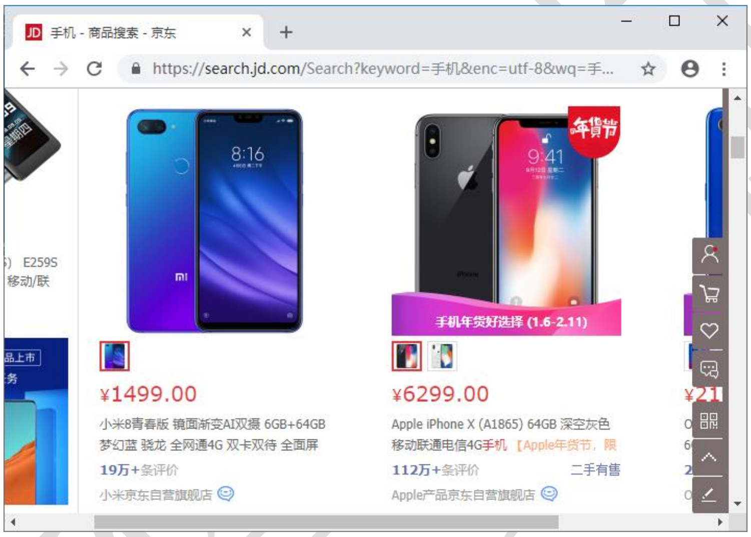
5.8 实战项目 爬取京东商城网站数据

任务：

京东商城有大量的商品数据，在搜索框中输入某类商品，例如"手机"，就可以看到近百页手机的信息。现在我们使用 Selenium 编写一个爬虫程序，自动在输入框输入"手机"，自动翻页爬取所有手机的数据与图像，并保存到数据库。

5.8.1 解析网页 HTML 代码

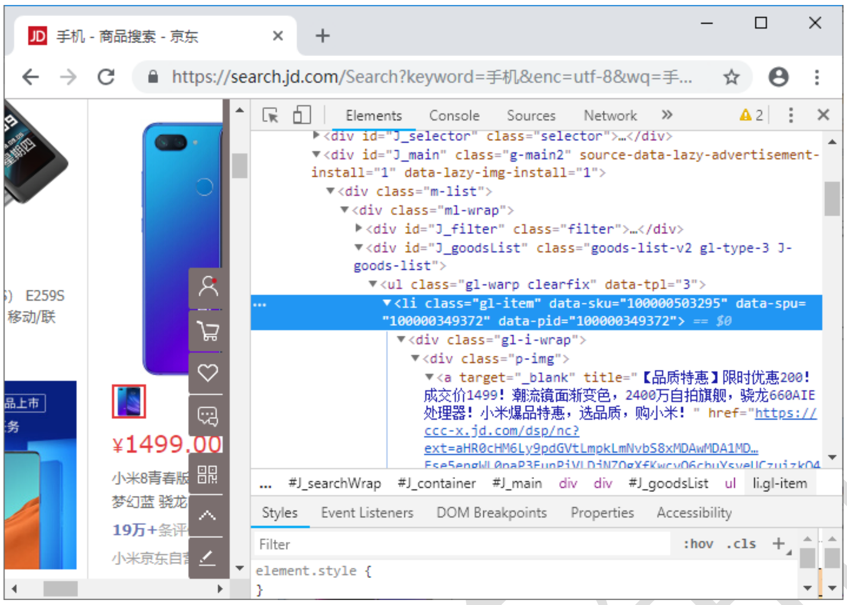
京东网站有很多各种各样的手机，用 chrome 浏览器进入京东网站输入“手机”，那么就会看到如图 5-8-1 所示的画面。



手机 - 商品搜索 - 京东
https://search.jd.com/Search?keyword=手机&enc=utf-8&wq=手...
年货节
手机年货好选择 (1.6-2.11)
¥1499.00
小米8青春版 镜面渐变AI双摄 6GB+64GB 梦幻蓝 骁龙 全网通4G 双卡双待 全面屏
19万+条评价
小米京东自营旗舰店
¥6299.00
Apple iPhone X (A1865) 64GB 深空灰色 移动联通电信4G手机 【Apple年货节，限
112万+条评价
二手有售
Apple产品京东自营旗舰店

