**数据采集工具**

**班 级： 计算机21103**

**学 号： 051321103023**

**姓 名： 张航**

**完成日期： 2022-12-18**

1.程序编写Urllib访问百度官网，并输出当前网页信息，状态码等内容。

**第一问：获取网页信息.py**

*import* urllib.request  
url = 'https://www.baidu.com'  
headers = {  
 'User-Agent':'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/103.0.5060.114 Safari/537.36 Edg/103.0.1264.62'  
}  
request = urllib.request.Request(url=url, headers=headers)  
  
response = urllib.request.urlopen(request)  
  
content = response.read().decode('utf-8')  
print(content)

**部分运行结果：**

D:\python-yuan\python.exe E:/python/爬虫/4\_爬虫\_urllib\_请求对象的定制.py

<!DOCTYPE html><!--STATUS OK--><html><head><meta http-equiv="Content-Type" content="text/html;charset=utf-8"><meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1"><meta content="always" name="referrer"><meta name="theme-color" content="#ffffff"><meta name="description" content="全球领先的中文搜索引擎、致力于让网民更便捷地获取信息，找到所求。百度超过千亿的中文网页数据库，可以瞬间找到相关的搜索结果。"><link rel="shortcut icon" href="/favicon.ico" type="image/x-icon" /><link rel="search" type="application/opensearchdescription+xml" href="/content-search.xml" title="百度搜索" /><link rel="icon" sizes="any" mask href="//www.baidu.com/img/baidu\_85beaf5496f291521eb75ba38eacbd87.svg"><link rel="dns-prefetch" href="//dss0.bdstatic.com"/><link rel="dns-prefetch" href="//dss1.bdstatic.com"/><link rel="dns-prefetch" href="//ss1.bdstatic.com"/><link rel="dns-prefetch" href="//sp0.baidu.com"/><link rel="dns-prefetch" href="//sp1.baidu.com"/><link rel="dns-prefetch" href="//sp2.baidu.com"/><link rel="apple-touch-icon-precomposed" href="https://psstatic.cdn.bcebos.com/video/wiseindex/aa6eef91f8b5b1a33b454c401\_1660835115000.png"><title>百度

**第二问：获取状态码.py**

*import* urllib.request  
  
url = 'http://www.baidu.com/'  
  
response = urllib.request.urlopen(url)  
  
print(response.getcode())

**运行结果：**

D:\python-yuan\python.exe E:/.Firstsnow/study/数据采集工具/数据采集\_爬虫实例\_file/Urllib/状态码.py

200

进程已结束,退出代码0

2.请用BeautifulSoup解析如下HTML代码（该HTML文件名为index.html），并完成如下功能。

（1）运用下述HTML文件创建 BeautifulSoup 对象

1. 输出<title>Hello BeautifulSoup</title>内容
2. 输出Hello Soup内容
3. 输出所有超链接标签<a>
4. 通过CSS选择器找到ID为a\_id\_2的元素内容

<!DOCTYPE **html**>  
<**html lang="en"**>  
<**head**>  
 <**title**>Hello BeautifulSoup</**title**>  
</**head**>  
<**body**>  
<**p class="soup"** >  
 <**b**>Hello Soup</**b**>  
</**p**>  
<**p class="story"**>  
 <**a href="#" class="a\_class\_1" id="a\_id\_1"**>Test1</**a**>  
 <**a href="#" class="a\_class\_2" id="a\_id\_2"**>Test2</**a**>  
</**body**>  
</**html**>

解：

1. 、Index.html文件:

<!DOCTYPE *html*>  
<html *lang*="en">  
<head>  
 <title>Hello BeautifulSoup</title>  
</head>  
<body>  
<p *class*="soup" >  
 <b>Hello Soup</b>  
</p>  
<p *class*="story">  
 <a *href*="#" *class*="a\_class\_1" *id*="a\_id\_1">Test1</a>  
 <a *href*="#" *class*="a\_class\_2" *id*="a\_id\_2">Test2</a>  
</body>  
</html>

1. 、（3）、（4）、（5）解：

Main.py文件

*# import beautifulsoup4  
from* bs4 *import* BeautifulSoup  
  
*# soup = BeautifulSoup(html)*soup = BeautifulSoup(open('index.html'), features="html.parser")  
  
*# print(soup.prettify())*print(soup.title)

print(soup.find\_all(text=re.compile('Hello Soup')))  
print(soup.select('a'))  
print(soup.select('#a\_id\_2'))

运行结果如下：

D:\python-yuan\python.exe "E:/.Firstsnow/study/数据采集工具/数据采集\_爬虫实例\_file/ beautifulsoup4/index.py"

<title>Hello BeautifulSoup</title>

['Hello Soup']

[<a class="a\_class\_1" href="#" id="a\_id\_1">Test1</a>, <a class="a\_class\_2" href="#" id="a\_id\_2">Test2</a>]

[<a class="a\_class\_2" href="#" id="a\_id\_2">Test2</a>]

进程已结束,退出代码0