Packt>

学习MLlib、DL4j和Weka等开源库，
掌握实用的Java数据科学技能

Java数据科学指南

Java Data Science
Cookbook

[加]鲁什迪·夏姆斯（Rushdi Shams）著
武传海 译

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

Java数据科学指南

Java Data Science
Cookbook

[加]鲁什迪·夏姆斯（Rushdi Shams）著
武传海 译

人民邮电出版社
北京

图书在版编目（CIP）数据

Java数据科学指南 /（加）鲁什迪·夏姆斯（Rushdi Shams）著；武传海译. -- 北京：人民邮电出版社，2018.6
ISBN 978-7-115-48163-4

Ⅰ. ①J… Ⅱ. ①鲁… ②武… Ⅲ. ①JAVA语言－程序设计－指南 Ⅳ. ①TP312.8-62

中国版本图书馆CIP数据核字(2018)第057957号

◆ 著 [加] 鲁什迪·夏姆斯（Rushdi Shams）
译 武传海
责任编辑 胡俊英
责任印制 沈 蓉 焦志炜

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
固安县铭成印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：20
字数：396 千字 2018 年 6 月第 1 版
印数：1－2 400 册 2018 年 6 月河北第 1 次印刷
著作权合同登记号 图字：01-2017-4028 号

定价：79.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

内容提要

现如今，数据科学已经成为一个热门的技术领域，它涵盖了人工智能的各个方面，例如数据处理、信息检索、机器学习、自然语言处理、数据可视化等。而 Java 作为一门经典的编程语言，在数据科学领域也有着卓越的表现。

本书旨在通过 Java 编程来引导读者更好地完成数据科学任务。本书通过 9 章内容，详细地介绍了数据获取与清洗、索引的建立和检索数据、统计分析、数据学习、信息的提取、大数据处理、深度学习、数据可视化等重要主题。

本书适合想通过 Java 解决数据科学问题的读者，也适合数据科学领域的专业人士以及普通 Java 开发者阅读。

作者简介

Rushdi Shams 毕业于加拿大韦仕敦大学，获得了机器学习应用博士学位，主攻方向是自然语言处理（Natural Language Processing，NLP）。在成为机器学习与 NLP 领域的专家之前，他讲授本科生与研究生课程。在 YouTube 上，他一直运营着一个名为“跟 Rushdi 一起学”（Learn with Rushdi）的频道，并且做得有声有色，该频道主要面向想学习计算机技术的朋友。

感谢我的家人、朋友与同事，谢谢你们不断地给予我支持、鼓励，以及中肯的批评与建议。

此外，还要感谢 Packt 公司的 Ajith 与 Cheryl，谢谢他们自发地与我保持持续的合作关系。

审稿人简介

Prashant Verma 从 2011 年就开始了他的 IT 生涯，当时他是爱立信公司的一名 Java 开发人员，面向的是电信领域。在有了几年的 Java EE 开发经验之后，他转战大数据领域，几乎用过所有流行的大数据技术，比如 Hadoop、Spark、Kafka、Flume、Mongo、Cassandra 等，而且还熟悉 Scala 与 Python 编程语言。目前，他供职于 QA Infotech 公司，是一名首席数据工程师，致力于使用数据分析与机器学习解决 E-Learning 领域中的问题。

此外，Prashant 也是 Packt 出版的 *Apache Spark for Java Developers* 一书的技术审稿人。

首先感谢 Packt Publishing 给我审阅本书的机会，还要感谢我的雇主、家人，谢谢他们在我忙于审读本书时所表现出的耐心。

谨以此书献给我漂亮的妻子 Mah-Zereen 与可爱的女儿 Ruayda!