图 5-8-1 京东商城网站

右键点击第一个手机弹出菜单，选择“检查”可以看到每一个手机的信息是包含在一组<li>的元素中的，再仔细分析这些<li>都包含在一个<div id="J_goodsList">中，而且每个<li>的形式都是<li class="gl-item">，因此分析<li>中的结构就可以找到手机的各种信息，如图 5-8-2 所示。



```
▶<div id="J_selector" class="selector">…</div>
▼<div id="J_main" class="g-main2" source-data-lazy-advertisement-install="1" data-lazy-img-install="1">
  ▼<div class="m-list">
    ▼<div class="ml-wrap">
      ▶<div id="J_filter" class="filter">…</div>
      ▼<div id="J_goodsList" class="goods-list-v2 gl-type-3 J-goods-list">
        ▼<ul class="gl-warp clearfix" data-tpl="3">
          ▼<li class="gl-item" data-sku="100000503295" data-spu="100000349372" data-pid="100000349372"> == $0
            ▼<div class="gl-i-wrap">
              ▼<div class="p-img">
                ▼<a target="_blank" title="【品质特惠】限时优惠200！成交价1499！潮流镜面渐变色，2400万自拍旗舰，骁龙660AIE处理器！小米爆品特惠，选品质，购小米！" href="https://ccc-x.jd.com/dsp/nc?ext=aHR0cHM6Ly9pdGVtLmpkLmNvbS8xMDAwMDA1MD…
                EseSengWL0paP3EunPiVLDjNZ0gXfKwcvQ6cbuYsveUCzuizkO4

… #J_searchWrap #J_container #J_main div div #J_goodsList ul li.gl-item
Styles  Event Listeners  DOM Breakpoints  Properties  Accessibility
Filter                                   :hov .cls +
element.style {
}
```

