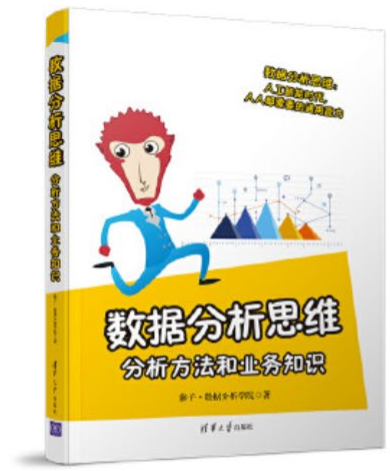
Homework 4

- 图书搜索引擎

  目标：

  - 1. 写一个Web爬虫，爬取图书网站（当当、京东等）的网页；
  - 2. 解析网页内容，对内容进行结构化，并存储到文件中；
    - 包括标题、作者、分类、出版社、图书照片、编辑推荐、内容简介、作者简介、目录、价格等信息。
  - 3. 为内容建立索引；
  - 4. 通过命令行进行内容检索，并展示内容列表
  - 可通过标题、作者、分类、出版社等来检索



当当自营 **数据分析思维：分析方法和业务知识**

数据分析思维：AI时代的通用能力，升职加薪的高维武器！ 预售商品

作者:猴子·数据分析学院　出版社:清华大学出版社　出版时间:2020年11月

在当当计算机/网络新书榜排名3位　★★★★★ 2条评论

预售价　降价通知

¥68.30　(6.9折)

定价¥99.00

促　销　电子书加价购　+20.6元换购《产品经理进阶之路：从小白到专家的》作者1：杨俊

加价购　购买本商品可加价换购以下任意一件商品　收起

【掌门1对1】 小学4-6年级数学1对1在线辅导试听课

+¥1

【掌门1对1】小学4-6年级英语专项口语课

+¥1

商品详情　商品评论(2)　商品问答(0)

开 本：16开　　纸 张：胶版纸　　包 装：平装-胶订

是否套装：否　　国际标准书号ISBN：9787302563839

所属分类：　图书>计算机/网络>数据库>数据库理论

http://product.dangdang.com/29149764.html

编辑推荐

面对工作，你是否经常遇到以下问题：有一堆繁杂数据，怎么利用？软件很熟练，怎么跟需求结合？结论太简单，领导不满意，怎么深入？分析完数据，如何得结论、提建议？找到新工作，如何快速掌握该行业的知识？阅读本书，给你答案！

内容简介

《数据分析思维：分析方法和业务知识》分为两大部分："方法篇"和"实战篇"。"方法篇"介绍了数据分析中常用的业务指标、分析方法以及如何用数据分析解决问题的步骤。"实战篇"讲解了如何应用*篇的方法来解决工作中的问题，分享十二个行业（国内电商、跨境电商、金融信贷、金融第三方支付、家政、旅游、在线教育、运营商、内容、房产、汽车、零售）的业务知识，以及该行业内用数据分析解决问题的实例。每个行业都包括业务模式、业务指标、案例分析三方面的内容。 通过本书的学习，你会熟悉数据分析的方法，并将其灵活应用在自己所处的行业中。

作者简介

本书由猴子·数据分析学院的成员共同编写。猴子，中国科学院大学硕士，"猴子·数据分析学院"创始人，公众号"猴子数据分析"创始人，前IBM工程师。其"分析方法"课程入围知乎年度口碑榜TOP 10，首创的"闯关游戏学习数据分析模式"深受用户喜欢。

目　　录

Homework 4

- 关键技术：
  - 爬虫
  - 信息抽取
  - 索引建立
  - 查询

Homework 4

- Tips：
- 1. 如何在Eclipse中引入jar包

Homework 4

- Tips
- 2. JAVA爬虫
  - crawler4j
    - https://github.com/yasserg/crawler4j

crawler4j

build passing | maven-central v4.4.0 | chat online

crawler4j is an open source web crawler for Java which provides a simple interface for crawling the Web. Using it, you can setup a multi-threaded web crawler in few minutes.

  - JSOUP
    - https://blog.csdn.net/zbx931197485/article/details/78582407

    - jsoup 是一款 Java 的HTML 解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，CSS以及类似于jQuery的操作方法来取出和操作数据，可以看作是java版的jQuery。
      jsoup的主要功能如下：
      从一个URL，文件或字符串中解析HTML；
      使用DOM或CSS选择器来查找、取出数据；
      可操作HTML元素、属性、文本；
      jsoup是基于MIT协议发布的，可放心使用于商业项目。官方网站：http://jsoup.org/

Homework 4

- 基于jsoup: Java HTML Parser来抽取信息 (如标题等，相同的网站同一个模板)，利用正则表达式来建立模板
  - https://jsoup.org/

```
File input = new File("/tmp/input.html");
Document doc = Jsoup.parse(input, "UTF-8", "http://example.com/");

Elements links = doc.select("a[href]"); // a with href
Elements pngs = doc.select("img[src$=.png]");
  // img with src ending .png

Element masthead = doc.select("div.masthead").first();
  // div with class=masthead

Elements resultLinks = doc.select("h3.r > a"); // direct a after h3
```

Homework 4

- Tips
- 3. 利用Lucene对文本进行索引，并进行检索

Apache Lucene™ is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

http://lucene.apache.org/core/

Homework 4

- 3. 利用Lucene对文本进行索引，并进行检索（输入检索词，查询得到相关的问题（或课程）列表，并显示详细信息。
  - 建索引和检索的简例

Homework 4

- 作业包括：java文件 + 文档 + 数据
- 作业打包上传到ftp homework/homework4下
- 文件：学号_姓名_homework4.rar

Homework 4

- 代码要求：
  - 遵守编程规范，如命名、注释等规范
  - 遵守面向对象的设计原则
  - 考虑异常处理等应用

Homework 4

- 文档要求：
  - 按附件格式样例，至少包括：引用、总体设计、详细设计、测试与运行、总结
  - 包括：数据格式说明
  - 附加：程序中包含的其他特色或改进
  - 附加：数据的丰富程度