前言

当今，数据科学是一个专业化的热门领域，它涵盖人工智能的各个方面，比如数据处理、信息检索、机器学习、自然语言处理、大数据、深度神经网络、数据可视化。在本书中，我们将讲解数据科学领域中既流行又智能的技术，这些技术分散在全书各个章节中，涉及 70 多个问题。

请记住，目前各个领域对高级数据科学家的需求是非常旺盛的，我们主要使用 Java 编写了本书各个章节，包括那些使用 Java 编写的著名的、经典的、最新的数据科学库。首先我们介绍数据采集与清洗流程，而后了解一下如何对所获取的数据建立索引以及进行检索。随后，我们讲解数据的统计描述、统计推断及其应用。接着，安排两个连续的章节讲解面向数据的机器学习应用，这些内容是创建智能系统的基础。除此之外，所讲解的内容还包括现代信息检索与自然语言处理技术。大数据是一个新兴的热门领域，本书内容会涉及其中几个方面。并且，我们还会讲解使用深度神经网络进行深度学习的基础知识。最后，我们学习如何使用有意义的视觉方式或图形表示数据以及从数据中获取的信息。

本书面向的读者是那些对数据科学感兴趣，或者打算应用 Java 数据科学技术来进一步理解底层数据的朋友。

本书内容安排

第 1 章 “获取数据与清洗数据”，介绍各种读写数据的方法，以及对数据进行清洗去除其中噪声的方法。本章所涉及的数据文件类型广为人知，比如 PDF、ASCII、CSV、TSV、XML、JSON。此外，本章还介绍用来提取 Web 数据的方法。

第 2 章 “为数据建立索引与搜索数据”，讲解如何使用 Apache Lucene 为数据建立索

引以实现快速检索。本章介绍的技术是现代搜索技术的基础。

第 3 章 “数据统计分析”，讲解应用 Apache Math API 进行数据收集乃至分析统计指标的内容。本章还包含一些高级概念，比如统计显著性检验这一标准的工具，科研人员可以通过它将得到的结果与基准数据进行比较。

第 4 章 “数据学习 I ”，包含使用 Weka 机器学习库进行分类、聚类、特征选择的内容。

第 5 章 “数据学习 II ”，本章是前一章的后续内容，讲解使用另一个 Java 库——Java-ML 库进行数据导入导出、分类、特征选择的内容。本章还包含使用斯坦福分类器（Stanford Classifier）与 MOA（Massive Online Access）进行基本分类的内容。

第 6 章 “从文本数据提取信息”，介绍对文本数据应用数据科学技术以提取信息的方法。内容涉及 Java 核心应用以及 OpenNLP、Stanford CoreNLP、Mallet、Weka 等著名的机器学习库，学习如何使用它们来完成信息提取与检索任务。

第 7 章 “处理大数据”，本章涵盖机器学习大数据平台应用的内容，比如 Apache Mahout、Spark-MLib。

第 8 章 “数据深度学习”，包含使用 DL4j 库进行深度学习的基础内容，介绍 word2vec 算法、信念网络与自动编码器。

第 9 章 “数据可视化”，讲解如何使用 GRAL 包为数据生成具有吸引力的信息展示图表。这个包功能众多，我们只讲解其中最基础的绘图功能。

阅读本书需要具备的知识

在本书中，我们使用 Java 来解决各种实际的数据科学问题。对于那些想了解如何使用 Java 解决问题的朋友，本书所讲解的内容正是他们所需要的。阅读本书需要你具备最基本的 Java 知识，比如懂得 Java 类、对象、方法、实参与形参、异常、导出 JAR 文件等内容。书中给出的代码都配有相应的讲解、介绍以及提示，这有助于各位读者更好地理解它们。对于书中所解决问题的背后原理，大部分我们都不会进行详细讲解，但必要时我们会提供相应的参考内容，以供感兴趣的读者进一步学习。

本书的目标读者

如果你想了解如何使用 Java 解决现实世界中与数据科学相关的问题，那么本书正是为

你准备的。在内容覆盖方面，由于本书内容涵盖数据科学的方方面面，因此对于那些正在从事数据学习相关工作，并寻求使用 Java 解决项目问题的朋友而言，本书也具有十分重要的参考价值。