图 5-8-2 网站 HTML 代码

复制出其中一个<li>的代码如下：

```
<li class="gl-item" data-pid="100000349372" data-sku="100000503295" data-spu="100000349372">
    <div class="gl-i-wrap">
      <div class="p-img">
        <a href="https://ccc-x.jd.com/dsp/nc?ext=aHR0cHM6Ly9pdGVtLmpkLmNvbS8xMDAwMDA1MDMyOTUuaHRtbA&amp;log=GGwsfgzPEa6GsUjtjtmlPqCcsJJ8_FyzP-oiMm40ZPdMyy4BUaQho5X09iHiOTaBXb5lRyCbqngulgWB4gciEzLOs2zl26geClF2loq8CClJghvlPDDTFItXp_NNt4-Irh4DErYAGKCBfUN09WUKYdtWrjpnRnJzb2V06AwJIRtBYFC-670vScpMxUHRUtgYyT9BrMur50VuMqVl0Wi74c3P67g8CKz0EbDQpSCm87-mn824KGDGtIUHTIfqbVGrclUkU_7haMxMz-jxV2maXdj9qSWTjnXT1Beoc0hl3xOiYxu8131zQ_raOiNa17DKxdTdfLioFnhNlnixJG9MC-tR0BnWMBIBl52Q_kZSuD9e7BLIgid8tNR9ANb4eOUhOXG7eku-06_mEkLhqEuxvdXN2Pk-agFclOULzpWulu4yP_VZz64ghL_hujkJ8JN4x_zGH_xTNSomO__PGBEse5engWL0paP3EunPiVLDjNZQgXfKwcvQ6cbuYsveUCzuizkO47q-Ra5-ZuK2iu2R5w&amp;v=404" onclick="searchlog(1,100000503295,0,2,'','adwClk=1')" target="_blank" title="【品质特惠】限时优惠 200！成交价 1499！潮流镜面渐变色，2400 万自拍旗舰，骁龙 660AIE 处理器！小米爆品特惠，选品质，购小米！">
          <img class="" data-img="1" data-lazy-img="done" height="220" source-data-lazy-img="" src="//img10.360buyimg.com/n7/jfs/t1/2617/6/6143/237736/5ba1f42aE71124526/e242e3e39ec95d66.jpg" width="220"/>
        </a>
        <div data-catid="655" data-done="1" data-lease="" data-presale="" data-venid="1000004123">
        </div>
        </div>
        <div class="p-scroll">
         <span class="ps-prev">
          &lt;
         </span>
         <span class="ps-next">
          &gt;
         </span>
         <div class="ps-wrap">
          <ul class="ps-main">
           <li class="ps-item">
            <a class="curr" href="javascript:;" title="梦幻蓝">
             <img class="" data-done="1" data-img="1" data-lazy-img="done" data-presale=""
data-sku="100000503295"                                              data-url="https://ccc-
x.jd.com/dsp/nc?ext=aHR0cHM6Ly9pdGVtLmpkLmNvbS8xMDAwMDA1MDMyOTUuaHRtbA&am
p;log=GGwsfgzPEa6GsUjtjtmIPqCcsJJ8_FyzP-
oiMm40ZPdMyy4BUaQho5X09iHiOTaBXb5lRyCbqnguIgWB4gciEzLOs2zl26geClF2Ioq8CClJghvlPDD
TFItXp_NNt4-Irh4DErYAGKCBfUN09WUKYdtWrjpnRnJzb2V06AwJIRtBYFC-
670vScpMxUHRUtgYyT9BrMur50VuMqVl0Wi74c3P67g8CKz0EbDQpSCm87-
mn824KGDGtIUHTIfqbVGrclUkU_7haMxMz-
jxV2maXdj9qSWTjnXT1Beoc0hl3xOiYxu8131zQ_raOiNa17DKxdTdfLioFnhNlnixJG9MC-
tR0BnWMBIBl52Q_kZSuD9e7BLIgid8tNR9ANb4eOUhOXG7eku-06_mEkLhqEuxvdXN2Pk-
agFcIOULzpWulu4yP_VZz64ghL_hujkJ8JN4x_zGH_xTNSomO__PGBEse5engWL0paP3EunPiVLDjNZ
QgXfKwcvQ6cbuYsveUCzuizkO47q-Ra5-ZuK2iu2R5w&amp;v=404"                    height="25"
src="//img10.360buyimg.com/n9/jfs/t1/2617/6/6143/237736/5ba1f42aE71124526/e242e3e39e
c95d66.jpg" width="25"/>
            </a>
           </li>
          </ul>
         </div>
        </div>
        <div class="p-price">
         <strong class="J_100000503295" data-done="1">
          <em>
           ￥
          </em>
          <i>
           1499.00
          </i>
         </strong>
        </div>
        <div class="p-name p-name-type-2">
         <a                                                              href="https://ccc-
x.jd.com/dsp/nc?ext=aHR0cHM6Ly9pdGVtLmpkLmNvbS8xMDAwMDA1MDMyOTUuaHRtbA&am
```

```
p;log=GGwsfgzPEa6GsUjtjtmIPqCcsJJ8_FyzP-
oiMm40ZPdMyy4BUaQho5X09iHiOTaBXb5lRyCbqnguIgWB4gciEzLOs2zl26geClF2Ioq8CClJghvlPDD
TFItXp_NNt4-Irh4DErYAGKCBfUN09WUKYdtWrjpnRnJzb2V06AwJIRtBYFC-
670vScpMxUHRUtgYyT9BrMur50VuMqVl0Wi74c3P67g8CKz0EbDQpSCm87-
mn824KGDGtIUHTIfqbVGrclUkU_7haMxMz-
jxV2maXdj9qSWTjnXT1Beoc0hl3xOiYxu8131zQ_raOiNa17DKxdTdfLioFnhNlnixJG9MC-
tR0BnWMBIBl52Q_kZSuD9e7BLIgid8tNR9ANb4eOUhOXG7eku-06_mEkLhqEuxvdXN2Pk-
agFcIOULzpWulu4yP_VZz64ghL_hujkJ8JN4x_zGH_xTNSomO__PGBEse5engWL0paP3EunPiVLDjNZ
QgXfKwcvQ6cbuYsveUCzuizkO47q-Ra5-ZuK2iu2R5w&amp;v=404"
onclick="searchlog(1,100000503295,0,1,'','adwClk=1')" target="_blank" title="【品质特惠】限时
优惠 200！成交价 1499！潮流镜面渐变色，2400 万自拍旗舰，骁龙 660AIE 处理器！小米爆
品特惠，选品质，购小米！">
          <em>
           小米 8 青春版  镜面渐变 AI 双摄  6GB+64GB  梦幻蓝  骁龙  全网通 4G  双卡双待
全面屏拍照游戏智能
           <font class="skcolor_ljg">
            手机
           </font>
          </em>
          <i class="promo-words" id="J_AD_100000503295">
           【品质特惠】限时优惠 200！成交价 1499！潮流镜面渐变色，2400 万自拍旗舰，
骁龙 660AIE 处理器！小米爆品特惠，选品质，购小米！
          </i>
         </a>
        </div>
        <div class="p-commit" data-done="1">
         <strong>
          <a                                                                href="https://ccc-
x.jd.com/dsp/nc?ext=aHR0cHM6Ly9pdGVtLmpkLmNvbS8xMDAwMDA1MDMyOTUuaHRtbA&am
p;log=GGwsfgzPEa6GsUjtjtmIPqCcsJJ8_FyzP-
oiMm40ZPdMyy4BUaQho5X09iHiOTaBXb5lRyCbqnguIgWB4gciEzLOs2zl26geClF2Ioq8CClJghvlPDD
TFItXp_NNt4-Irh4DErYAGKCBfUN09WUKYdtWrjpnRnJzb2V06AwJIRtBYFC-
670vScpMxUHRUtgYyT9BrMur50VuMqVl0Wi74c3P67g8CKz0EbDQpSCm87-
mn824KGDGtIUHTIfqbVGrclUkU_7haMxMz-
jxV2maXdj9qSWTjnXT1Beoc0hl3xOiYxu8131zQ_raOiNa17DKxdTdfLioFnhNlnixJG9MC-
tR0BnWMBIBl52Q_kZSuD9e7BLIgid8tNR9ANb4eOUhOXG7eku-06_mEkLhqEuxvdXN2Pk-
agFcIOULzpWulu4yP_VZz64ghL_hujkJ8JN4x_zGH_xTNSomO__PGBEse5engWL0paP3EunPiVLDjNZ
QgXfKwcvQ6cbuYsveUCzuizkO47q-Ra5-ZuK2iu2R5w&amp;v=404"
id="J_comment_100000503295"          onclick="searchlog(1,100000503295,0,3,'','adwClk=1')"
target="_blank">
           19 万+
          </a>
          条评价
         </strong>
        </div>
        <div class="p-focus">
         <a class="J_focus" data-sku="100000503295" href="javascript:;"
onclick="searchlog(1,100000503295,0,5,'','adwClk=1')" title="点击关注">
          <i>
          </i>
          关注
         </a>
        </div>
        <div class="p-shop" data-done="1" data-reputation="99" data-score="0" data-
selfware="1" data-verderid="1000004123">
         <span class="J_im_icon">
          <a href="//mall.jd.com/index-1000004123.html"
onclick="searchlog(1,1000004123,0,58)" target="_blank" title="小米京东自营旗舰店">
           小米京东自营旗舰店
          </a>
          <b class="im-02" onclick="searchlog(1,1000004123,0,61)"
style="background:url(//img14.360buyimg.com/uba/jfs/t26764/156/1205787445/713/9f715eaa
/5bc4255bN0776eea6.png) no-repeat;" title="联系客服">
          </b>
         </span>
        </div>
        <div class="p-icons" data-done="1" id="J_pro_100000503295">
         <i class="goods-icons J-picon-tips J-picon-fix" data-idx="1" data-tips="京东自营，品质
保障">
          自营
         </i>
        </div>
        <span class="p-promo-flag">
         广告
        </span>
        <img src="https://im-x.jd.com/dsp/np?log=GGwsfgzPEa6GsUjtjtmIPqCcsJJ8_FyzP-
oiMm40ZPdMyy4BUaQho5X09iHiOTaBXb5lRyCbqnguIgWB4gciEzLOs2zl26geClF2Ioq8CClJghvlPDD
TFItXp_NNt4-Irh4DErYAGKCBfUN09WUKYdtWrjpnRnJzb2V06AwJIRtBYFC-
670vScpMxUHRUtgYyT9BrMur50VuMqVl0Wi74bCEGdD_q75HGSzPQGw6hJ4HlaZ8W28vNYOfkLp
xilFoRc4qcGHvvWHOiOzImfnEepOqdi4fDVK_-
FB3_3oJQA7sQmNUBqdsLNuuYfNe9a4MlImbykNTzzAKxOD42W8vJO1zfl2NpV7UMZA9eUBGv4rK
FVn4PYB-WCFzT6SzeH1LUR3ZjJuNp38LQ2ZGFpK_JHhuYDLPgucuEtpSdIUJSV7NaX-VyBg6-
eg5oqoIYlFaxfqxYYqR5dGOYIBY_W2CuGdoNvynF09oa4nBudXWOdMUJ4GG-
KQ0U6vmwU4IuCEsywY-rl2-sAgHoxJLklZwvpk2XpJ8VXYcVbHTPFEZzU0~&amp;v=404&amp;rt=3"
style="display:none;"/>
        <img class="err-poster" source-data-lazy-advertisement="done"
src="//misc.360buyimg.com/lib/img/e/blank.gif" style="display: none;"/>
       </div>
```

```
    </li>
```

