项目 5 爬取商城网站数据

5.1 商城网站项目背景与目标

任务：

一个真实的网站的很多网站的网页都不是静态的 HTML 文档，大部分都包含 JsvaScript 程序，很多信息都是通过 JavaScript 程序处理后才显示出来的，使用普通的爬虫程序不能爬取网站的数据，我们必须使用一种能执行网页中 JavaScript 程序的工具，才能编写爬虫程序爬取这类网站的数据，Selenium 就是这样一种工具。在这个项目中我们使用 Selenium 编写爬虫程序爬取京东商城的手机数据。

5.1.1 爬取模拟商城网站数据

1、准备网站数据

创建一个 project5 的项目，在 project5 文件夹中有一个 phones.csv 文件，存储了手机的数据，前面几行如下：

ID,mMark,mPrice,mNote,,mImage

000001,荣耀 9i,1198.0,荣耀 9i 4GB+64GB 幻夜黑 移动联通电信 4G 全面屏手机 双卡双待,000001.jpg

000002,荣耀 8X,1399.0,荣耀 8X 千元屏霸 91%屏占比 2000 万 AI 双摄 4GB+64GB 幻夜黑 移动联通电信 4G 全面屏 双卡双待,000002.jpg

000003,小米 8,2299.0,小米 8 全面屏游戏智能手机 6GB+64GB 蓝色 全网通 4G 双卡双待,000003.jpg

.....

其中每行都是一款手机的数据，各个段之间用逗号","分开，第一段为 ID 编号，第二段是品牌，第三段是价格，第四段是说明，第五段是图像名称。

在 project5\images 文件夹中存储了各个手机的图像，如图 5-1-1 所示。

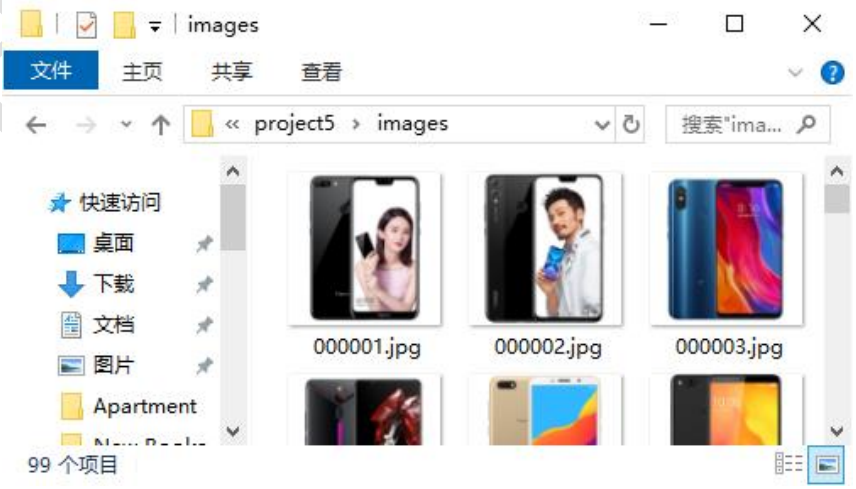

图 5-1-1 手机图像

2、创建商城网站

使用 phones.csv 的数据建立一个手机商城网站，如图 5-1-2 所示，其中网页的很多数据是 JavaScript 控制的，例如各个翻页按钮是<input type='button'>按钮，点击按钮时执行对应的 JavaScript 函数实现翻页，我们使用 Selenium 设计一个爬虫程序爬取所有的手机数据与图像。

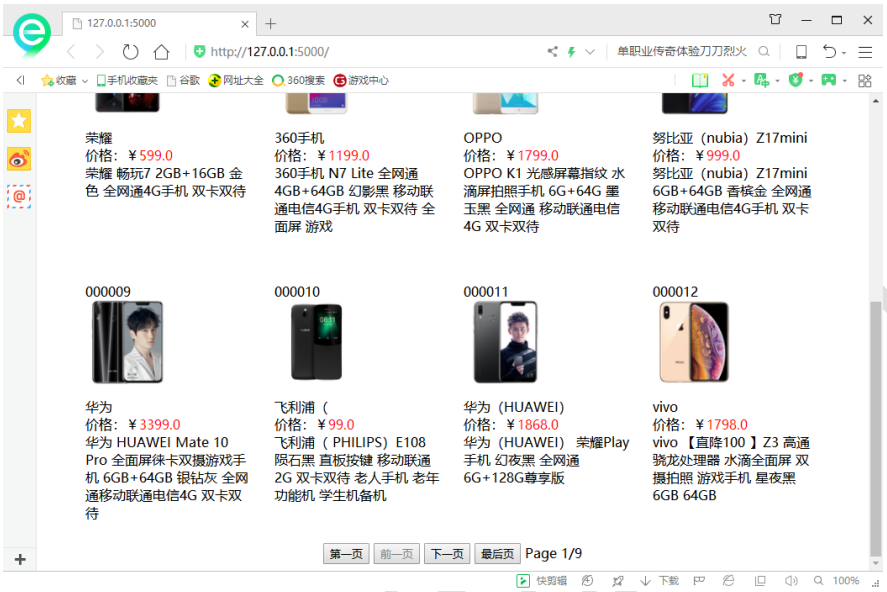

图 5-1-2 模拟商城网站

5.1.2 爬取京东商城网站数据

我们进入京东商城网站，在关键字输入框输入"手机"后回车可以看到如图 5-1-3 所示的页面，分析京东商城的网页发现很多数据是 JavaScript 控制的，我们使用 Selenium 设计一个爬虫程序爬取所有的手机数据与图像。

图 5-1-3 京东商城网站