001-爬虫简单了解

一、什么是爬虫？

简言之，爬虫可以帮助我们把网站上的信息快速提取并保存下来。

我们可以把互联网比作一张大网，而爬虫（即网络爬虫）便是在网上爬行的蜘蛛（Spider）。把网上的节点比作一个个网页，爬虫爬到这个节点就相当于访问了该网页，就能把网页上的信息提取出来。我们可以把节点间的连线比作网页与网页之间的链接关系，这样蜘蛛通过一个节点后，可以顺着节点连线继续爬行到达下一个节点，即通过一个网页继续获取后续的网页，这样整个网的节点便可以被蜘蛛全部爬行到，网页的数据就可以被抓取下来了。

二、爬虫有什么用？

通过上面的简单了解，你可能大致了解爬虫能够做什么了，但是一般要学一个东西，我们得知道学这个东西是来做什么的吧！

- 比如，我们在网上看到了很多精美的图片，想要保存下来，但是一次次的右键另存为就显得非常的费时费力，那么我们就可以利用爬虫将这些图片快速的抓取下来，极大地节省时间和精力。
- 比如，我们想收集一些新闻门户上的新闻，看一下每天都发生了哪些事情，我们可以写个爬虫把新闻爬取下来，每天运行一次或者设置定时任务定时运行，这样我们可以不用进入网页就能看到新闻，也可以根据关键词进行热点分析。

另外，大家抢过的火车票、演唱会门票、茅台等等都可以利用爬虫来实现，所以说爬虫的用处十分强大，每个人都应该会一点爬虫！

三、爬虫的分类

我们常见的爬虫有`通用爬虫`和`聚焦爬虫`。

- 通用爬虫：针对于百度、谷歌、必应这类搜索引擎类的爬虫程序
- 聚焦爬虫：又名定向爬虫，就是我们平时写的针对某个需求或者某个问题而写的程序

四、所谓的“爬虫学的好，牢饭吃到饱！”

时不时冒出一两个因为爬虫入狱的新闻，是不是爬虫是违法的呀，爬虫目前来说是灰色地带的东西，所以大家还是要区分好`小人`和`君子`，避免牢底坐穿！

👇👇👇👇👇

👆👆👆👆👆

网上有很多关于爬虫的案件，就不一一截图，大家自己上网搜索吧。

有朋友说，“为什么我学个爬虫都被抓，我犯法了吗？”这个目前还真的不好说，主要是什么，目前爬虫相关的就只有一个网站的`robots协议`，这个robots是网站跟爬虫间的协议，用简单直接的txt格式文本方式告诉对应的爬虫被允许的权限，也就是说`robots.txt`是搜索引擎访问网站的时候要查看的第一个文件。当一个搜索蜘蛛访问一个站点时，它首先会检查该站点根目录下是否存在robots.txt，如果存在，搜索机器人

就会按照该文件中的内容来确定访问的范围；如果该文件不存在，所有的搜索蜘蛛将能够访问网站上所有没有被口令保护的页面。也就是说robots协议是针对于通用爬虫而言的，而聚焦爬虫（就是我们平常写的爬虫程序）则没有一个严格法律说禁止什么的，但也没有说允许，所以目前的爬虫就处在了一个灰色地带，这个robots协议也就仅仅起到了一个”防君子不防小人“的作用，而很多情况下是真的不好判定你到底是违法还是不违法的。所以大家使用爬虫尽量不从事商业性的活动吧！好消息是，据说有关部门正在起草爬虫法，不久便会颁布，后续就可以按照这个标准来进行了。

再说一下学习爬虫的具体流程吧！

五、爬虫的大致流程

1. 获取网页

   爬虫首先要做的工作就是获取网页，这里就是获取网页的源代码。源代码里包含了网页的部分有用信息，所以只要把源代码获取下来，就可以从中提取想要的信息了。

   我们用浏览器浏览网页时，其实浏览器就帮我们模拟了这个过程，浏览器向服务器发送了一个个请求，返回的响应体便是网页源代码，然后浏览器将其解析并呈现出来。所以，我们要做的爬虫其实就和浏览器类似，将网页源代码获取下来之后将内容解析出来就好了，只不过我们用的不是浏览器，而是 Python。

   刚才说，最关键的部分就是构造一个请求并发送给服务器，然后接收到响应并将其解析出来，那么这个流程怎样用 Python 实现呢？

   Python 提供了许多库来帮助我们实现这个操作，如urllib、requests等。我们可以用这些库来实现 HTTP 请求操作，请求和响应都可以用类库提供的数据结构来表示，得到响应之后只需要解析数据结构中的`body`部分即可，即得到网页的源代码，这样我们可以用程序来实现获取网页的过程了。

2. 提取信息

   获取网页的源代码后，接下来就是分析网页的源代码，从中提取我们想要的数据。首先，最通用的方法便是采用正则表达式提取，这是一个万能的方法，但是在构造正则表达式时比较复杂且容易出错。

   另外，由于网页的结构有一定的规则，所以还有一些根据网页节点属性、CSS 选择器或 XPath 来提取网页信息的库，如 BeautifulSoup4、pyquery、lxml 等。使用这些库，我们可以高效快速地从中提取网页信息，如节点的属性、文本值等。

   提取信息是爬虫非常重要的部分，它可以使杂乱的数据变得条理、清晰，以便我们后续处理和分析数据。

3. 保存数据

提取信息后，我们一般会将提取到的数据保存到某处以便后续使用。这里保存形式有多种多样，如可以简单保存为TXT文件或JSON文件，也可以保存为我们常用的CSV文件或Excel文件，还可以保存到数据库，如 MySQL 和 MongoDB 等，这个需要看你自己的具体需求，怎样再提取数据方便就保存为什么样的数据。

经过本节内容的讲解，大家肯定对爬虫有了基本了解，接下来让我们一起迈进学习爬虫的大门吧！