5.8.2 爬取网页数据

1、爬取手机<li>元素

从 HTML 中可以看到　　每个<li>都包含在<div id="J_goodsList">下面的，而且每个<li>都是<li class="gl-item">的格式，因此使用：

```
lis = self.driver.find_elements_by_xpath("//div[@id='J_goodsList']//li[@class='gl-item']")
for li in lis:
    # 从 li 爬取数据
```

在循环可以得到每个<li>元素，每部手机的数据从 li 对象中进一步爬取。

2、爬取价格

价格存在于<div class='p-price'>下面的一个<i>元素下，因此 price 价格这样爬取：

```
try:
    price = li.find_element_by_xpath(".//div[@class='p-price']//i").text
except:
    price = "0"
```

3、爬取品牌与简介

品牌与简介存在于<div class='p-name p-name-type-2'>下面的<em>元素中，而且品牌时这个<em>下面文字中的第一个部分（用空格分开），因此品牌 mark 与简介 note 如下：

```
try:
    note = li.find_element_by_xpath(".//div[@class='p-name p-name-type-2']//em").text
    mark = note.split(" ")[0]
    mark = mark.replace("爱心东东\n", "")
    mark = mark.replace(",", "")
    note = note.replace("爱心东东\n", "")
    note = note.replace(",", "")
except:
    note = ""
    mark = ""
```

