**2022春《python程序设计》实验指导书**

**python网页爬虫与信息提取之——**

**中国大学排名**

**班级：\_\_\_\_\_\_计实验20\_\_\_\_\_\_**

**姓名：\_\_\_\_\_\_\_苏靖博\_\_\_\_\_\_\_**

**学号：\_\_20105050110\_\_\_**

**成绩：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**实验目的**

熟悉python使用正则表达式进行字符串提取的方法，熟悉requests库、beautifulSoup库及正则表达式的基本运用，掌握使用python开发定向网页爬虫的方法。

**时间要求**：课内4学时+课外4学时

**问题描述**：

软科中国大学排名”前身是“中国最好大学排名”，自2015年首次发布以来，以专业、客观、透明的优势赢得了高等教育领域内外的广泛关注和认可，已经成为具有重要社会影响力和权威参考价值的中国大学排名领先品牌。

1. 2019、2020、2021年上海交通大学推出的最好大学排名网址分别是：

https://www.shanghairanking.cn/rankings/bcur/2019

https://www.shanghairanking.cn/rankings/bcur/2020

https://www.shanghairanking.cn/rankings/bcur/2021

1. 利用requests库、BeautifulSoup库以及正则表达式，使用format()方法进行字符串格式化输出，使用python编写程序以下图所示的方式分别显示2019、2020、2021三年的中国大学排名(以下数据均为模拟数据)：

2019年中国大学排名（前18）

排名 　　　学校名称　　　 总分 科学研究

1 　　　清华大学　　　 95.3 100.0

2 　　　北京大学　　　 88.6 96.4

3 　　　浙江大学　　　 76.9 86.3

4 　　上海交通大学　　 73.1 90.5

5 　　　复旦大学　　　 66.0 91.5

6 　中国科学技术大学　 61.9 92.1

7 　　　南京大学　　　 59.8 88.4

8 　　华中科技大学　　 59.1 82.4

9 　　　中山大学　　　 58.6 81.1

10 　哈尔滨工业大学　　 57.4 78.2

11 　　　同济大学　　　 56.4 87.0

12 　　　武汉大学　　　 55.5 84.4

13 　　　东南大学　　　 55.3 81.8

14 　　西安交通大学　　 54.2 83.6

15 　北京航空航天大学　 54.0 85.4

16 　　　南开大学　　　 53.9 85.9

17 　　　四川大学　　　 53.3 75.7

18 　　　天津大学　　　 52.4 82.5

1. 输出的大学排名数量（如：10，18、20等）以**参数**的形式提交给程序；完成上述功能后，将以2019、2020、2021三年的学科水平、科学研究、人才培养三项总分之和作为排名依据(若某项数据为百分数，则自己约定从1到100分，对应百分之多少即可，需确保逻辑正确)，再次以上述方式输出大学三年综合排名（输出综合排名时无学生项）。
2. 已知**炒股一点通**网站可以查询深市和沪市的股票代码信息，网址分别为：

沪市： 'http://www.cgedt.com/stockcode/hushi.asp 及

深市： 'http://www.cgedt.com/stockcode/shenshi.asp'

请使用**正则表达式**从上述两个网址获取所有股票代码。

1. 结合要求3，从要求4获得的全部股票代码中随机选取的n支股票代码分配给大学排名的前n所大学，n为键盘输入值。达到如下效果（数据均为示意，此处n=8）：

2019-2021年中国大学排名（前8）

排名 　　　学校名称　　　 总分 股票代码

1 　　　清华大学　　　 295.3 600567

2 　　　北京大学　　　 288.6 000008

3 　　　浙江大学　　　 276.9 000301

4 　　上海交通大学　　 273.1 600033

5 　　　复旦大学　　　 266.0 000001

6 　中国科学技术大学　 261.9 300099

7 　　　南京大学　　　 259.8 600993

8 　　华中科技大学　　 259.1 000008

**试设计一个程序完成上述要求**。

**Code:**

# author: Jingbo Su

# created: 17.05.2022 09:09:56

import io

import os

import sys

import requests

from bs4 import BeautifulSoup

import bs4

def get(url):

try:

re = requests.get(url, timeout = 30)

re.raise\_for\_status()

re.encoding = re.apparent\_encoding

html = re.text

except:

html = "Fail"

return html

def info1(a, html):

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

data = soup.tbody.contents

for t in data:

if isinstance(t, bs4.element.Tag):

info = t.find\_all('td')

a.append([info[0].div.string,

info[1].find(name = 'a', attrs = {'class': 'name-cn'}).string,

info[2].get\_text(), info[3].get\_text(),

info[4].string, info[5].string])

else:

pass

def info2(b, html):

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

data = soup.div.contents

for t in data:

if isinstance(t, bs4.element.Tag) and t.find(name = 'div', attrs = {'id': 'stockcodelist'}):

info = t.find\_all('a')

for x in info:

b.append(x.get\_text())

