爬虫组测试文档

## 编写目的

对爬虫组的爬虫代码进行测试，查看爬虫的数据是否满足需求，对于爬取的不同的网站存在的特殊情况进行反馈，为软件设计人员提供更正系统问题的依据。

## 任务概述

* 1. 目标

对爬虫的代码进行详细的测试，爬取的数据存在txt文件中，针对txt文件中存在的特殊情况进行排查，排除基本的bug，并将情况反馈给代码开发人员。

* 1. 运行环境

操作系统：Windows 7

数据库：MySQL

客户端：CPU在1GHz以上，内存2G以上。

服务端：CPU在4GHz以上，内存6G以上。

* 1. 测试项目

测试1：进行黑盒测试，就是输入爬取的起始地址，输出爬取的书的信息到txt文件中，对于输出的信息进行排查。

## 测试分析

* 1. 存在的问题：

问题一：需爬取处当当网上的分类信息，方便数据库建设，以及后期图书分类的处理。

问题二：如下的书的属性列缺失，需在代码中添加相应的爬取代码：

1. 包装是精装还是平装
2. 版次
3. 评分星级
4. 抢购价；抢购时间限制
5. 商品评论数（好评数、差评数）

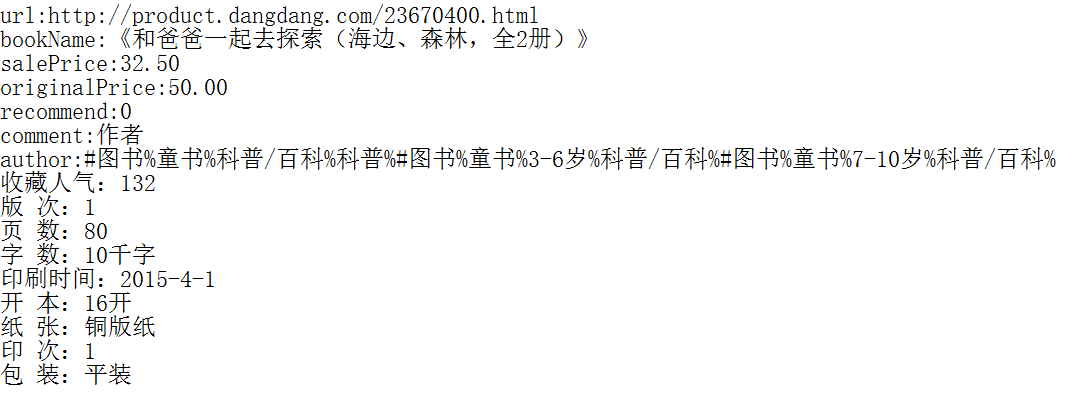
问题三：书名内容太杂，需清理

问题四：爬取处的属性列需要在其值之前标注出属性，如：将《平凡的世界》修改为“书名：《平凡的世界》”或“bookName：《平凡的世界》”

问题五：爬虫存入txt的正确的数据如下图所示：



图一

但部分数据格式仍存在如下问题，如下图所示：

图二

其中的图二中的classify属性缺失，导致图书的分类信息被存入的author里面。

问题六：当当网的数据爬取速度比较快，但是京东的数据爬取速度比较慢。

## 总结

基本上的功能已经实现，但是对于一些细节问题还需完善，对于缺失属性列还需添加爬虫代码，爬虫代码还可继续优化。