4、爬取图像地址

仔细分析 HTML 代码可以看到手机图像存在于<div class="p-img">下面的一个<a>超级链接的<img>元素中，图像要么存储于<img>的 src 属性，要存储在<img>的 data-lazy-img 属性，因此我们在这个两个属性中去取两个地址 src1 与 src2，程序如下：

```
try:
    src1 = li.find_element_by_xpath(".//div[@class='p-img']//a//img").get_attribute("src")
except:
    src1=""
try:
    src2 = li.find_element_by_xpath(".//div[@class='p-img']//a//img").get_attribute("data-lazy-img")
except:
    src2=""
```

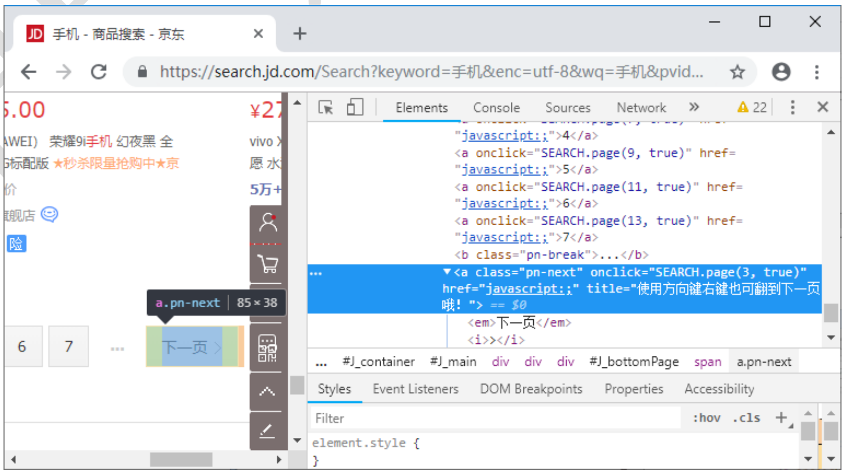
地址 src1 与 src2 中一般总有一个有图像存在，编写 download 下载函数如下：

```
def download(self, src1, src2, mFile):
    data = None
    if src1:
        try:
            req = urllib.request.Request(src1, headers=MySpider.headers)
            resp = urllib.request.urlopen(req, timeout=400)
            data = resp.read()
        except:
            pass
    if not data and src2:
        try:
            req = urllib.request.Request(src2, headers=MySpider.headers)
            resp = urllib.request.urlopen(req, timeout=400)
            data = resp.read()
        except:
            pass
    if data:
        fobj = open(MySpider.imagePath + "\\" + mFile, "wb")
        fobj.write(data)
        fobj.close()
        print("download ", mFile)
```

如果 src1 存在 download 试图先从 src1 下载，如果下载成功就完成，如果 src1 不存在或者下载失败就从 src2 下载，一般情况下总有一个地址可以下载。

5.8.3 实现网页翻页

手机很多，有很多个网页，找到网页的翻页控制，发现翻页不是通过简单的 HTML 代码控制的，而是通过 JavaScript 控制的，如图 5-8-3 所示。



```
"javascript:;">4</a>
<a onclick="SEARCH.page(9, true)" href=
"javascript:;">5</a>
<a onclick="SEARCH.page(11, true)" href=
"javascript:;">6</a>
<a onclick="SEARCH.page(13, true)" href=
"javascript:;">7</a>
<b class="pn-break">...</b>
▼<a class="pn-next" onclick="SEARCH.page(3, true)"
href="javascript:;" title="使用方向键右键也可翻到下一页
哦! "> == $0
  <em>下一页</em>
  <i>></i>
... #J_container #J_main div div div #J_bottomPage span a.pn-next
```

图 5-8-3 网页翻页

由此可见要爬取下一个页面的手机就必须获取控制翻页的超级链接元素<a>，并模仿鼠

标点击那样去点击该链接，这样就转去下一个页面了。

复制出翻页的 HTML 如下：

```
<span class="p-num">
 <a class="pn-prev disabled">
  <i>
   &lt;
  </i>
  <em>
   上一页
  </em>
 </a>
 <a class="curr" href="javascript:;">
  1
 </a>
 <a href="javascript:;" onclick="SEARCH.page(3, true)">
  2
 </a>
 <a href="javascript:;" onclick="SEARCH.page(5, true)">
  3
 </a>
 <a href="javascript:;" onclick="SEARCH.page(7, true)">
  4
 </a>
 <a href="javascript:;" onclick="SEARCH.page(9, true)">
  5
 </a>
 <a href="javascript:;" onclick="SEARCH.page(11, true)">
  6
 </a>
 <a href="javascript:;" onclick="SEARCH.page(13, true)">
  7
 </a>
 <b class="pn-break">
  ...
 </b>
 <a class="pn-next" href="javascript:;" onclick="SEARCH.page(3, true)" title="使用方向键右键也可翻到下一页哦！">
  <em>
   下一页
  </em>
  <i>
   &gt;
  </i>
 </a>
```

```
    </span>
```

