书稿体例结构如下：

第1章——一级标题

1.1——二级标题

1.1.1——三级标题

1.——四级标题

（1） 可做标题，也可做内容（做内容，句末加标点）

① 做内容，句末加标点

注意事项：

1. 尽量避免口语化叙述，采用书面语进行叙述。

为使书稿语言简洁、连贯、通顺，不致产生歧义或逻辑错误，叙述时一定要有统一的叙述人称。一般要求全部采用第三人称的叙述角度，并且尽量使用祈使句的表达方式。即在书中尽量不用“你”“我”“我们”等代词，“读者”“用户”也尽量少用，但要注意不要让句子缺成份。必要时可用“笔者”代替“我”。

1. 图、表要先在文中用话引出，再出现图、表。
2. 英文第一次出现，要加中文解释；英文字母注意大小写问题。
3. 句子句末要加标点。
4. 图片要清晰，自己做的图直接用可编辑的源文件，便于修改。

# 第1章 爬虫基础

互联网上有浩瀚的数据资源，要想爬取这些数据就离不开爬虫。本章将对Python爬虫的一些技术内容进行介绍。

本章重点讲解以下内容。

* 了解爬虫的相关概念。
* Python语法的一些基础知识。
* 网页相关的知识包括HTML和HTTP。

## 1.1 认识爬虫

网络爬虫又被称为网页蜘蛛，网络机器人，通俗来讲，爬虫就是一段程序，通过这段程序可以在网站上获取需要的信息，如文字、视频、图片等数据。此外，网络爬虫还有些不常用的名字，如蚂蚁、自动索引、模拟程序或者蠕虫等。

爬虫的设计思路包含以下3步。

（1）明确需要爬取的网页的URL地址。

（2）通过HTTP请求来获取对应的HTML页面。

（3）提取HTML里面的内容。这里有两种情况：如果有用的数据，就保存起来；如果是需要继续爬取页面，重新指定第2步。

为了更方便理解爬虫，首先介绍大数据时代获取数据的方式。

### 1.1.1 大数据时代获取数据的方式

随着社会高速发展，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据成为高科技时代的产物。那么在大数据时代，获取数据的方式有哪些？

（1）企业产生的数据：百度搜索指数、腾讯公司业绩数据、阿里巴巴集团财务及运营数据、新浪微博微指数等。数据平台购买数据：数据堂、国云数据市场、贵阳大数据交易所等。

（2）政府/机构公开的数据：国家统计局数据、中国人民银行调查统计司、世界银行公开数据、联合国数据库、纳斯达克、新浪美股实时行情等。这些数据通常都是由各地政府统计上报，或者由行业内专业的网站、机构等提供。

（3）数据管理咨询公司：麦肯锡、埃森哲、尼尔森、中国互联网络信息中心、艾瑞咨询等。这样的公司通常拥有庞大的数据团队，一般通过市场调研、问卷调查、固定的样本检测与各行各业的公司合作、专家对话来获取数据，并根据客户需求制定商业解决方案。

（4）爬取网络数据：如果数据市场上没有需要的数据，或者价格太高不愿意购买，那么可以利用爬虫技术，获取网站上的数据。

第4种方式爬取网络数据就是本书要重点介绍的内容，下面就一起认识爬虫。

### 1.1.2 爬虫的分类

爬虫有很多种类型，根据使用场景的不同，可以将爬虫分为通用爬虫和聚焦爬虫两种。下面了解下这两种爬虫。

1.通用爬虫

随着网络的迅速发展，万维网成为大量信息的载体，如何有效地提取并利用这些信息成为巨大的挑战。搜索引擎的核心就是通用爬虫，例如，传统的通用搜索引擎谷歌、百度等，作为一个辅助人们检索信息的工具成为用户访问万维网的入口和指南。

通用网络爬虫是搜索引擎爬取系统的重要组成部分，主要目的是将互联网上的网页下载到本地，形成一个互联网内容的镜像备份。

通用网络爬虫是从互联网中搜集网页、采集信息。采集的网页信息可以为搜索引擎建立索引提供支持，决定着整个引擎系统的内容是否丰富，信息是否即时，因此其性能的优劣直接影响着搜索引擎的效果。搜索引擎网络爬虫的基本工作流程如下。

（1）爬取网页。

如图1-1所示，取一部分种子URL，将这些种子放入待爬取URL队列。取出待爬取URL，解析DNS得到主机的IP，并将URL对应的网页下载下来，存储到已下载的网页库中，再将这些URL放进已爬取的URL队列中。将已爬取的URL队列中的URL进行分析，分析其他的URL，并且将URL放入待爬取的URL队列中，进入下一个循环。



图1-1 搜索引擎网络爬虫的基本工作流程

在图1-1中，搜索引擎获取新网站的URL，是输入了一定的规则，如标注为nofollow的链接或Robots协议。

提示：网络爬虫排除标准（Robots Exclusion Protocol，Robots），也叫作爬虫协议、机器人协议等，网站通过Robots协议告诉搜索引擎哪些页面可以爬取，哪些页面不能爬取。例如，淘宝网：https://www.taobao.com/robots.txt，腾讯网：http://www.qq.com/robots.txt。

（2）数据存储。

搜索引擎通过爬虫爬取的网页，将数据存入原始页面数据库。其中，页面数据与用户浏览器得到的HTML是一致的。搜索引擎蜘蛛在爬取页面的同时，也做重复内容检测，一旦遇到访问权重很低的网站上有大量抄袭、采集或复制的内容，将有可能不再继续爬行。

（3）预处理。

将搜索引擎将爬虫爬取的页面，进行预处理，比如提取文字、中文分词、索引处理等。

除了HTML文件外，搜索引擎通常还能爬取和索引以文字为基础的多种文件类型，如 PDF、Word、WPS、XLS、PPT、TXT文件等。

（4）提供检索服务，网站排名。

搜索引擎在对信息进行组织和处理后，为用户提供关键字检索服务，将用户检索相关的信息展示给用户。同时会根据页面的PageRank值，也就是链接的访问量排名，来进行网站排名，Rank值高的网站在搜索结果中会排名较前。

通用爬虫虽然功能很强大，但是也存在一定的局限性。

（1）不同领域、不同背景的用户往往具有不同的检索目的和需求，通用搜索引擎返回的结果包含大量用户不关心的网页内容。

（2）通用搜索引擎的目标是尽可能地扩大网络覆盖范围，有限的搜索引擎服务器资源与无限的网络数据资源之间的矛盾将进一步加深。

（3）网络技术不断发展，万维网数据形式越来越丰富，如图片、数据库、音频、视频多媒体等不同数据大量出现，通用搜索引擎往往对这些信息含量密集且具有一定结构的数据无能为力，不能很快地发现、获取。

（4）通用搜索引擎大多提供基于关键字的检索，难以支持根据语义信息提出的查询。

2.聚焦爬虫

为了解决通用爬虫的局限性问题，定向爬取相关网页资源的聚焦爬虫应运而生。聚焦爬虫是一个自动下载网页的程序，它根据既定的爬取目标，有选择地访问万维网上的网页与相关的链接，获取所需信息。与通用爬虫不同的是，聚焦爬虫并不追求大的覆盖范围，而将目标定为爬取与某一特定主题内容相关的网页，为面向主题的用户查询获取数据资源。

提示：从第2章到第21章要学习的网络爬虫，就是聚焦爬虫。