# 常见的爬虫框架

**一、什么是Python爬虫框架**

简单来说，Python的爬虫框架就是一些爬虫项目的半成品。比如我们可以将一些常见爬虫功能的实现代码写好，然后留下一些接口，在做不同的爬虫项目时，我们只需要根据实际情况，手写少量需要变动的代码部分，并按照需要调用这些接口，即可以实现一个爬虫项目。

**二、常见的Python爬虫框架**

**1、Scrapy框架**

Scrapy框架是一套比较成熟的Python爬虫框架，是使用Python开发的快速、高层次的信息爬取框架，可以高效的爬取web页面并提取出结构化数据。

Scrapy应用范围很广，爬虫开发、数据挖掘、数据监测、自动化测试等。

Scrapy的官网：<https://scrapy.org/>

**2、Crawley框架**

Crawley也是Python开发出的爬虫框架，该框架致力于改变人们从互联网中提取数据的方式。可以高速爬取对应网站的内容，支持关系和非关系数据库，数据可以导出为JSON、XML等。

官网：<http://project.crawley-cloud.com/>

3、**PySpider：**

一个国人编写的强大的网络爬虫系统并带有强大的WebUI。采用Python语言编写，分布式架构，支持多种数据库后端，强大的WebUI支持脚本编辑器，任务监视器，项目管理器以及结果查看器。Python脚本控制，可以用任何你喜欢的html解析包。

**4、Portia：**

是一个开源可视化爬虫工具，可让使用者在不需要任何编程知识的情况下爬取网站！简单地注释自己感兴趣的页面，Portia将创建一个蜘蛛来从类似的页面提取数据。

简单来讲，它是基于scrapy内核；可视化爬取内容，不需要任何开发专业知识；动态匹配相同模板的内容。

**5、Cola：**

是一个分布式的爬虫框架，对于用户来说，只需编写几个特定的函数，而无需关注分布式运行的细节。任务会自动分配到多台机器上，整个过程对用户是透明的。项目整体设计有点糟，模块间耦合度较高。