第11关

学习到这里，我们已经做过不少爬虫项目，不过我们爬取的数据都不算太大，如果我们想要爬取的是成千上万条的数据，那么就会遇到一个问题：

###

因为程序是一行一行依次执行的缘故，要等待很久，我们才能拿到想要的数据。那我们能不能让多个爬虫一起爬取？

###

那么，在这一关，我们会学习【多线程】，来帮助我们解决这个问题~

###

多协程，是一种非抢占式的异步方式。使用多协程的话，就能让多个爬取任务用异步的方式交替执行。

###

就跟下载电影一样，一部完成下载另外一部，这样效率并不高，而且速度不一定达到最快，这就是【同步】进行。

###

【异步】就是多部电影，同时下载，下载完了哪部就可以立马观看，提高任务进行效率。

###

另外，真正大型的爬虫程序不会单单只靠多协程来提升爬取速度的。比如，百度，可以说是超大型的爬虫程序，它除了靠多协程，一定还会靠多进程，甚至是分布式爬虫，而这是同学可以额外去了解的内容了~

第12关

在这一关我们会进行一个项目，爬取食物热量的信息。既能将上一关学到的协程知识实践起来，又能获得一份食物热量表，还是蛮两全其美的。

###

那么，我们这一关的项目就可以定为：用多协程爬取薄荷网的食物热量。

从爬虫四步（获取数据→解析数据→提取数据→存储数据）入手，开始逐一分析。

###



###

不知道你在做这次项目时的感受是怎样的？会不会因为看不懂一行代码而苦恼许久，但在代码运行通过时又欣喜雀跃呢？

第13关

这一关，我们要学习的就是一个功能强大的爬虫框架——Scrapy。

###

一个爬虫框架里包含了能实现爬虫整个流程的各种模块，就像PPT模板一开始就帮你设置好了主题颜色和排版方式一样。

###



###

我们可以把整个Scrapy框架看成是一家爬虫公司。最中心位置的Scrapy Engine(引擎）就是这家爬虫公司的大boss，负责统筹公司的4大部门，每个部门都只听从它的命令，并只向它汇报工作。

###

在Scrapy爬虫公司里，每个部门都各司其职，形成了很高效的运行流程。

###

不过这一关的内容没有涉及到存储数据的步骤。存储数据需要修改pipelines.py文件。这个知识点我们会到下一关再再进行学习呢。

第14关

这一关，我们会实操一个更大的项目——用Scrapy爬取招聘网站的招聘信息。大家可以借此体验一把当Scrapy爬虫公司CEO的感觉，用代码控制并操作整个Scrapy的运行。

###

在这里，再次会涉及一个爬虫的全程：获取数据，解析数据，提取数据和存储数据。

###

我们先会对项目进行分析，分析企业排行榜的公司信息以及公司详情页面的招聘信息。

###

接着代码实现环节，会创建我们的项目，以及定义item

###

最后则是创建和便携爬虫文件，包括存储文件和修改设置

###

到这里，我们已经学习过用BS解析数据、带参数请求数据、加入cookeies、发邮件、携程，也明白它们的工作原理。

###

另外还有Scrapy框架的组成和工作原理。

只要将两者加以组合，辅以练习，就能轻松掌握。

第15关

很开心能和大家走到第15关。至此，我们已经完成爬虫精进所有关卡的知识学习，大家把掌声送给努力的自己！

###

这一关，我们主要会进行四个目标，大家要保持学习状态，顺利完成噢。

###



###

除了对我们课程做一个总复习以外，吴枫老师还会给大家爬虫进阶路线的指引。

###

另外，所谓的反爬虫，从不是将爬虫完全杜绝；而是想办法将爬虫的访问量限制在一个可接纳的范围，不要让它过于肆无忌惮。这也是一个很重要的方向呢。

###

在学习python中，大家不要忘了我们做项目的思路：确认目标-分析过程-先面向过程一行行实现代码-代码封装。

###

同时分析过程还分为了爬虫四步：获取数据-解析数据-提取数据-存储数据。

###

逢山开路与不甘庸碌，希望大家可以记住这两个词语，关于爬虫的故事我们就到此就将告一段落啦~