数据的几种来源



**一．开源数据表：**



**二．爬虫**

在 Python 爬虫中，基本上会经历三个过程。

1. 使用 Requests 爬取内容。我们可以使用 Requests 库来抓取网页信息。Requests 库可以说是 Python 爬虫的利器，也就是 Python 的 HTTP 库，通过这个库爬取网页中的数据，非常方便，可以帮我们节约大量的时间。
2. 使用 XPath 解析内容。XPath 是 XML Path 的缩写，也就是 XML 路径语言。它是一种用来确定 XML 文档中某部分位置的语言，在开发中经常用来当作小型查询语言。XPath 可以通过元素和属性进行位置索引。

在这一步中，主要用到了两个工具。针对 HTML 页面，可以使用 XPath 进行元素定位，提取数据；针对 JSON 数据，可以使用 JSON 进行解析。

1. 使用 Pandas 保存数据。Pandas 是让数据分析工作变得更加简单的高级数据结构，我们可以用 Pandas 保存爬取的数据。最后通过 Pandas 再写入到 XLS 或者 MySQL 等数据库中。

Requests、XPath、Pandas 是 Python 的三个利器。当然做 Python 爬虫还有很多利器，比如 Selenium，PhantomJS，或者用 Puppeteer 这种无头模式。

另外可以利用一些工具爬虫，比如八爪鱼。

1. **日志**

埋点，日志搜集等等。