第一关 网页追踪者：信息的请求与收集

2022年11月29日　21:20

1. 今天学什么？

首先欢迎你来到爬虫课程。在你踏上爬虫课程的学习之旅前，先来通过这节导学课程来了解一下爬虫这门课程吧。
正所谓知己知彼百战不殆，你对爬虫这门课程的了解越充分，你之后的学习也就能越顺利。
这节课会告诉你爬虫课程的相关信息，为后面的学习做好准备，让你更好更快地适应新的课程学习。
那么让我们开始本节课的学习吧~

2. 爬虫是什么？

我先问你一个问题，你知道网络爬虫是什么吗?
单选题
请选择你觉得正确的答案。
A.
模仿人类自动访问网站的程序。
B.
窃取网络信息的病毒程序。

C.
一款消灭害虫的网络游戏。
回答正确
爬虫的本质就是模仿人类自动访问网站的程序，你在浏览器中做的大部分动作基本都可以通过网络爬虫程序来实现。
网络爬虫指的是能够自动化访问**网站**的程序，其目的一般是提取和保存**网页信息**。
你可以把互联网想象成一张大网，而爬虫便是在网上爬行的蜘蛛。
把网页比作网的一个个节点，爬虫爬到节点就相当于访问了该页面，获取了信息。所以有些时候我们也称之为爬取信息。
那么，爬虫能做什么?
爬虫能做很多事，它结合数据分析可以做商业分析，还可以给应用程序的开发提供数据支持，比如：爬一爬北京近两年二手房成交均价是多少？情人节期间深圳酒店的价格…等等。
在过去，数据的获取是通过一个个的点击访问，人工复制实现的。这种重复性的手动工作不仅浪费时间，还容易出差错。
并且受限于数据收集的速度，在瞬息万变的互联网世界，手动获取的数据往往不具备时效性。
网络爬虫技术的飞速发展完美解决了手动收集数据的痛点，不但实现了自动化还保证了数据的实时获取。
在数据量爆发式增长的互联网时代，网站与用户的沟通，本质上就是数据的交换。
以百度为例，你在搜索的时候仔细看，会发现每个搜索结果下面都有一个百度快照。

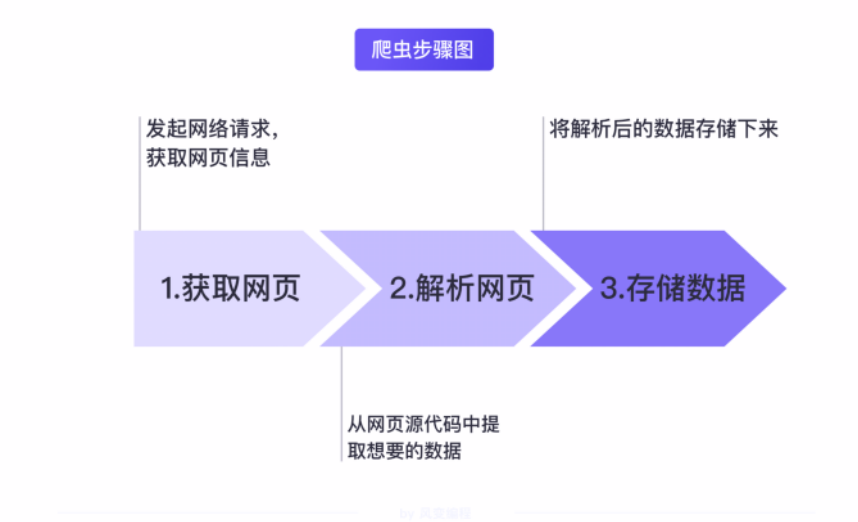
点击百度快照，你会发现网址的开头有 baidu 这个词，也就是说这个网页属于百度。

这是因为，百度这家公司会源源不断地把千千万万个网站爬取下来，存储在自己的服务器上。
你在百度搜索的本质就是在它的服务器上搜索信息，你搜索到的结果是一些超链接，在超链接跳转之后你就可以访问其它网站了。
爬虫让这些搜索巨头有机会朝着人工智能的未来迈进，因为人工智能的发展离不开海量的数据。而每天使用这些搜索网站的用户都是数以亿计的，产生的数据自然也是难以计量的。

