12–建立爬虫军队

1.同步和异步

a. 同步：同步就是一个任务结束才能启动下一个
b. 异步：在一个任务未完成时，就可以执行其他多个任务，彼此不受影响，这就是异步

2.实现异步爬虫

1.协程

一种非抢占式的异步技术

原理是：一个任务在执行过程中，如果遇到等待，就先去执行其他的任务，当等待结束，再回来继续之前的那个任务。在计算机的世界，这种任务来回切换得非常快速，看上去就像多个任务在被同时执行一样。

2.实现

使用gevent,可以在python中实现异步任务

gevent并不是python的标准库，使用前你得先安装它。

安装方法：

window电脑：在终端输入命令：pip install gevent，按下enter键；

mac电脑：在终端输入命令：pip3 install gevent，按下enter键

代码示例：

```
from gevent import monkey
#从gevent库里导入monkey模块。
monkey.patch_all()
#monkey.patch_all()能把程序变成协作式运行，就是可以帮助程序实现异步。
import gevent
import time
import requests
#导入gevent、time、requests。
start = time.time()
#记录程序开始时间。
url_list = ['https://www.baidu.com/',
'https://www.sina.com.cn/',
'http://www.sohu.com/',
'https://www.qq.com/',
'https://www.163.com/',
'http://www.iqiyi.com/',
'https://www.tmall.com/',
```

```
'http://www.ifeng.com/']
#把8个网站封装成列表。
def crawler(url):
#定义一个crawler()函数。
    r = requests.get(url)
    #用requests.get()函数爬取网站。
    print(url,time.time()-start,r.status_code)
    #打印网址、请求运行时间、状态码。
tasks_list = [ ]
#创建空的任务列表。
for url in url_list:
#遍历url_list。
    task = gevent.spawn(crawler,url)
    #用gevent.spawn()函数创建任务。
    tasks_list.append(task)
    #往任务列表添加任务。
gevent.joinall(tasks_list)
#执行任务列表里的所有任务，就是让爬虫开始爬取网站。
end = time.time()
#记录程序结束时间。
print(end-start)
#打印程序最终所需时间。
```

2.实现异步的思路

1.从gevent库里导入了monkey模块，这个模块能将程序转换成可异步的程序。

2.monkey.patch_all()，它的作用就是让程序变成是异步模式。

3.gevent.spawn()来创建任务对象

gevent.spawn()的参数需为要调用的函数名及该函数的参数。比如，gevent.spawn(crawler,url)就是创建一个执行crawler函数的任务，参数为crawler函数名和它自身的参数url。

4.调用gevent库里的joinall方法，能启动执行所有的任务。gevent.joinall(tasks_list)就是执行tasks_list这个任务列表里的所有任务，开始爬取。