O'REILLY®

2nd Edition

Mining the Social Web

DATA MINING FACEBOOK, TWITTER, LINKEDIN, GOOGLE+, GITHUB, AND MORE

Matthew A. Russell

www.it-ebooks.info

O'REILLY®

Mining the Social Web

How can you tap into the wealth of social web data to discover who's making connections with whom, what they're talking about, and where they're located? With this expanded and thoroughly revised edition, you'll learn how to acquire, analyze, and summarize data from all corners of the social web, including Facebook, Twitter, LinkedIn, Google+, GitHub, email, websites, and blogs.

- Employ IPython Notebook, the Natural Language Toolkit, NetworkX, and other scientific computing tools to mine popular social websites
- Apply advanced text-mining techniques, such as clustering and TF-IDF, to extract meaning from human language data
- Bootstrap interest graphs from GitHub by discovering affinities among people, programming languages, and coding projects
- Build interactive visualizations with D3.js, an extraordinarily flexible HTML5 and JavaScript toolkit
- Take advantage of more than two-dozen Twitter recipes, presented in O'Reilly's popular "problem/solution/discussion" cookbook format

The example code for this unique data science book is maintained in a public GitHub repository. It's designed to be easily accessible through a turnkey virtual machine that facilitates interactive learning with an easy-to-use collection of IPython Notebooks.

Matthew Russell, Chief Technology Officer at Digital Reasoning Systems, Principal at Zaffra, is a computer scientist who is passionate about data mining, open source, and creating technology to amplify human intelligence.

"Mining insights through an API is an essential skill to have, whether or not you consider yourself a programmer. This book exposes you to a breadth of key information sources while using tools that make the coding easily accessible."

—**Kevin Makice**
author of *Twitter API: Up and Running*

"This book offers all readers a fresh perspective of social web data through illustrative and concise code—all within in the comfort of a web browser! Readers get a fantastic tour of computer science concepts by example, including algorithmic complexity, natural language processing, and the future of the Internet of Things."

—**Jason Yee**
Data Scientist at Digital Reasoning

Strata
Making Data Work

Strata is the emerging ecosystem of people, tools, and technologies that turn big data into smart decisions. Find information and resources at oreilly.com/data.

DATA/DATA MINING

US $44.99 CAN $47.99

ISBN: 978-1-449-36761-9

Twitter: @oreillymedia
facebook.com/oreilly

www.it-ebooks.info

O'REILLY®
Strata
Making Data Work

Learn how to turn data into decisions.

From startups to the Fortune 500, smart companies are betting on data-driven insight, seizing the opportunities that are emerging from the convergence of four powerful trends:

- New methods of collecting, managing, and analyzing data
- Cloud computing that offers inexpensive storage and flexible, on-demand computing power for massive data sets
- Visualization techniques that turn complex data into images that tell a compelling story
- Tools that make the power of data available to anyone

Get control over big data and turn it into insight with O'Reilly's Strata offerings. Find the inspiration and information to create new products or revive existing ones, understand customer behavior, and get the data edge.

O'REILLY®

Visit oreilly.com/data to learn more.

©2011 O'Reilly Media, Inc. O'Reilly logo is a registered trademark of O'Reilly Media, Inc.

www.it-ebooks.info

www.it-ebooks.info

SECOND EDITION

Mining the Social Web

Matthew A. Russell

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

www.it-ebooks.info

Mining the Social Web, Second Edition

by Matthew A. Russell

Copyright © 2014 Matthew A. Russell. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Mary Treseler
Production Editor: Kristen Brown
Copyeditor: Rachel Monaghan
Proofreader: Rachel Head

Indexer: Lucie Haskins
Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator: Rebecca Demarest

October 2013: Second Edition

Revision History for the Second Edition:

2013-09-25: First release

See *http://oreilly.com/catalog/errata.csp?isbn=9781449367619* for release details.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. *Mining the Social Web*, the image of a groundhog, and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and author assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-36761-9

[LSI]

www.it-ebooks.info

If the ax is dull and its edge unsharpened, more strength is needed,
but skill will bring success.

—Ecclesiastes 10:10

www.it-ebooks.info

www.it-ebooks.info

Table of Contents

Preface xiii

Part I. A Guided Tour of the Social Web

Prelude 3

1. Mining Twitter: Exploring Trending Topics, Discovering What People Are Talking About, and More 5

1.1. Overview 6
1.2. Why Is Twitter All the Rage? 6
1.3. Exploring Twitter's API 9
1.3.1. Fundamental Twitter Terminology 9
1.3.2. Creating a Twitter API Connection 12
1.3.3. Exploring Trending Topics 15
1.3.4. Searching for Tweets 20
1.4. Analyzing the 140 Characters 26
1.4.1. Extracting Tweet Entities 28
1.4.2. Analyzing Tweets and Tweet Entities with Frequency Analysis 29
1.4.3. Computing the Lexical Diversity of Tweets 32
1.4.4. Examining Patterns in Retweets 34
1.4.5. Visualizing Frequency Data with Histograms 36
1.5. Closing Remarks 41
1.6. Recommended Exercises 42
1.7. Online Resources 43

2. Mining Facebook: Analyzing Fan Pages, Examining Friendships, and More 45

2.1. Overview 46
2.2. Exploring Facebook's Social Graph API 46
2.2.1. Understanding the Social Graph API 48
2.2.2. Understanding the Open Graph Protocol 54

www.it-ebooks.info

2.3. Analyzing Social Graph Connections 59
2.3.1. Analyzing Facebook Pages 63
2.3.2. Examining Friendships 70
2.4. Closing Remarks 85
2.5. Recommended Exercises 85
2.6. Online Resources 86

3. Mining LinkedIn: Faceting Job Titles, Clustering Colleagues, and More. 89
3.1. Overview 90
3.2. Exploring the LinkedIn API 90
3.2.1. Making LinkedIn API Requests 91
3.2.2. Downloading LinkedIn Connections as a CSV File 96
3.3. Crash Course on Clustering Data 97
3.3.1. Clustering Enhances User Experiences 100
3.3.2. Normalizing Data to Enable Analysis 101
3.3.3. Measuring Similarity 112
3.3.4. Clustering Algorithms 115
3.4. Closing Remarks 131
3.5. Recommended Exercises 132
3.6. Online Resources 133

4. Mining Google+: Computing Document Similarity, Extracting Collocations, and More 135
4.1. Overview 136
4.2. Exploring the Google+ API 136
4.2.1. Making Google+ API Requests 138
4.3. A Whiz-Bang Introduction to TF-IDF 147
4.3.1. Term Frequency 148
4.3.2. Inverse Document Frequency 150
4.3.3. TF-IDF 151
4.4. Querying Human Language Data with TF-IDF 155
4.4.1. Introducing the Natural Language Toolkit 155
4.4.2. Applying TF-IDF to Human Language 158
4.4.3. Finding Similar Documents 160
4.4.4. Analyzing Bigrams in Human Language 167
4.4.5. Reflections on Analyzing Human Language Data 177
4.5. Closing Remarks 178
4.6. Recommended Exercises 179
4.7. Online Resources 180

5. Mining Web Pages: Using Natural Language Processing to Understand Human Language, Summarize Blog Posts, and More. 181
5.1. Overview 182

www.it-ebooks.info