从搜索巨头到人工智能巨头，这是一条波澜壮阔的路。而我们应该看到，事情的源头，是我们今日所要学习的“爬虫”。
现在，我们对爬虫有了初步的印象，知道了爬虫是什么，能做什么，那我们接下来看看，爬虫是如何做到这些事的。

2.1 网络信息的爬取流程

网络爬虫的流程其实十分简单，主要可以分为三步，分别是：**获取网页**、**解析网页**以及**存储数据**。
获取网页，顾名思义就是获取网页信息，在网络爬虫技术中这里获取的就是网页源代码。
解析网页，指的是从网页源代码中提取想要的数据，由于网页的结构有一定的规则，配合 Python 的一些第三方库我们可以高效地从中提取网页数据。
存储数据，这一步很简单，就是将数据存储下来。而且我记得你之前有写过 csv 模块的代码，所以我后面也会跟你复习一遍这些知识。



大部分网络信息的爬取都绕不开这三个步骤，所以接下来的课程都将会围绕这三个步骤带你入门爬虫技术。
学完入门爬虫技术，你就可以编写简单的爬虫程序来自动化爬取闪光读书网站里的信息。
下面我来考察一下你对爬虫程序的认识~
单选题

以下选项的描述中，哪一项不能通过爬虫去实现。

A.
自动获取下厨房网菜谱信息的程序。
B.
自动获取豆瓣网电影信息的程序。

C.
自动获取硬盘中 Excel 表格数据汇总信息的程序。
回答正确
爬虫的本质就是模仿人类**自动访问网站**的程序，C 选项自动获取的是硬盘上文件的信息，所以不属于爬虫技术。
好了，关于爬虫知识我先说到这里，后面的课程还会有更详细介绍。下面让我们来到学习指南环节，了解爬虫课程的学习路径、课程结构以及爬虫的学习误区、学习方法。

3. 学习指南

3.1 爬虫学习路径

如上图所示，在这次的爬虫课程，你将由导学课——知识课——项目课——实操课的顺序经历从小蜘蛛到蜘蛛侠的成长历程。
在导学课之后，你会先在知识课中按照爬虫的流程学习必备的基础知识，再通过几节项目课逐步提高爬虫技术。
最后，当你掌握了一定的爬虫技术，你就可以来到我们的实操课大展身手，成为一代蜘蛛侠！

这些知识课、项目课、实操课具体是什么样的呢？接下来我们来一起了解一下课程结构。

3.2 爬虫课程结构

3.2.1 知识课

知识课顾名思义，就是学习爬虫基础知识的地方。你需要完成网络请求、网页代码、网页解析工具等必备知识的学习才能去上手爬虫项目。所以知识课的结构会和你学习基础语法时一样，注重基础知识的讲解，具体结构如下：

1）**课前复习**：对上一关卡的知识，或前面你已学过，且当前关卡会使用到知识进行复习；
2）**今天学什么？**：提前告知关卡的学习内容，以及哪些知识点是本关的重难点；
3）**当天关卡知识**：当天所学关卡的知识讲解及课内练习；
4）**总结与复习**：回顾总结关卡所学的新知识，并对这些知识进行练习。

3.2.2 项目课

项目课就是我们学习爬虫实操的地方，具体结构如下：

1）**项目代码**：明确项目需求，体验项目代码，拆解代码功能；
2）**课前复习**：对上一关卡的知识，或前面你已学过，且当前关卡会使用到知识进行复习；
3）**当天关卡知识**：当天所学关卡的知识讲解及课内练习；
4）**程序实现与总结**：梳理代码流程，整合、复现项目代码并对本关内容进行总结。

3.2.3 实操课

实操课是我们修炼爬虫技术，成为一代蜘蛛侠的练功房！具体结构如下：

1）**项目分析**：说明项目目标；
2）**网页分析**：手动访问目标网站，分析完成这一项目的项目目标，都需要做什么样的操作；
3）**代码实现**：根据目标分步骤实现代码；

4）**项目总结**：梳理代码流程，整合、复现项目代码并对本关内容进行总结。

可以看到，我们这次的课程形式十分丰富，那么在学习中我们有哪些误区需要避免呢?