从中看只要找到<span class="p-num">，然后找到"下一页"的超级链接，在正常能翻页时超级链接是<a class='pn-next'>，到最后一页不能翻页时变成<a class='pn-next-disabled'>，因此编写下列程序找到 nextPage 就能实现翻页：

```
try:
    self.driver.find_element_by_xpath("//span[@class='p-num']//a[@class='pn-next disabled']")
except:
    nextPage = self.driver.find_element_by_xpath("//span[@class='p-num']//a[@class='pn-next']")
    nextPage.click()
```

如果没有找到<a class='pn-next-disabled'>元素就找<a class='pn-next'>元素，然后使用 nextPage.click()实现翻页。

5.8.4 编写爬虫程序

根据前面的分析，编写爬虫程序如下：

```
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
import urllib.request
import threading
import sqlite3
import os
import datetime
from selenium.webdriver.common.keys import Keys
import time

class MySpider:
    headers = {
        "User-Agent": "Mozilla/5.0 (Windows; U; Windows NT 6.0 x64; en-US; rv:1.9pre) Gecko/2008072421 Minefield/3.0.2pre"}
    imagePath = "download"

    def startUp(self,url,key):
        # Initializing Chrome browser
        chrome_options = Options()
        chrome_options.add_argument('--headless')
        chrome_options.add_argument('--disable-gpu')
        self.driver = webdriver.Chrome(chrome_options=chrome_options)

        # Initializing variables
        self.threads = []
        self.No = 0
        self.imgNo=0
        # Initializing database
        try:
            self.con = sqlite3.connect("phones.db")
            self.cursor = self.con.cursor()
            try:
                # 如果有表就删除
                self.cursor.execute("drop table phones")
            except:
                pass
            try:
                # 建立新的表
                sql = "create table phones (mNo varchar(32) primary key,mMark varchar(256),mPrice varchar(32),mNote varchar(1024),mFile varchar(256))"
                self.cursor.execute(sql)
            except:
                pass
        except Exception as err:
            print(err)

        # Initializing images folder
        try:
            if not os.path.exists(MySpider.imagePath):
                os.mkdir(MySpider.imagePath)
            images = os.listdir(MySpider.imagePath)
            for img in images:
                s = os.path.join(MySpider.imagePath, img)
                os.remove(s)
        except Exception as err:
            print(err)

        self.driver.get(url)
        keyInput=self.driver.find_element_by_id("key")
        keyInput.send_keys(key)
        keyInput.send_keys(Keys.ENTER)

    def closeUp(self):
        try:
            self.con.commit()
            self.con.close()
            self.driver.close()
        except Exception as err:
            print(err);

    def insertDB(self, mNo, mMark, mPrice, mNote, mFile):
        try:
            sql = "insert into phones (mNo,mMark,mPrice,mNote,mFile) values (?,?,?,?,?)"
            self.cursor.execute(sql, (mNo, mMark, mPrice, mNote, mFile))
        except Exception as err:
            print(err)

    def showDB(self):
        try:
            con=sqlite3.connect("phones.db")
            cursor=con.cursor()
            print("%-8s %-16s %-8s %-16s %s" % ("No", "Mark", "Price", "Image", "Note"))
            cursor.execute("select mNo,mMark,mPrice,mFile,mNote from phones order by mNo")
            rows = cursor.fetchall()
            for row in rows:
                print("%-8s %-16s %-8s %-16s %s" % (row[0], row[1], row[2], row[3], row[4]))
            con.close()
        except Exception as err:
            print(err)

    def download(self, src1,src2,mFile):
        data=None
        if src1:
            try:
                req = urllib.request.Request(src1, headers=MySpider.headers)
                resp = urllib.request.urlopen(req, timeout=400)
                data = resp.read()
            except:
                pass
        if not data and src2:
            try:
                req = urllib.request.Request(src2, headers=MySpider.headers)
                resp = urllib.request.urlopen(req, timeout=400)
                data = resp.read()
            except:
                pass
        if data:
            fobj = open(MySpider.imagePath + "\\" + mFile, "wb")
            fobj.write(data)
            fobj.close()
            print("download ",mFile)

    def processSpider(self):
            try:
                time.sleep(1)
                print(self.driver.current_url)
                lis                                                                                  =
self.driver.find_elements_by_xpath("//div[@id='J_goodsList']//li[@class='gl-item']")
                for li in lis:
                    # We find that the image is either in src or in data-lazy-img attribute
                    try:
                        src1        =        li.find_element_by_xpath(".//div[@class='p-
img']//a//img").get_attribute("src")
                    except:
                        src1=""
                    try:
                        src2        =        li.find_element_by_xpath(".//div[@class='p-
img']//a//img").get_attribute("data-lazy-img")
                    except:
                        src2=""
                    try:
                        price = li.find_element_by_xpath(".//div[@class='p-price']//i").text
                    except:
                        price="0"
                    try:
                        note  =  li.find_element_by_xpath(".//div[@class='p-name  p-name-
type-2']//em").text
                        mark = note.split(" ")[0]
                        mark = mark.replace("爱心东东\n", "")
                        mark = mark.replace(",", "")
                        note = note.replace("爱心东东\n", "")
                        note = note.replace(",", "")
                    except:
                        note=""
                        mark=""

                    self.No = self.No + 1
                    no = str(self.No)
                    while len(no) < 6:
                        no = "0" + no
                    print(no,mark,price)
                    if src1:
                        src1=urllib.request.urljoin(self.driver.current_url,src1)
                        p = src1.rfind(".")
                        mFile = no + src1[p:]
                    elif src2:
                        src2=urllib.request.urljoin(self.driver.current_url,src2)
                        p = src2.rfind(".")
                        mFile = no + src2[p:]
                    if src1 or src2:
                        T = threading.Thread(target=self.download, args=(src1,src2,mFile))
                        T.setDaemon(False)
                        T.start()
                        self.threads.append(T)
                    else:
                        mFile = ""
                    self.insertDB(no, mark, price, note, mFile)
                try:
                    self.driver.find_element_by_xpath("//span[@class='p-
    num']//a[@class='pn-next disabled']")
                except:
                    nextPage    =    self.driver.find_element_by_xpath("//span[@class='p-
    num']//a[@class='pn-next']")
                    nextPage.click()
                    self.processSpider()
            except Exception as err:
                print(err)

        def executeSpider(self, url,key):
            starttime = datetime.datetime.now()
            print("Spider starting......")
            self.startUp(url,key)
            self.processSpider()
            self.closeUp()
            for t in self.threads:
                t.join()
            print("Spider completed......")
            endtime = datetime.datetime.now()
            elapsed = (endtime - starttime).seconds
            print("Total ", elapsed, " seconds elapsed")

    url = "http://www.jd.com"
    spider = MySpider()
    while True:
        print("1.爬取")
        print("2.显示")
        print("3.退出")
        s=input("请选择(1,2,3):")
        if s=="1":
            spider.executeSpider(url,"手机")
        elif s=="2":
            spider.showDB()
        elif s=="3":
            break
```

