## 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

实验内容：

1.程序练习题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 range(num):

u = Univ[i]

print("{1:^4}{2:{0}^10}{3:{0}^5}{4:{0}^8.1f}{5:{0}^10}".format(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 = BeautifulSoup(html, 'html.parser')

fillUnivList(soup)

print(len(allUniv))

printUnivList(len(allUniv))

main(10)

2. 程序练习题10.2

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。