3.3 爬虫课程的学习误区

误区1：认为同一个代码可以爬取不同网页的信息

有这么一部分学员为了尽快解决自身的实际问题，所有的练习都去复制粘贴答案，或者死记硬背项目代码，一节课学下来什么也不记得。

然后以为自己能爬虫了，直接把代码里的网络链接替换成了自己想要爬的网站，结果各种报错。这里你需要注意，爬虫程序不是万能钥匙。

不同网页结构的爬虫代码也是不一样的，我们要学习的是探索网页结构，在各色各样的网站中找到它的爬取方法。

我们的课程介绍了很多爬虫的项目，这些项目是为了让你熟悉网络爬虫的流程，掌握代码逻辑，你在学习这些网站的爬取方法时一定要注意举一反三。

这样之后当你遇到想要爬取的网站和我们课程的项目网站结构类似时，就可以根据代码逻辑自主完成这些爬虫程序的开发。

误区2：认为网络上的所有信息都可以使用爬虫技术获取

很多学员刚接触爬虫，还不是很了解网络爬虫的伦理规范，其实网络上的信息并非都能随意使用。

滥用爬虫程序可能会侵犯别人隐私，占用网站资源，甚至会触犯法律风险，引发牢狱之灾。

在网络世界中，有一个专门的 **Robots 协议**来规范爬虫，维护网络秩序。它可以告诉网络爬虫程序哪些内容是可以获取的，哪些内容是不能获取的。

具体查看协议的方法以及协议内容到了后面的课程我们自然会学到，接下来我们再来了解一下爬虫的学习方法。

3.4 爬虫学习方法

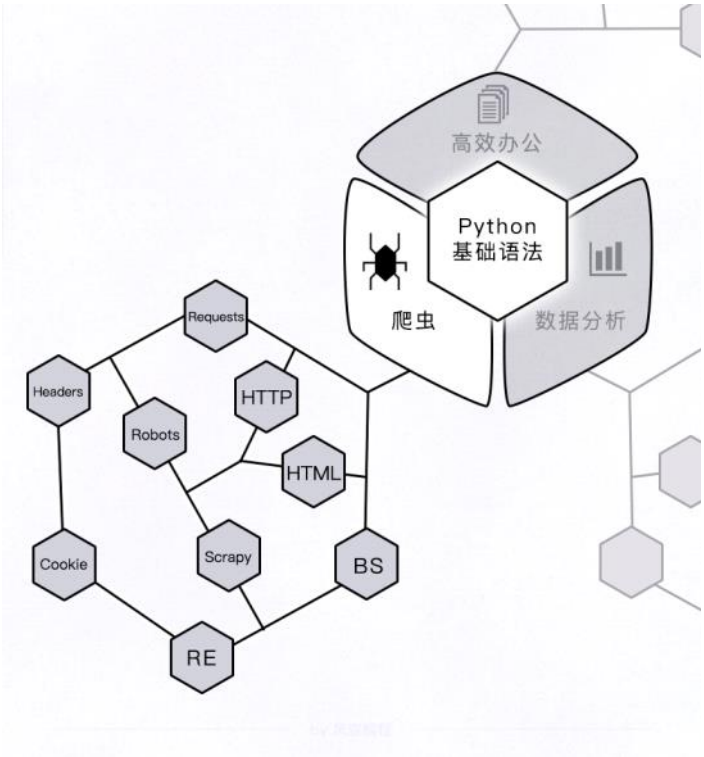
在基础语法课程中，我们学习的是跟 Python 语法相关的基础知识，结构比较单一，目的也是让你掌握编程的基本能力。

