

Rushdi Shams

Java Data Science Cookbook

Explore the power of MLlib, DL4j, Weka, and more



Packt>

Java Data Science Cookbook

If you are looking to build data science models that are good for production, Java has come to the rescue. With the aid of strong libraries such as MLlib, Weka, DL4j, and more, you can efficiently perform all the data science tasks you need to.

This unique book provides modern recipes to solve your common and not-so-common data science-related problems. We start with recipes to help you obtain, clean, index, and search data. Then you will learn a variety of techniques to analyze, learn from, and retrieve information from data. You will also understand how to handle big data, learn deeply from data, and visualize data.

Finally, you will work through unique recipes that solve your problems while taking data science to production, writing distributed data science applications, and much more—things that will come in handy at work.

Things you will learn:

- Find out how to clean and make datasets ready so you can acquire actual insights by removing noise and outliers
- Develop the skills needed to use modern machine learning techniques to retrieve information and transform data to knowledge
- Familiarize yourself with cutting-edge techniques to store and search large volumes of data and retrieve information from large amounts of data in text format
- Develop the basic skills needed to apply big data and deep learning technologies to large volumes of data
- Evolve your data visualization skills and gain valuable insights from your data
- Learn about a step-by-step formula to help you develop an industry-standard, large-scale, real-life data product

Packt
www.packtpub.com

\$ 49.99 US
£ 41.99 UK

Prices do not include local sales
Tax or VAT where applicable



Java Data Science Cookbook

Rushodi Shams



Java Data Science Cookbook

Explore the power of MLlib, DL4j, Weka, and more

Rushdi Shams



BIRMINGHAM - MUMBAI

Java Data Science Cookbook

Copyright © 2017 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: March 2017

Production reference: 1240317

Published by Packt Publishing Ltd.

Livery Place

35 Livery Street

Birmingham

B3 2PB, UK.

ISBN 978-1-78712-253-6

www.packtpub.com

Credits

Author

Rushdi Shams

Copy Editors

Vikrant Phadke

Manisha Sinha

Reviewer

Prashant Verma

Project Coordinator

Nidhi Joshi

Commissioning Editor

Veena Pagare

Proofreader

Safis Editing

Acquisition Editor

Ajith Menon

Indexer

Aishwarya Gangawane

Content Development Editor

Cheryl Dsa

Graphics

Tania Dutta

Technical Editor

Dharmendra Yadav

Production Coordinator

Arvindkumar Gupta

About the Author

Rushdi Shams has a PhD on application of machine learning in Natural Language Processing (NLP) problem areas from Western University, Canada. Before starting work as a machine learning and NLP specialist in industry, he was engaged in teaching undergrad and grad courses. He has been successfully maintaining his YouTube channel named "Learn with Rushdi" for learning computer technologies.

I would like to acknowledge the Almighty Allah for giving me the strength, support, and knowledge to finish the book.

I extend my thanks to my family members, friends, and colleagues for continuous support, encouragement, and constructive criticism.

I would also like to thank Ajith and Cheryl from Packt for their continuous and spontaneous collaboration with me.

About the Reviewer

Prashant Verma started his IT career in 2011 as a Java developer at Ericsson, working in the telecom domain. After a couple of years of Java EE experience, he moved into the big data domain, and has worked on almost all the popular big data technologies such as Hadoop, Spark, Kafka, Flume, Mongo, Cassandra, and so on. He has also worked in Scala and Python. Currently, he works with QA Infotech as Lead Data Engineer, working on solving e-learning domain problems using data analytics and machine learning.

Prashant has also worked on *Apache Spark for Java Developers*, Packt as a Technical Reviewer.

I want to thank Packt Publishing for giving me the chance to review the book, as well as my employer and my family for their patience while I was busy working on this book.

www.PacktPub.com

For support files and downloads related to your book, please visit www.PacktPub.com.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details.

At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.



<https://www.packtpub.com/mapt>

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at <https://www.amazon.com/dp/1787122530>.

If you'd like to join our team of regular reviewers, you can e-mail us at customerreviews@packtpub.com. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products! .

To my lovely wife, Mah-Zereen, and adorable daughter, Ruayda.

Table of Contents

| | |
|---|----|
| Preface | 1 |
| Chapter 1: Obtaining and Cleaning Data | 9 |
| Introduction | 9 |
| Retrieving all filenames from hierarchical directories using Java | 11 |
| Getting ready | 11 |
| How to do it... | 11 |
| Retrieving all filenames from hierarchical directories using Apache Commons IO | 13 |
| Getting ready | 13 |
| How to do it... | 13 |
| Reading contents from text files all at once using Java 8 | 15 |
| How to do it... | 15 |
| Reading contents from text files all at once using Apache Commons IO | 16 |
| Getting ready | 16 |
| How to do it... | 16 |
| Extracting PDF text using Apache Tika | 17 |
| Getting ready | 18 |
| How to do it... | 18 |
| Cleaning ASCII text files using Regular Expressions | 20 |
| How to do it... | 20 |
| Parsing Comma Separated Value (CSV) Files using Univocity | 21 |
| Getting ready | 22 |
| How to do it... | 22 |
| Parsing Tab Separated Value (TSV) file using Univocity | 24 |
| Getting ready | 25 |
| How to do it... | 25 |
| Parsing XML files using JDOM | 26 |
| Getting ready | 27 |
| How to do it... | 27 |
| Writing JSON files using JSON.simple | 30 |
| Getting ready | 30 |
| How to do it... | 30 |
| Reading JSON files using JSON.simple | 33 |
| Getting ready | 33 |

