SPRINGERBRIEFS IN STATISTICS

Dominique Haughton
Mark-David McLaughlin
Kevin Mentzer
Changan Zhang

Movie Analytics

A Hollywood Introduction to Big Data

Springer

SpringerBriefs in Statistics

More information about this series at http://www.springer.com/series/8921

Dominique Haughton
Mark-David McLaughlin
Kevin Mentzer
Changan Zhang

Movie Analytics

A Hollywood Introduction to Big Data

Springer

Dominique Haughton
Bentley University
Université Paris 1 Panthéon-Sorbonne (SAMM) and Université Toulouse 1 (GREMAQ)
Waltham, MA, USA

Kevin Mentzer
Business
Bentley University and Bryant University
Waltham, MA, USA

Mark-David McLaughlin
Information and Process Management
Bentley University and Cisco Systems
Waltham, MA, USA

Changan Zhang
Business
Bentley University and CTrip
Waltham, MA, USA

ISSN 2191-544X ISSN 2191-5458 (electronic)
SpringerBriefs in Statistics
ISBN 978-3-319-09425-0 ISBN 978-3-319-09426-7 (eBook)
DOI 10.1007/978-3-319-09426-7

Library of Congress Control Number: 2015946241

Springer Cham Heidelberg New York Dordrecht London
© The Author(s) 2015
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Acknowledgements

With our most sincere gratitude to Nathalie Villa-Vialaneix for contributing Chap. 4 to this monograph and to Veo (veocinemas.fr) for most kindly providing the data used in Chap. 4. We also wish to warmly thank Ignacio Alvarez-Hamelin for very useful conversations about the LaNet-vi visualization technique.

Contents

What Do We Know About Analyzing Movie Data? 1

What Does "Big Data" Mean? The Data Scientist Point of View 3

Visualization of Very Large Networks: The Co-starring Social Network 5
1 Introduction 5
2 The IMDb Data 5
3 Issues with Large Datasets 8
4 Configuring Python and Gephi to Use the 64-Bit Libraries 9
4.1 Python 9
4.2 Gephi 12
5 Visualizations of the Co-starring Network 14
6 The Pajek and Gephi Approaches 15
6.1 The Gephi Approach 15
6.2 Loading the Data 15
6.3 Partitioning the Data 16
6.4 Displaying the Network 16
6.5 Interpreting the Data 17
7 K-Core Representations of the IMDb Co-starring Network 19
7.1 K-Core Decomposition 19
8 K-Core Representations of the IMDb Co-starring Network Using LaNet-vi2 21

Movie Attendance and Trends 25
1 Importing the Data 25
2 Basic Description of the Data 27
3 Weekly Number of Movie Goers 28
4 Evolution of the Weekly Number of Movie Goers 30
5 Trends 33