优秀的武术大师往往都要经历十年如一日的扎马步、站桩等基本功的练习，基本功练扎实了才有资格习武。

编程也是如此，基础语法熟练了才能继续往下学习编程在各领域的应用。

如今你已投入爬虫“流派”，你要提前做好准备，因为这里的知识结构跟你学习基础语法时的结构截然不同。

在爬虫课程中你将会继续深入学习一些 Python 模块与库的使用，除此之外你还会学习大量的网络请求、爬虫的原理知识以及工具使用。



对于这些不同的知识种类如果你还是沿用之前的方法去学那就大错特错了。

下面我来向你介绍几个适用于不同知识的学习方法~

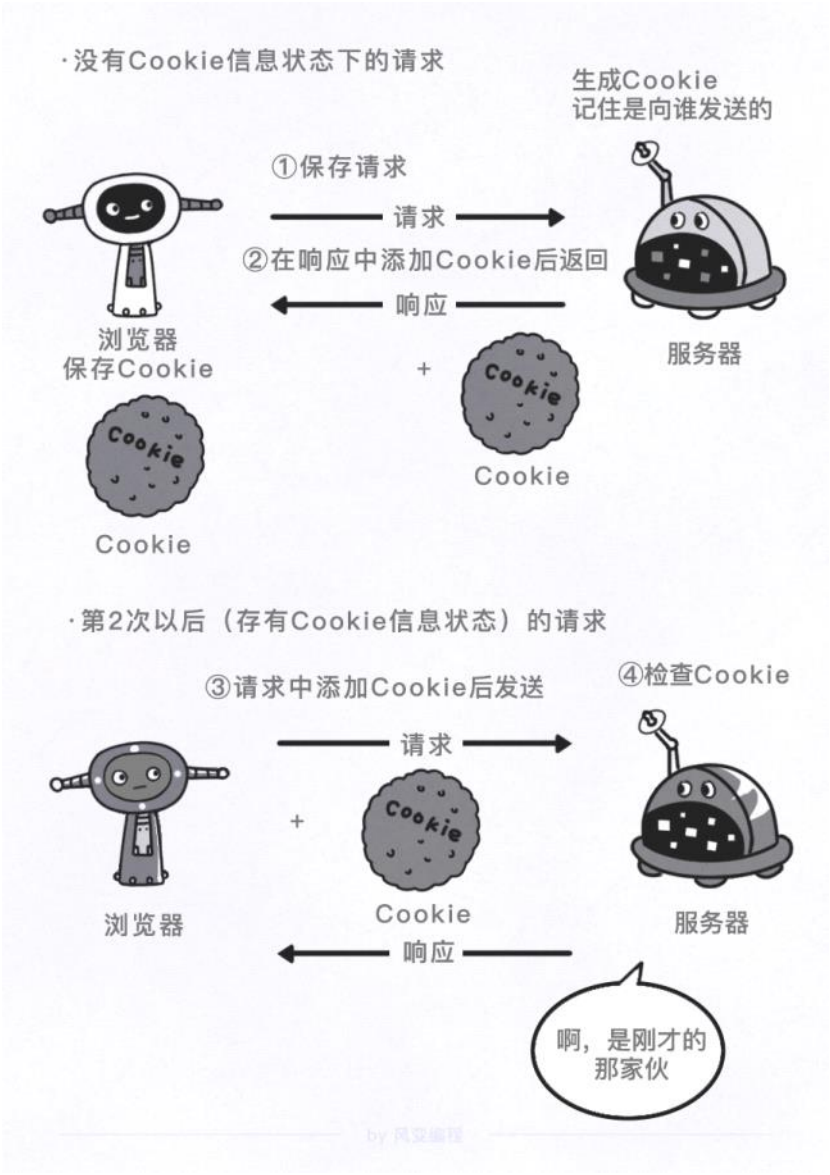
1. 使用画图的方法去学习网络原理

爬虫的本质是通过程序模仿人类上网的过程，你必须了解一些基本的网络原理才能写好爬虫程序。

对于这些网络原理，你更需要的是去**理解**，而不是死记硬背。当你感觉理解起来很痛苦时，你可以动手将你的理解画出来。

比如之后我们会学到的网络请求，它指的是我们从浏览器点开一个网络链接再到我们看到实际网页这一过程中的工作原理。

这一原理只看文字的解释很抽象，不太好记。但如果动手将这些工作原理画出来，你会发现它实际也没有那么难理解。



以上图为例，这一方法并不需要什么绘画技巧，重要的是将你的想法画出来，以此来加深你对这一知识点的印象。

2. 查询网络文档去学习 HTML 语言

如上图所示，由于爬虫获取的信息大部分都是网页的源代码，这些源代码基本都是使用 HTML 语言编写的，所以 HTML 语言对于爬虫的学习十分重要。
看到这里先不要慌，后面的课程并不是要你像学习 Python 似的去学习一门新的编程语言。
我们课程的主角是爬虫，不是网页开发，所以对于 HTML 语言的掌握要求并没有那么高。
你只需要在课程上简单理解 HTML 语言的标签结构，之后遇到不熟悉的标签再去网上查询即可满足爬虫对于 HTML 语言掌握水平的要求。

