**数据结构实验报告10**

**学号：** 117060400225 **姓名**： 池艳 **班级：应用统计学二班**

**指导老师：林卫中**

**实验名称**：爬虫

**实验要求：（1）** 大学排名爬取

（2）HTML网页阅读

（3）编写爬虫代码抓取图片并下载

**实验题目：**

1.爬大学排名

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()

2.美国大学排名

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])

        div2 = div.find('div',{'class':'block-normal text-small'})

        singleUniv.append(div2.string.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","计划培养人数"))

    for u in allUniv:

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

def main():

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

    html = getHTMLText(url)

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

    fillUnivList(soup)

    printUnivList()

main()

3.照片下载

import requests

from bs4 import BeautifulSoup

import re

def 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\_code

def getImg(html):

imgre = re.compile('"objURL":"(.\*?)"')

imglist = re.findall(imgre,html)

return imglist

def 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 += 1

page = 'http://image.baidu.com/search/index?tn=baiduimage&ps=1&amp;ct=201326592&lm=-1&cl=2&amp;nc=1&ie=utf-8&;word=%E8%8C%83%E5%86%B0%E5%86%B0'

html= getHTMLText(page,'utf-8')

download(getImg(html),'c:\\1（同上，你要保存到的地址）')

**算法实现：**

**实验结果：**