startUp 函数中初始化数据库 phones.db 并建立一张空的 phones 表以备存储数据，同时创建 download 文件夹并清空文件夹的文件，准备存储下载的图像。函数中还查找到网页的输入框<input id="key">并模拟键盘输入要爬取的商品关键字 key，并模拟键盘回车转去商品的网页，关键程序如下：

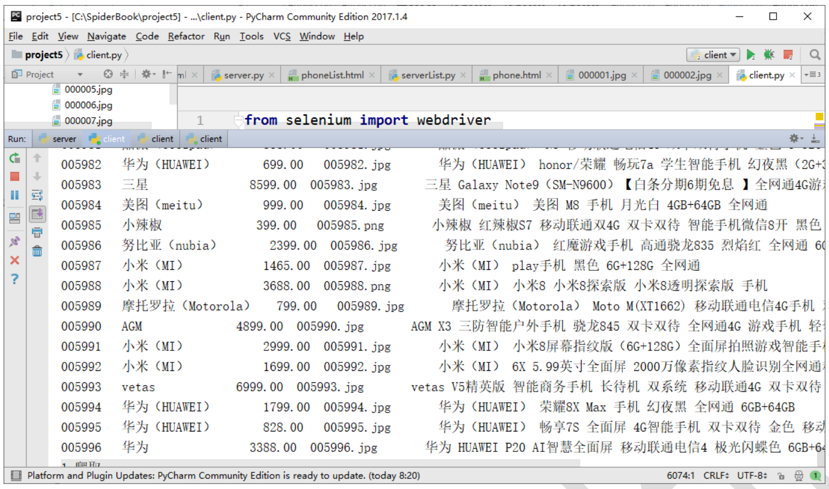
```
keyInput=self.driver.find_element_by_id("key")
keyInput.send_keys(key)
keyInput.send_keys(Keys.ENTER)
```

5.8.5 执行爬虫程序

执行这个爬虫程序，耗时 921 秒，结果总共爬取到近 100 个页面的近 6000 部手机的数据与图像，下面是部分结果：

```
Spider starting......
https://search.jd.com/Search?keyword=%E6%89%8B%E6%9C%BA&enc=utf-8&pvid=13986d0072404844b2dae1417d9191b0
000001 Apple 6299.00
000002 OPPO 1699.00
000003  一加手机 6T 3599.00
000004  荣耀 10 青春版  1399.00
000005  荣耀 10 2199.00
000006  小米 8 2299.00
000007  荣耀 8X 1399.00
000008 Apple 3998.00
......
download    005993.jpg
005994  华为（HUAWEI）  1799.00
download    005994.jpg
005995  华为（HUAWEI）  828.00
download    005995.jpg
005996  华为  3388.00
No          Mark                Price      Image                Note
download    005996.jpg
Spider completed......
Total    921    seconds elapsed
```