上图展示的是 MDN 网站上的 HTML 教程，现在看不懂没关系，后面的课程也会向你介绍具体网络文档的查阅方法。

3. 通过实践去学习浏览器开发者工具的使用

后面的课程会向你介绍浏览器开发者工具的使用，目的是在浏览器查看网页结构以及一些网络请求信息。

上图展示的就是浏览虎扑网时，打开浏览器开发者工具查看网页元素。

这部分的学习可以专注于实践，因为几乎所有的浏览器都有开发者工具，我们可以在日常网上冲浪的时候打开它，熟悉基本操作。

4. 爬虫相关的模块与库需要坚持代码练习

最后，我们学习的爬虫技术毕竟是要通过 Python 语言来实现，所以之后课程依然会包含大量 Python 模块与库的使用。

对于 Python 模块与库的知识拓展，你就可以跟之前一样通过练习、实操的方式熟悉这些代码，毕竟基本功的训练是一个持之以恒的过程。

这些学习方法都记住了吗？我来出道题考考你~

单选题

以下哪个选项没有按照我刚才介绍的方法去学习爬虫课程的知识？

A.

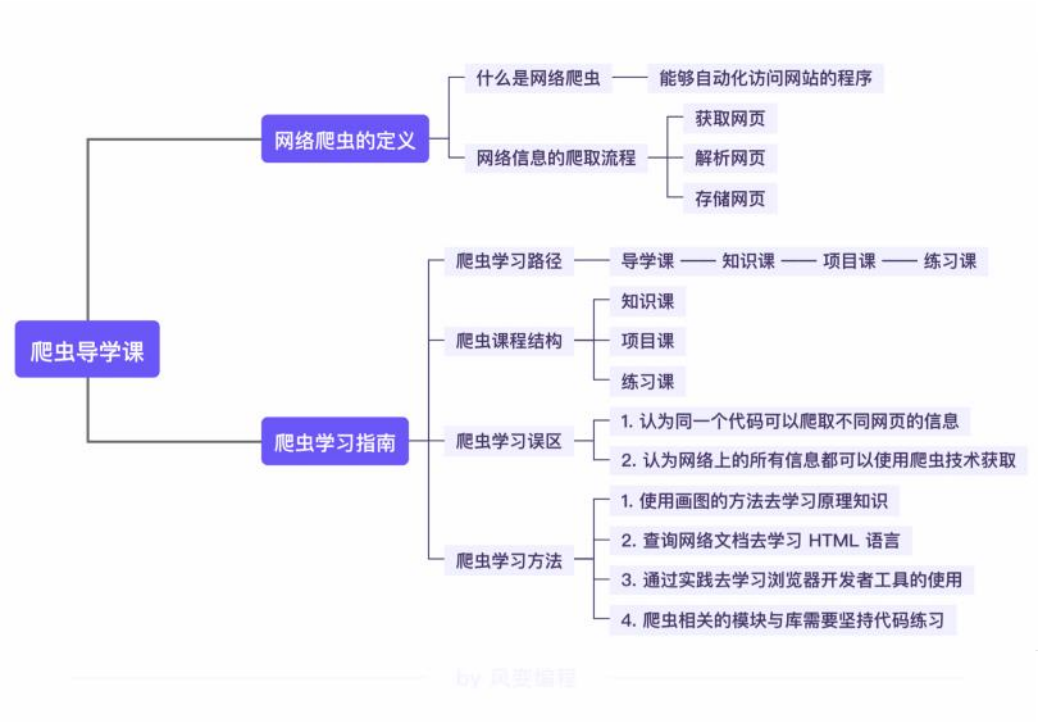
画草图去理解网络请求的基本原理

B.

通过开发一个新的网页去学习爬虫需要的 HTML 语言基础

C.
在浏览器刷知乎的时候打开浏览器的开发者工具，熟悉它的操作

D.
通过代码练习去熟悉 Python 中与爬虫相关的模块与库的使用
回答正确
HTML 语言的学习我推荐你用的时候查阅网络文档就好了
好了，今天的学习到这里就结束了。
最后，我们一起来回顾一下，你今天又 get 到了什么新知识。



导学课程已经结束了，你可以在接下来的正式课程里“施展拳脚”啦！新的故事正在开启，值得期待哦。