结构安排

在本书中，你将经常看到如下几个标题：“准备工作”“操作步骤”“工作原理”“更多内容”“另见”。

对于本书的每一节内容，我们会使用如下几个小标题来组织相关内容。

准备工作

本部分指出学习本节内容需要做的准备，也包含安装一些必需软件或做一些预先设置的内容。

操作步骤

本部分包含跟学中要做的具体步骤。

工作原理

本部分通常讲解与前一部分内容相关的更多细节。

更多内容

本部分讲解与前面内容相关的更多知识，通过阅读本部分内容，让读者掌握更多相关知识。

另见

本部分提供了一些有用的页面链接，从中读者可以获取更多与当前主题相关的有用

内容。

本书使用说明

在本书中，在不同类型的信息之间，你将看到大量不同的文本类型。下面给出了这些类型的一些示例，并对它们所代表的含义进行了说明。

正文中出现的代码用语、数据表名、文件夹名、文件名、文件扩展名、路径名、虚拟URL、用户输入、推特标签显示如下："在它们之间，你会发现一个名为 `lib` 的文件夹，它就是感兴趣的文件夹。"

代码块设置如下：

```
classVals = new ArrayList<String>();
  for (int i = 0; i < 5; i++){
    classVals.add("class" + (i + 1));
}
```

命令行输入或输出写成如下形式：

```
@relation MyRelation

@attribute age numeric
@attribute name string
@attribute dob date yyyy-MM-dd
@attribute class {class1,class2,class3,class4,class5}

@data
35,'John Doe',1981-01-20,class3
30,'Harry Potter',1986-07-05,class1
```

正文中的新术语与关键词以粗体形式标识出来。你在屏幕截图中看到的词，比如在菜单或对话框中，出现在正文中的形式如下："从 Administration 面板选择 System info"。

警告或重要注释出现在这里

提示与技巧出现在这里

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

配套资源

本书提供如下资源：

- 本书源代码；
- 书中彩图文件。

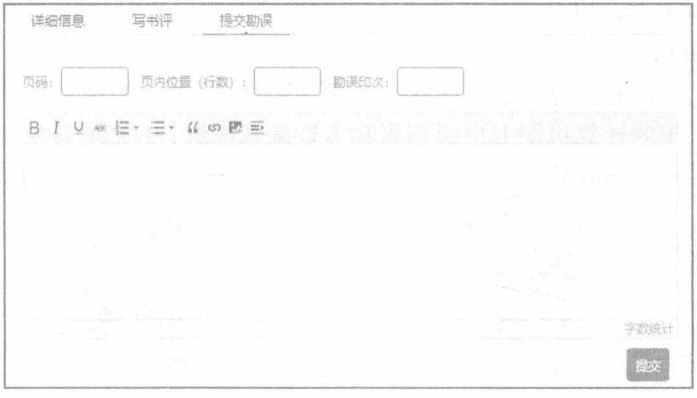
要获得以上配套资源，请在异步社区本书页面中点击 配套资源，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

如果您是教师，希望获得教学配套资源，请在社区本书页面中直接联系本书的责任编辑。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，按书名搜索，进入本书页面，点击“提交勘误”，输入勘误信息，单击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券、样书或奖品。



扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们；有意出版图书的作者也可以到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权益的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下 IT 专业图书社区，致力于出版精品 IT 技术图书和相关学习产品，为作译者提供优质出版服务。异步社区创办于 2015 年 8 月，提供大量精品 IT 技术图书和电子书，以及高品质技术文章和视频课程。更多详情请访问异步社区官网 https://www.epubit.com。

“异步图书”是由异步社区编辑团队策划出版的精品 IT 专业图书的品牌，依托于人民邮电出版社近 30 年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的 LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

异步社区

微信服务号

目录