第二关 第二关练习

2022年11月30日 12:13

1. 知识回顾

本节课为爬虫第二关的练习课。在正式开始练习前，让我们通过下面的思维导图回顾一下第二关的重要知识点：

为了让你能更好的巩固第二关的知识点，做好学习下一关的准备。请你试着完成下方的练习吧：

2. 牛刀小试

选择题 1 - 网络信息的爬取流程

单选题

以下选项中，哪一项展示了正确的爬虫流程？

A.
解析网页 -> 获取网页 -> 存储数据

B.
获取网页 -> 下载网页 -> 解析网页

C.
获取网页 -> 下载网页 -> 存储数据

D.
获取网页 -> 解析网页 -> 存储数据

回答正确

网络爬虫的流程主要可以分为三步，分别是：**获取网页**、**解析网页**以及**存储数据**。

本题的知识包含 网络信息的爬取流程，如果有所遗忘，课后请适当进行回顾复习。

by 风变编程

选择题 2 - 统一资源定位符

已知书籍《全球通史》的网址为： https://wp.forchange.cn/history/6704/ **。**

单选题

请问，上方网址的**协议**、**主机域名**和**路径**，分别为哪部分？

A.
https；wp.forchange.cn；/history/6704/

B.
https；wp.forchange.cn/history；/6704/

C.
https://wp.forchange.cn；/history；/6704/

D.
https；wp.forchange.cn/history；/6704/

分析

答案是 A

（1）https 是协议；
（2）双斜杠 // 后的 wp.forchange.cn 是主机域名，可以理解为服务器在网络上的位置；
（3）/history/6704/ 是路径，表示资源在主机域名下的位置，各级目录会用 / 隔开。

本题的知识包含 统一资源定位符（URL），**如果有所遗忘，课后请适当进行回顾复习。**

选择题 3 - Robots 协议

单选题

想要查看某个网站的 Robots 协议，需要在网站的主机域名后加上以下哪个选项的内容？

A.
/robots.text

B.
/robot.txt

C.
/robots.txt

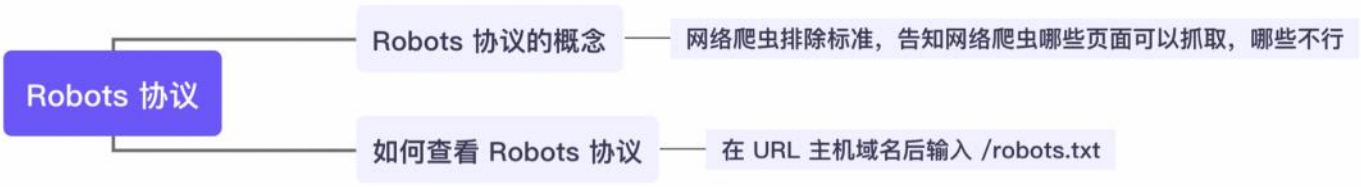
D.
/Robots.txt

分析

答案是 C

在网站的主机域名后加上 /robots.txt，可以查看网站的 Robots 协议。

本题的知识包含 Robots 协议，如果有所遗忘，课后请适当进行回顾复习。



选择题 4 - 爬虫伦理

以下是《今日头条》网站的部分 Robots 协议：

```
User-agent: *
Disallow: /
Allow: /complain/
Allow: /media_partners/
Allow: /about/
Allow: /user_agreement/
Allow: /$
```

多选题

下面哪几个选项是不符合爬虫伦理的？

A.
爬取页面 https://www.toutiao.com/friend_link/ 里的信息

B.
频繁地爬取页面 https://www.toutiao.com/user_agreement/ 里的信息

C.
随意爬取《今日头条》网站的所有信息，并用以商业盈利

D.
爬取页面 https://www.toutiao.com/about/ 里的信息

分析

答案是 A、B、C

选项 A：该网站的 Robots 协议中，Allow 允许的路径中没有提到 /friend_link/，因此不允许爬取该路径的信息；
选项 B：过于频繁地爬取网站，会给服务器产生巨大的压力，网站可能封锁 IP，甚至采取进一步的法律行动；
选项 C：服务器上的数据有产权归属，网络爬虫获取数据如果用来牟利会带来法律风险。

本题的知识包含 爬虫伦理，**如果有所遗忘，课后请适当进行回顾复习。**

选择题 5 - 响应状态码

单选题

服务器成功接收请求并响应的状态码为？页面未找到而导致请求失败的状态为？

A.
200；400

B.
200；404

C.
400；200

D.
204；400
分析
服务器成功接收请求并响应，会返回状态码 200；
页面未找到而导致请求失败，会返回状态码 404。
本题的知识包含 常用响应状态码，如果有所遗忘，课后请适当进行回顾复习。

常见响应状态码解释			
响应状态码	说明	举例	说明
1xx	请求收到	100	继续提出请求
2xx	请求成功	200	成功
3xx	重新定向	305	应使用代理访问
4xx	客户端错误	404	无法提供信息
5xx	服务器错误	503	服务不可用

3. 代码实战
实战 1 - 网页请求
网页请求
练习介绍
已知书籍《全球通史》的网址为：https://wp.forchange.cn/history/6704/ 。
题目要求
请你补充代码，实现以下功能：
1）使用 Requests 库中的 GET 请求，向该网站发起网络请求；
2）打印服务器返回的响应状态码。

```
# 导入 requests 库


# 使用 Get 请求向网站发起请求


# 打印服务器返回的响应状态码
```

答案

```python
# 导入 requests 库

import requests
# 使用 Get 请求向网站发起请求
res = requests.get(' https://wp.forchange.cn/history/6704/ ')

# 打印服务器返回的响应状态码

print(res.status_code)
```

实战 2 - 打印网页响应内容

打印网页响应内容

练习介绍

已知《电影天堂》的网址： https://www.dytt8.net/index2.htm 。

题目要求

请你补充代码，实现以下功能：

1）将请求该网页返回的 Response 对象的编码格式设置为 gbk；

2）获取网页的响应内容。

```python
# 导入 requests 库


# 使用 Get 请求向网站发起请求


# 将 Response 对象的编码格式设置为 gbk


# 打印响应内容
```

答案

```python
# 导入 requests 库

import requests

# 使用 Get 请求向网站发起请求
res = requests.get('https://www.dytt8.net/index2.htm ')

# 将 Response 对象的编码格式设置为 gbk
res.encoding = 'gbk'

# 打印响应内容
print(res.text)
```

恭喜你完成本节课的练习题。