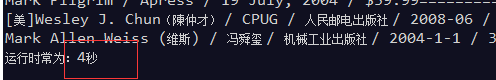
**爬虫基本流程及实现：**

1. 首先创建BookInfo实体类用于保存书籍的相关信息，包含字段有：

序号，书名，评分，评论数，作者，出版社，出版日期，价格

1. 创建定长线程池，处理前5页的内容。因按评价排序，每页有20条数据，5页则筛选出评价最高的前100条数据。这样便于控制线程最大并发数，超出的线程会再队列中等待。
2. 采集数据。使用Jsoup实现网页抓取，因它提供了较多功能可供我们从URL中加载HTML，并生成Document对象实例，方便查找并提取和书籍相关的信息，如作者、评论数、评分、出版社和出版日期等数据，所以非常高效，且实用性强。
3. 将采集来的5页数据内容保存到Excel中，通过POI中的HSSF实现对Excel格式表的读写功能。基本过程为：创建Excel 2007工作簿对象，创建工作表，再根据数据量生成相应的行，并把每个数据的相应字段写入该行对应的列。（评论数大于1000才进行写操作）
4. 抓取前5页满足要求的数据总共运行时间为：4秒。



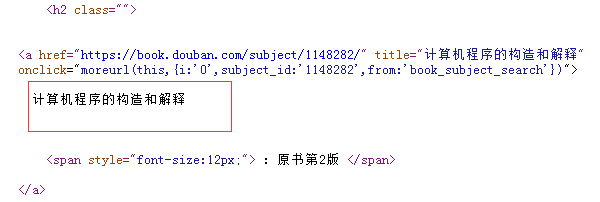
**相关源码及页面如下：**

1. 豆瓣读书有关“编程”方面的书籍按评价排序网页链接为：

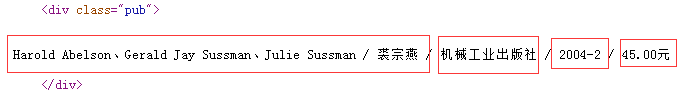




1. 书籍名称对应源码为：



1. 作者及出版社、出版时间及价格源码，提取出保存在String数组中进行处理



1. 评分



1. 评论数

