爬虫修改意见：

1.所爬取的网页有些是有重复的，即有的url是不同的，但是其所对应的页面是同一个网页，如对于所爬取到的《大数据时代》，，共有如下四个网址：

http://product.dangdang.com/22928335.html#ddclick\_reco\_product\_alsobuy

http://product.dangdang.com/22928335.html?ref=book-65152-9162\_1-473628-2

http://product.dangdang.com/22928335.html

http://product.dangdang.com/22928335.html?ddclick\_reco\_product\_alsobuy#comment

查询集合大小为：4

但可以看到其中有些网址是由于对页面<http://product.dangdang.com/22928335.htm>做出了不同的请求而出现的url，其在浏览器中及所爬出的数据都是一样的，所以导致了数据库的冗余。

解决方法：多找出类似的问题，寻找共性，对于上面的例子URL1和3其实是同一个页面，只不过它们是不同的页面段而已，可以处理为同一个页面。

2.每次存储至第4000条数据时会出现代码停止的情况，所报出的错误为nullpointerexception，且无错误代码行数提醒，此处问题一直未解决，正尝试将爬取网页深度修改为无限制，看是否能解决该问题。

3.在爬取网页时会经常出现如下所示的一些类型的错误：

org.htmlparser.util.ParserException: Error in opening a connection to showMsgBox('lcase22905590','22905590','http://customer.dangdang.com/wishlist/remote\_addtofavorlist.aspx');;

org.htmlparser.util.ParserException: D:\wpc\myeclipseWorkpace\PriceComparsion\_01\AddToShoppingCart(22868441) (系统找不到指定的文件。);

org.htmlparser.util.ParserException: D:\wpc\myeclipseWorkpace\PriceComparsion\_01\goto\_product(23424363) (系统找不到指定的文件。);

java.io.FileNotFoundException: D:\wpc\myeclipseWorkpace\PriceComparsion\_01\goto\_product(23424363) (系统找不到指定的文件。)

以上错误不影响程序执行

解决方法：默默的修改程序中的bug，换个人员进行调试。