| | |
|--|----|
| How to do it ... | 33 |
| Extracting web data from a URL using JSoup | 35 |
| Getting ready | 36 |
| How to do it... | 36 |
| Extracting web data from a website using Selenium Webdriver | 38 |
| Getting ready | 39 |
| How to do it... | 39 |
| Reading table data from a MySQL database | 42 |
| Getting ready | 42 |
| How to do it... | 43 |
| Chapter 2: Indexing and Searching Data | 47 |
| Introduction | 47 |
| Indexing data with Apache Lucene | 47 |
| Getting ready | 48 |
| How to do it... | 54 |
| How it works... | 63 |
| Searching indexed data with Apache Lucene | 65 |
| Getting ready | 66 |
| How to do it... | 67 |
| Chapter 3: Analyzing Data Statistically | 73 |
| Introduction | 74 |
| Generating descriptive statistics | 76 |
| How to do it... | 76 |
| Generating summary statistics | 77 |
| How to do it... | 78 |
| Generating summary statistics from multiple distributions | 79 |
| How to do it... | 79 |
| There's more... | 81 |
| Computing frequency distribution | 81 |
| How to do it... | 81 |
| Counting word frequency in a string | 82 |
| How to do it... | 83 |
| How it works... | 84 |
| Counting word frequency in a string using Java 8 | 84 |
| How to do it... | 85 |
| Computing simple regression | 86 |
| How to do it... | 86 |
| Computing ordinary least squares regression | 87 |

| | |
|---|-----|
| How to do it... | 87 |
| Computing generalized least squares regression | 90 |
| How to do it... | 90 |
| Calculating covariance of two sets of data points | 92 |
| How to do it... | 92 |
| Calculating Pearson's correlation of two sets of data points | 93 |
| How to do it... | 93 |
| Conducting a paired t-test | 94 |
| How to do it... | 94 |
| Conducting a Chi-square test | 96 |
| How to do it... | 96 |
| Conducting the one-way ANOVA test | 97 |
| How to do it... | 97 |
| Conducting a Kolmogorov-Smirnov test | 99 |
| How to do it... | 99 |
| Chapter 4: Learning from Data - Part 1 | 101 |
| Introduction | 101 |
| Creating and saving an Attribute-Relation File Format (ARFF) file | 102 |
| How to do it... | 106 |
| Cross-validating a machine learning model | 110 |
| How to do it... | 111 |
| Classifying unseen test data | 114 |
| Getting ready | 114 |
| How to do it... | 116 |
| Classifying unseen test data with a filtered classifier | 122 |
| How to do it... | 122 |
| Generating linear regression models | 125 |
| How to do it... | 125 |
| Generating logistic regression models | 127 |
| How to do it... | 127 |
| Clustering data points using the KMeans algorithm | 130 |
| How to do it... | 130 |
| Clustering data from classes | 133 |
| How to do it... | 133 |
| Learning association rules from data | 135 |
| Getting ready | 136 |
| How to do it... | 136 |
| Selecting features/attributes using the low-level method, the filtering method, and the meta-classifier method | 138 |

| | |
|--|------------|
| Getting ready | 139 |
| How to do it... | 139 |
| Chapter 5: Learning from Data - Part 2 | 145 |
| Introduction | 145 |
| Applying machine learning on data using Java Machine Learning (Java-ML) library | 146 |
| Getting ready | 146 |
| How to do it... | 150 |
| Classifying data points using the Stanford classifier | 159 |
| Getting ready | 159 |
| How to do it... | 163 |
| How it works... | 164 |
| Classifying data points using Massive Online Analysis (MOA) | 165 |
| Getting ready | 166 |
| How to do it... | 168 |
| Classifying multilabeled data points using Mulan | 171 |
| Getting ready | 171 |
| How to do it... | 175 |
| Chapter 6: Retrieving Information from Text Data | 179 |
| Introduction | 179 |
| Detecting tokens (words) using Java | 180 |
| Getting ready | 180 |
| How to do it... | 180 |
| Detecting sentences using Java | 185 |
| Getting ready | 185 |
| How to do it... | 185 |
| Detecting tokens (words) and sentences using OpenNLP | 187 |
| Getting ready | 187 |
| How to do it... | 189 |
| Retrieving lemma, part-of-speech, and recognizing named entities from tokens using Stanford CoreNLP | 193 |
| Getting ready | 193 |
| How to do it... | 195 |
| Measuring text similarity with Cosine Similarity measure using Java 8 | 198 |
| Getting ready | 198 |
| How to do it... | 199 |
| Extracting topics from text documents using Mallet | 203 |
| Getting ready | 204 |