'''

try:

b.append([t.find(name = 'a', attrs = {'target': '\_black'}).get\_text()])

except:

print("Attr Error\n", b)

'''

else:

pass

def printf(a, b, n):

col = "{0:{7}^0}\t{1:{7}^10}\t{2:{7}^10}\t{3:{7}^10}\t{4:{7}^7}\t{5:{7}^7}\t{6:{7}^6}\t"

print(col.format('rank','大学名称','省市','类型','Total Score','School Level', 'Stock Code', chr(12288)))

for i in range(n):

x = a[i]

y = b[i + 1]

s = ""

for c in y:

if c.isdigit():

s += c

print(col.format(str(x[0]).strip(), str(x[1]).strip(), str(x[2]).strip(), str(x[3]).strip(), str(x[4]).strip(), str(x[5]).strip(), str(s), chr(12288)))

if \_\_name\_\_ == '\_\_main\_\_':

a = []

b = []

school\_url = str(input())

school\_html = get(school\_url)

info1(a, school\_html)

n = int(input())

stock\_url = str(input())

stock\_html = get(stock\_url)

info2(b, stock\_html)

print("\n" + "[" + str(school\_url[-4:]) + ":]\n")

printf(a, b, n)

''' test1

https://www.shanghairanking.cn/rankings/bcur/2022

30

http://www.cgedt.com/stockcode/hushi.asp

'''

''' test2

https://www.shanghairanking.cn/rankings/bcur/2020

8

http://www.cgedt.com/stockcode/shenshi.asp

'''

**需求分析：**（包括对问题的理解，解决问题的策略、方法描述）

问题要求通过使用 beautifulsoap 分别对软科排名和炒股一点通两个网站进行信息爬取，按照学校得分排名为该学校“分配”股票；二者做法相似，仅需修改爬取 html 标签名称和对应搜索 tag 即可

**系统设计：**（包括函数定义、参数说明、主程序模块处理过程描述）

引入相应 library

import requests

from bs4 import BeautifulSoup

import bs4

定义解析 url 获取 html 的函数

def get(url):

try:

re = requests.get(url, timeout = 30)

re.raise\_for\_status()

re.encoding = re.apparent\_encoding

html = re.text

except:

html = "Fail"

return html

爬取网站1

def info1(a, html):

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

data = soup.tbody.contents

for t in data:

if isinstance(t, bs4.element.Tag):

info = t.find\_all('td')

a.append([info[0].div.string,

info[1].find(name = 'a', attrs = {'class': 'name-cn'}).string,

info[2].get\_text(), info[3].get\_text(),

info[4].string, info[5].string])

else:

pass

爬取网站2

def info2(b, html):

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

data = soup.div.contents

for t in data:

if isinstance(t, bs4.element.Tag) and t.find(name = 'div', attrs = {'id': 'stockcodelist'}):

info = t.find\_all('a')

for x in info:

b.append(x.get\_text())

else:

pass

打印爬取结果

def printf(a, b, n):

col = "{0:{7}^0}\t{1:{7}^10}\t{2:{7}^10}\t{3:{7}^10}\t{4:{7}^7}\t{5:{7}^7}\t{6:{7}^6}\t"

print(col.format('rank','大学名称','省市','类型','Total Score','School Level', 'Stock Code', chr(12288)))

for i in range(n):

x = a[i]

y = b[i + 1]

s = ""

for c in y:

if c.isdigit():

s += c

print(col.format(str(x[0]).strip(), str(x[1]).strip(), str(x[2]).strip(), str(x[3]).strip(), str(x[4]).strip(), str(x[5]).strip(), str(s), chr(12288)))

**main()**

if \_\_name\_\_ == '\_\_main\_\_':

a = []

b = []

school\_url = str(input())

school\_html = get(school\_url)

info1(a, school\_html)

n = int(input())

stock\_url = str(input())

stock\_html = get(stock\_url)

info2(b, stock\_html)

print("\n" + "[" + str(school\_url[-4:]) + ":]\n")

printf(a, b, n)

**测试结果：**（输入的测试数据及运行结果、正确性、在线测试情况）

**input1:**

**https://www.shanghairanking.cn/rankings/bcur/2022**

**30**

**http://www.cgedt.com/stockcode/hushi.asp**

**ouput1:**

[2022:]