如图 5-8-4 是程序查看查看 phone.db 数据库的结果，图 5-8-5 所示是爬取并存储到 download 文件夹中的图像，也有近 6000 个。



```
from selenium import webdriver
```

```
005982   华为（HUAWEI）       699.00   005982.jpg     华为（HUAWEI） honor/荣耀 畅玩7a 学生智能手机 幻夜黑（2G+
005983   三星                8599.00   005983.jpg     三星 Galaxy Note9（SM-N9600）【白条分期6期免息 】全网通4G游
005984   美图（meitu）        999.00   005984.jpg     美图（meitu） 美图 M8 手机 月光白 4GB+64GB 全网通
005985   小辣椒               399.00   005985.png     小辣椒 红辣椒S7 移动联通双4G 双卡双待 智能手机微信8开 黑色
005986   努比亚（nubia）      2399.00   005986.jpg     努比亚（nubia） 红魔游戏手机 高通骁龙835 烈焰红 全网通 6
005987   小米（MI）           1465.00   005987.jpg     小米（MI） play手机 黑色 6G+128G 全网通
005988   小米（MI）           3688.00   005988.png     小米（MI） 小米8 小米8探索版 小米8透明探索版 手机
005989   摩托罗拉（Motorola）   799.00   005989.jpg     摩托罗拉（Motorola） Moto M(XT1662) 移动联通电信4G手机
005990   AGM                4899.00   005990.jpg     AGM X3 三防智能户外手机 骁龙845 双卡双待 全网通4G 游戏手机 轻
005991   小米（MI）           2999.00   005991.jpg     小米（MI） 小米8屏幕指纹版（6G+128G）全面屏拍照游戏智能手
005992   小米（MI）           1699.00   005992.jpg     小米（MI） 6X 5.99英寸全面屏 2000万像素指纹人脸识别全网通
005993   vetas              6999.00   005993.jpg     vetas V5精英版 智能商务手机 长待机 双系统 移动联通4G 双卡双待
005994   华为（HUAWEI）      1799.00   005994.jpg     华为（HUAWEI） 荣耀8X Max 手机 幻夜黑 全网通 6GB+64GB
005995   华为（HUAWEI）       828.00   005995.jpg     华为（HUAWEI） 畅享7S 全面屏 4G智能手机 双卡双待 金色 移动
005996   华为                3388.00   005996.jpg     华为 HUAWEI P20 AI智慧全面屏 移动联通电信4 极光闪蝶色 6GB+6
```

project5 - [C:\SpiderBook\project5] - ...\client.py - PyCharm Community Edition 2017.1.4
File Edit View Navigate Code Refactor Run Tools VCS Window Help
project5 client.py
client
Project
server.py
phoneList.html
serverList.py
phone.html
000001.jpg
000002.jpg
client.py
000005.jpg
000006.jpg
000007.jpg
Run: server client client client
Platform and Plugin Updates: PyCharm Community Edition is ready to update. (today 8:20)
6074:1 CRLF÷ UTF-8÷

图 5-8-4 爬取到的数据

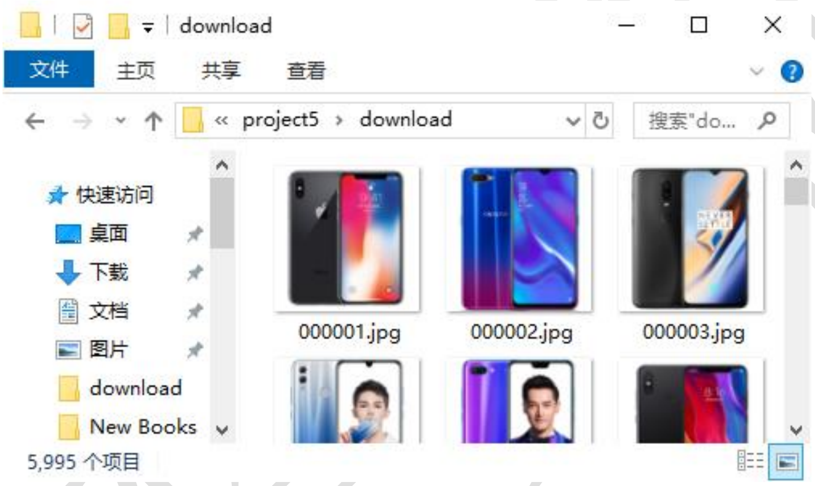


图 5-8-5 爬取到的图像