## 实验十、网络爬虫

实验目的：学习网络爬虫编写，获取数据

实验要求：

1. 能运用request库和beautifulsoup4库访问URL并解析获取的HTML

2. 能向百度等搜索引擎自动提交关键词并获取返回结果

实验学时：4学时

实验内容：

1. 程序练习题10.1

提示：修改教材实例代码20.1中的25-27行即可，只输出给定的省份的学校。要求输出江西省和北京市的高校

2. 程序练习题10.2

提示：

1）、美国大学排名网站：<https://www.usnews.com/best-colleges/rankings/national-universities>

2）、上述网站只列出了排名前10的高校，当鼠标往下翻滚时才会加载后10个高校，仔细分析html源代码，找到显示后10个高校的网址

3. 程序练习题10.6

提示：

1）下载网页上的图片代码如下函数：imgUrl是图片的网址；destUrl是存储在本地的地址；fname是图片保存时的名字，默认值为空，图片保存时按网络上图片的名字保存

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

2）注意分析百度图片的搜索结果，图片的结果是保存在json格式中，关键字为thumbURL所对应的值就是图片的网址。

3）百度图片返回的搜索结果只有30幅图像，只有当鼠标下移才能看到后面30张图像，依此类推。分析HTML，找到相关网址下载后30张图像。

4）选择一个自己喜欢的明星，下载TA的90张照片。

实验代码10.1

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(num):

Univ = sorted(allUniv, key = lambda location: location[2])

print("{1:^2}{2:{0}^10}{3:{0}^6}{4:{0}^4}{5:{0}^10}".form

at(chr(12288),'

排名

','

学校名称

','

省市

','

总分

','

培养规模

'))

for i in r

ange(num):

u = Univ[i]

print("{1:^4}{2:{0}^10}{3:{0}^5}{4:{0}^8.1f}{5:{0}^10}".f

ormat(chr(12288),u[0],u[1],u[2],eval(u[3]),u[6]))

def main(num):

url=

'http://www.zuihaodaxue.cn/zuihaodaxuepaiming2016.htm

l'

html = getHTMLText(url)

soup = BeautifulSou

p(html, 'html.parser')

fillUnivList(soup)

print(len(allUniv))

printUnivList(len(allUniv))

main(10)

实验代码10.2

import

requests

from bs4 import BeautifulSoup

# USNEWS 的大学排名默认是卡片显示，为了处理方便，这里将它设置成表格显示

# USNEWS 的大学排名分页，在每一页的读取里加入了读取下一页网址的功能，就是代码中的Button部分

# 回避其他因素，只进行了美国大学排名。

def getHtml(url):

response = requests.get(url)

response.raise\_for\_status()

response.encoding = 'utf

-

8'

return response.text

def parseHtml():

url =

'http://colleges.usnews.rankingsandreviews.com/best

-

colleges/rankings/national

-

universities?\_sort=rank&\_sort

-

direction=asc&\_mode=table'

while True:

text = getHtml(url)

soup = BeautifulSoup(text)

button = soup.find\_all('a', {'class': 'button

secondary radius shadow full

-

width '})

if len(button) == 0:

break

url = button[0]['href']

university =

soup.find\_all('tr', {'data

-

parent

-

id':'search

-

application

-

results

-

view'})

rankUniversities(university)

def rankUniversities(universities):

for i in universities:

name = i.div.a.string

rank = i.span.string[1:4]

if rank =

= 'ank' or rank == 'nra':

rank = 'N/A'

if rank[

-

1] == 'i':

rank = rank[:

-

1]

enroll = [j for j in i.find\_all('td',{'class':'text

-

right'})][1].div.string

print('{:<55}{:<5}{:<7}'.format(name, rank,

enroll.repla

ce('

\

n','').replace(' ','')))

parseHtml()

实验代码10.6

import requests

import json

import re

import os

import random

folder\_path = 'haha'

def getBaiduImage(url):

try:

rsp = requests.get(url, timeout = 10)

rsp.raise\_for\_status()

except:

print('

对不起，百度图片访问失败！程序退出

')

return

data = json.loads(rsp.text)

downLoadImage(data)

def tryPic(url):

form

= url[

-

4:]

pat = '.\*/(.\*?)' + form

pattern = re.compile(pat, re.S)

filename = re.findall(pattern, url)

return filename[0], form

def downLoadImage(data):

global folder\_path

os.mkdir(folder\_path)

imgs = data['data']

count = 0

for i in imgs:

try:

url = i['download\_url']

img = requests.get(url, timeout = 10)

img = img.content

image,form = tryPic(url)

FileName = folder\_path +'

\

\

'+image+form

file = open(FileName,'bw')

file.write(img)

print('No.%d success' % count)

except Ke

yError:

print('No.%d failed'%count)

count += 1

#

用女明星的图片进行测试

url =

'http://image.baidu.com/channel/listjson?pn=0&rn=30&tag1=%E6%98

%8E%E6%98%9F&tag2=%E5%85%A8%E9%83%A8&ftags=%E5%A5%B3%E6%98%8E%E

6%98%9F##

内地

&ie=utf8'+'1'+'&rn=60&ie=utf8&oe=utf

-

8&'

+str(random.random())

getBaiduImage(url）