使用Scrapy + Selenium框架抓取京东商品信息

目录

[使用Scrapy + Selenium框架抓取京东商品信息 1](#_Toc420682117)

[一、介绍 1](#_Toc420682118)

[二、Scrapy提取商品详情链接 1](#_Toc420682119)

[三、使用selenium 的webdriver 下载商品详情页面 1](#_Toc420682120)

[四、总结 2](#_Toc420682121)

### 一、介绍

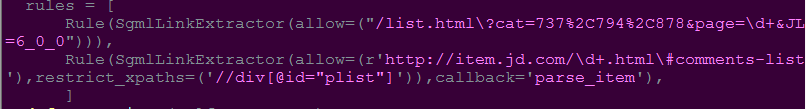
Scrapy是一个基于Twisted，纯Python实现的爬虫框架，用户只需要定制开发几个模块就可以轻松的实现一个爬虫，用来抓取网页内容、图片、视频等，非常方便。

Selenium本是用来对web application做自动化测试的。不过，它有个天大的好处：能让我们用python(当然不仅是python)代码模拟人对浏览器的操作。

由于scrapy的middleware的配置比较复杂，这里采用一个简单的方法：改用最简单的spider,在parse函数中用selenium下载详情页面，我们知道京东商品页面我们感兴趣的信息大部分是由javascript代码生成，因此可以先在浏览器中执行JavaScript代码，再从最终的页面上抓取信息（当然也有其他的方法）。

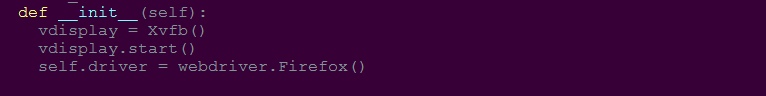
### 二、Scrapy提取商品详情链接

这里使用冰箱这个类别做为例子介绍，冰箱列表页面<http://list.jd.com/list.html?cat=737,794,878>，使用scrapy中的SmglLinkExtractor提取各个冰箱商品信息的链接，并指定回调函数。

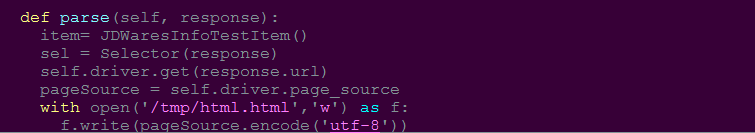


### 三、使用selenium 的webdriver 下载商品详情页面

1) .初始化Xvfb，创建Firefoxwebdriver实例

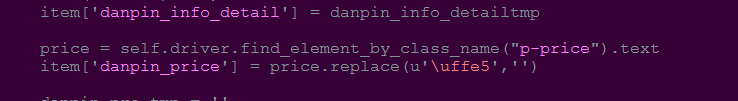


2).使用driver.get()尝试打开提取的商品详情页面，webdriver会等待网页元素加载完成之后才把控制权交回脚本，查看得到的源文件是不是执行过js的“源文件”



通过测试可以发现html.html中已经可以查看商品的价格、促销、商品颜色规格等信息(最初的源文件是没有的)，

3).使用 webdriver提供了很多如find\_element\_by\_\*的方法来匹配要查找的信息元素，例如价格信息



### 四、总结

在scrapy中使用selenium模拟浏览器下载页面需要新建虚拟窗口并在后台打开浏览器，因此在抓取页面相比默认下载器速度要慢。