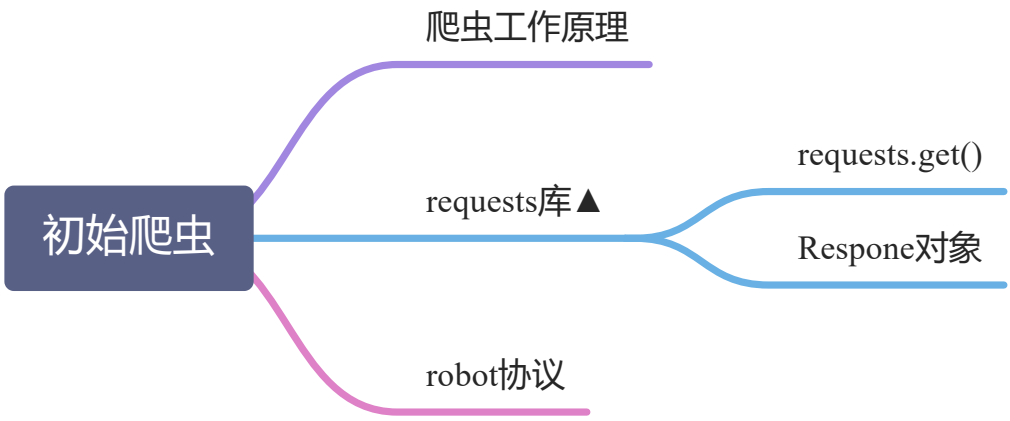
第0课 初识爬虫

一、课程结构导图



注：▲为重点知识点。

二、知识点讲解

2.1 爬虫工作原理

爬虫工作流程四步：

1. 获取数据：爬虫程序会根据我们提供的网址，向服务器发起请求，然后返回数据。
2. 解析数据：爬虫程序会把服务器返回的数据解析成我们能读懂的格式。
3. 提取数据：爬虫程序再从中提取出我们需要的数据。
4. 储存数据：爬虫程序把这些有用的数据保存起来，便于使用和分析。

2.2 requests库

使用场景： 需要使用爬虫向网站发起请求、获取网页数据时，可以使用requests 库来实现。

2.2.1 requests.get()

常用功能： 向网站发起请求，获取网址数据。

常用参数： requests.get() 方法中比较常用的参数有url 、headers、params。可以给参数url 传入网址；给参数headers 传入请求头（第5课学习）；给参数params 传入网址的参数（第5课学习）。**第一个参数url 是必须参数**，其它参数需要根据实际情况增加。

示例：

```python
import requests

url = 'http://……'    # 网址，必须参数
headers = {……}       # 请求头
params = {……}         # 网址参数
proxies = {……}        # IP代理

res = requests.get(url=url, headers=headers, params=params, proxies=proxies)
# 把响应结果赋值给变量res
```

注：调用requests.get() 返回Response对象。

2.2.2 Response 对象

常用功能：查看请求的数据是否成功、给请求成功的数据指定编码、把请求的数据转化为字符串或者二进制数据，方面后续使用。

常用属性1： 通过属性status_code 返回的值查看get() 请求是否成功。

示例1：

```
import requests
res = requests.get('https://www.baidu.com')
print(res.status_code)
# 结果为200时说明请求成功
```

常用属性2： 通过属性content 能把Response对象的内容以二进制数据的形式返回，适用于图片、音频、视频的下载。

示例2：

```
import requests
res = requests.get('https://res.pandateacher.com/2018-12-18-10-43-07.png')
pic = res.content       # 将图片以二进制形式返回

# 写入文件查看
f = open('ppt.png','wb')     # 二进制使用wb写入
f.write(pic)         # 写入返回数据
f.close()
```

注意：该代码运行之后在终端不会显示返回结果（因为没有print()语句）。

常用属性3： 通过属性text 能把Response对象的内容以字符串的形式返回，适用于文字、网页源代码的下载。

示例3：

```
# 代码一：获取小说内容
import requests
url = 'https://localprod.pandateacher.com/python-manuscript/crawler-html/sanguo.md'
res = requests.get(url)     # 获取小说内容
novel=res.text      # res.text 返回字符串给novel
print(novel[:100])  # 字符串支持切片，打印前面100个字符

# 代码二：获取网页源代码
import requests
url = 'https://www.baidu.com'
res = requests.get(url)     # 获取百度网页
print(res.text)
```

代码解析：获取小说内容和获取网页源代码本质上是一样的，把网页的内容爬取到，然后返回给res变量。

常用属性4： 通过属性encoding 可以修改Response对象的编码。一般当返回的内容出现乱码时，才会使用该属性对返回内容的编码进行指定。

示例4：

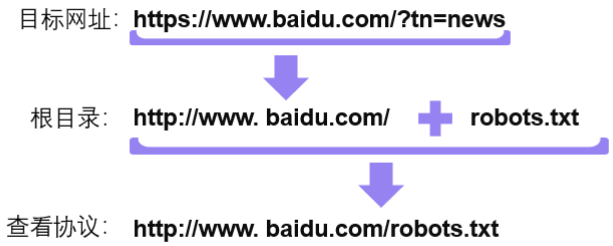
```
import requests
res = requests.get('https://www.baidu.com')
res.encoding = 'utf-8' # 若该行代码注释掉，打印的源代码会有乱码
print(res.text) # 打印网页源代码
```

2.3 robots协议

概念： robots协议是互联网爬虫的一项公认的道德规范，它的全称是“网络爬虫排除标准”（robots exclusion protocol）。该协议主要是用来告诉用户，哪些页面可以抓取，哪些不可以。

使用场景：在进行爬虫前，先查看该协议，了解什么内容不在允许范围内，确定需要爬取的内容在允许范围内之后再进行爬虫。

查看方式：在网页的根目录加上robots.txt，构成新的网址，访问网址即可查看。



协议解读：

- User-agent 表示爬虫类型。User-Agent: * 中 * 指向所有未被明确提及的爬虫。
- Allow 代表允许被访问，Disallow代表禁止被访问。
  - “Allow: /” 表示允许爬取所有目录。同理，“Disallow: /” 表示不允许爬取所有目录。
  - “Disallow: /bh ”表示不允许爬取 “https://www.baidu.com/bh <https://www.baidu.com/bh> ”下的所有内容。