1.1 爬虫程序开发环境

1.1.1 爬虫程序简介

爬虫程序是一组客户端程序，它的功能是访问 web 服务器，从服务器中获取网页代码，网页代码中包含了很多各种各样的数据信息，程序从中提取所关心的数据，把数据整理后存储在本地的数据库中，这些数据将应用在数据分析等领域中。

例如我们要想知道一个城市一段时间内的天气预报，就可以设计一组程序去访问有天气预报数据的网站，如图 1-1-1 是一个天气预报的网站代码。爬虫程序的任务就是要从这一大段复杂的代码中提取所关心的天气状况、温度、风力等数据，把这些数据存储在数据库中。

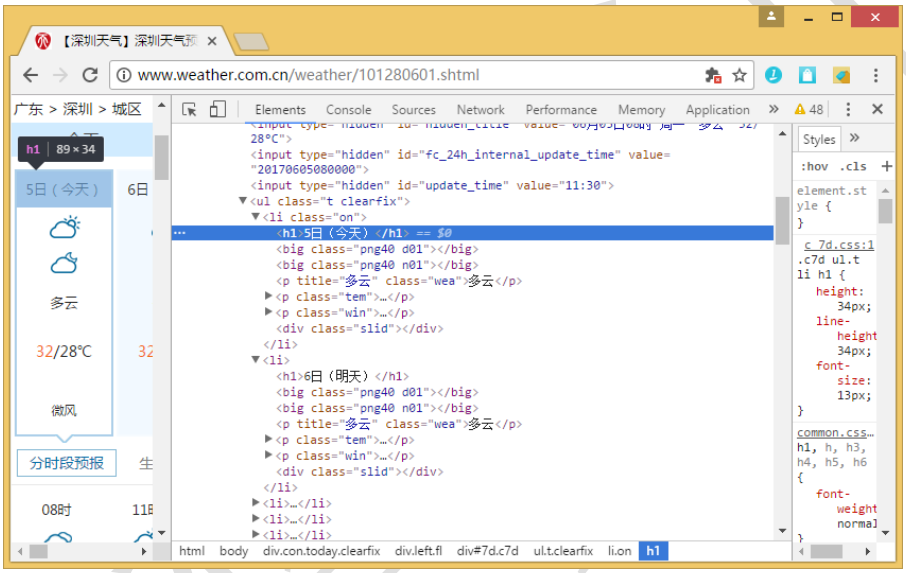

图 1-1-1 网页代码

编写一个爬虫程序可以使用 Python、Java、C++、C#等各种常用的开发语言，但是使用 Python 是比较简单也是比较流行的一种方法。

爬虫程序爬取的数据往往很多，而且相关的数据往往分布在很多不同的网页中，甚至分布在相关联的很多不同的网站中，爬虫程序必须能按链接自动往返于这些不同的网站中去爬取数据，一个爬虫程序爬取成百万上千万条的数据是常有的事，怎么样设计一个高效率的爬虫程序成了我们学习的重点。

1.1.2 Python 开发环境搭建

Python 是一种面向对象的解释型计算机程序设计语言，由荷兰人 Guido van Rossum 于 1989 年发明，第一个公开发行版发行于 1991 年。

Python 语言具有以下特点：

- 开源、免费、功能强大；
- 语法简洁清晰，强制用空白符(white space)作为语句缩进；
- 具有丰富和强大的库，无论你想实现什么功能几乎都有一个库能满足你的要求；
- 易读、易维护，被大量用户所欢迎的、用途广泛；
- 是解释性语言，变量类型可变，类似 JavaScript；

Python 安装后自己带有一个命令行工具与小的 IDE 程序，但是这个 IDE 很弱，因此在

此基础上可以搭配第三方的各种 IDE 开发工具，下面介绍几种主流的开发工具与环境。

1、Python 自带开发环境

Python 的开发环境也十分简单，用户可以到官网 https://www.python.org/中直接下载 Python 的程序包。目前 Python 有两个主流的版本，一个是 Python2.7，另外一个是 Python3.6，这两个版本在语法上有些差异，本教程主要使用 Python3.6。

下载 Python3.6 程序包后直接安装，选择安装目录，在短短几分钟类就可以完成安装。Python 安装完毕后在 WIndows 的启动菜单中就可以看到 Python36 的启动菜单，启动 Python36 可以看到 Python 的命令行界面。这个环境是命令行环境，只能运行一些简单的测试语句，显然不能用它来编写程序。Python 自带一个 IDE，但是这个 IDE 的功能十分有限，不适合开发 Python 工程项目。

2、PyCharm 与 Python 的开发环境

一个比较流行的开发环境是 PyCharm，它的风格类似 Eclipse，是一种专门为 Python 开发的 IDE，带有一整套可以帮助用户在使用 Python 语言开发时提高其效率的工具，比如调试、语法高亮、Project 管理、代码跳转、智能提示、自动完成、单元测试、版本控制。

读者可以到 PyCharm 的官网 http://www.jetbrains.com/pycharm/去下载免费的 PyCharm Community 版本，这个版本虽然不及收费的 Professional 专业版本功能强大，但对于一般应用已经足够了。

3、Anaconda 与 Python 的开发环境

一个比较流行的开发环境是 Anaconda，这个程序比较庞大，但它是一个十分强大的 Python 开发环境，它自己带 Python 的解释器，也就是说安装 Anaconda 时就自动安装 Python 了，同时它还带有一个功能强大的 IDE 开发工具 Spider。Anaconda 最大的好处是可以帮助用户找到与安装 Python 的各种各样的开发库，使得 Python 的开发十分方便与高效。另外 Anaconda 对 Windows 用户十分有用，因为 Python 的一些开发库在 Windows 环境下安装常常出现这样那样的问题，而 Anaconda 能顺利解决这些问题。读者可以到官网 https://www.continuum.io/downloads 下载 Anaconda。