rank 　　　大学名称　　　 　　　　省市　　　　 　　　　类型　　　　Total Score School Level 　股票代码

1 　　　清华大学　　　 　　　　北京　　　　 　　　　综合　　　　 999.4　　 　　　37.6　　　 600000

2 　　　北京大学　　　 　　　　北京　　　　 　　　　综合　　　　 912.5　　 　　　34.4　　　 600001

3 　　　浙江大学　　　 　　　　浙江　　　　 　　　　综合　　　　 825.3　　 　　　34.1　　　 600002

4 　　上海交通大学　　 　　　　上海　　　　 　　　　综合　　　　 783.3　　 　　　35.5　　　 600003

5 　　　复旦大学　　　 　　　　上海　　　　 　　　　综合　　　　 697.8　　 　　　35.9　　　 600004

6 　　　南京大学　　　 　　　　江苏　　　　 　　　　综合　　　　 683.4　　 　　　37.7　　　 600005

7 　中国科学技术大学　 　　　　安徽　　　　 　　　　理工　　　　 609.9　　 　　　40.0　　　 600006

8 　　华中科技大学　　 　　　　湖北　　　　 　　　　综合　　　　 609.3　　 　　　32.3　　　 600007

9 　　　武汉大学　　　 　　　　湖北　　　　 　　　　综合　　　　 607.1　　 　　　32.8　　　 600008

10 　　西安交通大学　　 　　　　陕西　　　　 　　　　综合　　　　 570.2　　 　　　34.2　　　 600009

11 　　　四川大学　　　 　　　　四川　　　　 　　　　综合　　　　 561.7　　 　　　32.0　　　 600010

12 　　　中山大学　　　 　　　　广东　　　　 　　　　综合　　　　 559.4　　 　　　31.4　　　 600011

13 　哈尔滨工业大学　　 　　　黑龙江　　　　 　　　　理工　　　　 549.8　　 　　　32.7　　　 600012

14 　　　同济大学　　　 　　　　上海　　　　 　　　　理工　　　　 548.4　　 　　　32.8　　　 600015

15 　北京航空航天大学　 　　　　北京　　　　 　　　　理工　　　　 546.7　　 　　　33.2　　　 600016

16 　　　东南大学　　　 　　　　江苏　　　　 　　　　综合　　　　 538.5　　 　　　34.9　　　 600017

17 　　北京师范大学　　 　　　　北京　　　　 　　　　师范　　　　 534.7　　 　　　34.6　　　 600018

18 　　北京理工大学　　 　　　　北京　　　　 　　　　理工　　　　 530.9　　 　　　33.1　　　 600019

19 　　中国人民大学　　 　　　　北京　　　　 　　　　综合　　　　 514.7　　 　　　34.2　　　 600020

20 　　　南开大学　　　 　　　　天津　　　　 　　　　综合　　　　 497.4　　 　　　31.3　　　 600021

21 　　　天津大学　　　 　　　　天津　　　　 　　　　理工　　　　 490.8　　 　　　33.0　　　 600022

22 　　　山东大学　　　 　　　　山东　　　　 　　　　综合　　　　 484.3　　 　　　29.2　　　 600023

23 　　　中南大学　　　 　　　　湖南　　　　 　　　　综合　　　　 475.8　　 　　　30.7　　　 600026

24 　　西北工业大学　　 　　　　陕西　　　　 　　　　理工　　　　 467.9　　 　　　32.7　　　 600027

25 　　华南理工大学　　 　　　　广东　　　　 　　　　理工　　　　 466.7　　 　　　30.8　　　 600028

26 　　　厦门大学　　　 　　　　福建　　　　 　　　　综合　　　　 458.8　　 　　　33.7　　　 600029

27 　　　吉林大学　　　 　　　　吉林　　　　 　　　　综合　　　　 450.3　　 　　　31.2　　　 600030

28 　　华东师范大学　　 　　　　上海　　　　 　　　　师范　　　　 440.7　　 　　　32.6　　　 600031

29 　　中国农业大学　　 　　　　北京　　　　 　　　　农业　　　　 433.7　　 　　　31.9　　　 600033

30 　　电子科技大学　　 　　　　四川　　　　 　　　　理工　　　　 432.0　　 　　　32.3　　　 600035

**input2:**

**https://www.shanghairanking.cn/rankings/bcur/2020**

**8**

**http://www.cgedt.com/stockcode/shenshi.asp**

**ouput2:**

[2020:]

rank 　　　大学名称　　　 　　　　省市　　　　 　　　　类型　　　　 Total Score School Level 股票代码

1 　　　清华大学　　　 　　　　北京　　　　 　　　　综合　　　　 852.5　　 　　　38.2　　　 000001

2 　　　北京大学　　　 　　　　北京　　　　 　　　　综合　　　　 746.7　　 　　　36.1　　　 000002

3 　　　浙江大学　　　 　　　　浙江　　　　 　　　　综合　　　　 649.2　　 　　　33.9　　　 000003

4 　　上海交通大学　　 　　　　上海　　　　 　　　　综合　　　　 625.9　　 　　　35.4　　　 000004

5 　　　南京大学　　　 　　　　江苏　　　　 　　　　综合　　　　 566.1　　 　　　35.1　　　 000005

6 　　　复旦大学　　　 　　　　上海　　　　 　　　　综合　　　　 556.7　　 　　　36.6　　　 000006

7 　中国科学技术大学　 　　　　安徽　　　　 　　　　理工　　　　 526.4　　 　　　40.0　　　 000007

8 　　华中科技大学　　 　　　　湖北　　　　 　　　　综合　　　　 497.7　　 　　　31.9　　　 000008