.parser')  
         fillUnivList(soup)          
    printUnivList()  
main(3)

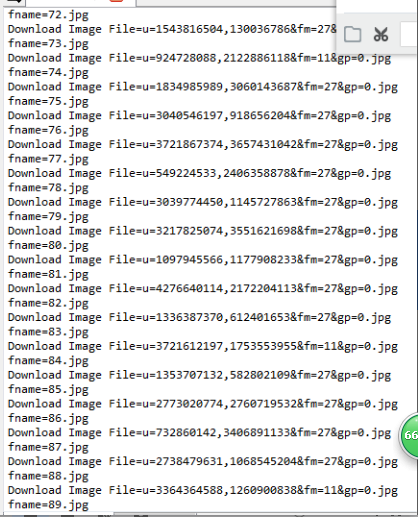
运行结果：  


2、从百度图片上爬取90张范冰冰（或其他明星）的照片 http://image.baidu.com/

代码如下：

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 = 'IMG_256https://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(‘李钟硕',3, 'd:/baidupic')

运行结果：



实验总结：通过这次实验课，能够在有限的条件下爬取更多的内容，了解到更多的网络信息。通过学习，能够自己爬取一些自己想知道的东西，激发学生的学习兴趣 。