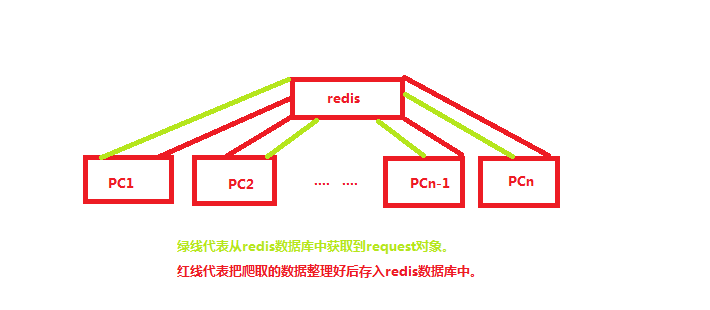
# Dangdang\_book\_log

本次爬虫项目所爬取的数据是当当网中的图书分类中的图书信息，图书这个分类算是比较复杂的分类，它有三个细分（包括大分类、中分类、小分类），因此爬取其他分类的数据也是比较容易的。

## 爬取数据所使用的框架

本次爬虫项目所使用的框架是scrapy\_redis，实现分布式爬虫，如何实现的呢？如图所示：



它的原理是这样的，每台电脑都有一份爬取某个网站数据的代码，并且都要连接上某台特定的redis服务器。当每台电脑开始运行代码的时候，就在等待request对象。

1. 每台电脑开始运行的时候，等在这start\_url，也就是起始网页，这个start\_url是在redis中的，预先把start\_url通过键值对的形式，存入redis中
2. 当每台电脑全部运行代码时，都会抢夺这个start\_url，谁先抢到了，谁开始访问这个网址，获取其他的url,然后框架会把url转换成request对象存入redis中，让其他的电脑进行访问获取数据。以此类推
3. 如果当前的电脑数还不够，那么可以增加电脑数，只要新增的电脑连接上特定的redis服务器中就可以了，根据scrapy\_redis框架的特性，不会重复访问网站，因此不会获取到重复的数据。

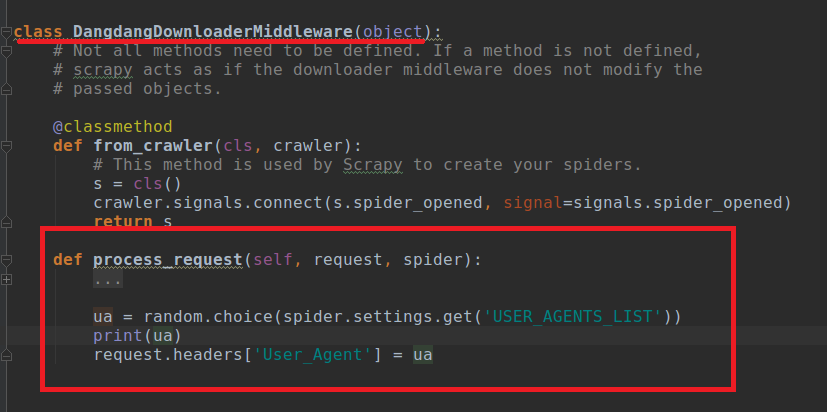
这样就实现了分布式爬虫，提高效率。

## 反爬处理

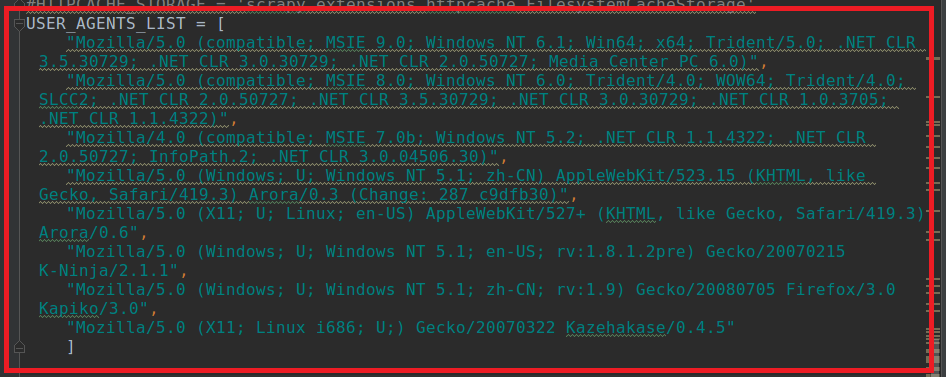
相信这么多台电脑访问同一个网站，必定会进行反爬，因此，需要做到反反爬处理工作。

这里就需要用到下载中间件了。

创建一个middlewares.py这个文件，然后随便定义一个类，这个类中定义一个方法，process\_request，这个方法中实现动态的浏览器代理（User-Agent）,代码如下：



可以看到通过spider.settings.get(‘USER\_AGENTS\_LIST’)这个方法获取到settings.py文件中的预先设置好的代理，如图所示：



这些代理自己可以去寻找的，这样就实现了动态的User-Agent。可以达到反爬的效果。

当然可以在这个middlewares.py文件中在定义一个类，实现动态的Ip代理，这些Ip可以在相应的免费网站上获得，或者自掏腰包花钱买。（我没买，哈哈，家里没矿）实现的相关教程参照各大技术门户网站吧，实现起来跟实现动态User-Agent差不多。

## Xpath的构思

要获取页面上的指定数据必须要通过xpath来获取，所以需要分析页面的源代码来编写相应的xpath。

当当图书有三个类别（大分类、中分类、小分类）如图所示：



因此需要看看源代码如何获取这些数据，以及在获取的时候判断数据是否在响应中，（这是十分必要的警惕）。

注意点：

1. 有些时候，我们编写xpath是按照Element中的代码进行编写的，但是一定要注意，必须要与网页的源代码进行对比，有些时候，Element中a标签的上一级有一个span标签，但是网页源代码中的a标签的上一级没有span标签，这样的话，如果按照Element来编写xpath会获取不到数据的，值为None
2. 除了上面的注意之外，还要判断网页中的某些数据的来源（产品的价格，产品的厂家等等），是否是在当前url地址的响应中，如果不是，就要分析了，是不是通过ajax动态生成，或者其它的方式。

以上的注意点是十分必要的，这个是爬取数据的核心吧。

## 本次爬虫的难点

爬取当当网的难点或是重点：

1. 实现分布式爬虫
2. 反爬的处理机制
3. 数据的获取

通过爬取当当网，觉得这个网站反爬的措施是有的，动不动就ban，因此爬取的时候要考虑动态的ip代理和User-Agent。

## 数据爬取过后，需要清洗

也就是通过正则来处理某些我们不想要的数据，比方说空格、&nbsp；等等。如下所示是清洗后的数据的格式：

{**'book\_name'**: ' 胡适：四十自述（图文典藏版）', **'href':** 'http://category.dangdang.com/cp01.38.25.00.00.00.html', **'desc'**: ['《胡适四十自述》秉承了胡适先生恳坦白的态度为我们展现了一个有血有肉、“鲜活”的胡适。\*令人震撼的是，他甚至把自己少年时代打麻将、喝酒、逛窑子、醉倒在马路上与巡捕打架等种种恶习，均毫无讳', **'press':** '华文出版社', 'price': '¥18.50', **'s\_cate':** '自传', 'publish\_date': '2013-01-01', **'img':** 'http://img3m5.ddimg.cn/53/22/23188625-1\_b\_1.jpg', ': '文艺', **'author':** ['胡适']}

## 存储数据

清洗后的数据可以存在本地的某个文件或者其他的数据库中，个人需求吧，本次的数据是默认的存放在redis数据库中的。