Distilling Rich Information from Messy Data

21 Recipes for Mining Twitter

O'REILLY®

Matthew A. Russell

21 Recipes for Mining Twitter

21 Recipes for Mining Twitter

Matthew A. Russell

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

21 Recipes for Mining Twitter
by Matthew A. Russell

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: Mike Loukides
Production Editor: Kristen Borg
Proofreader: Kristen Borg

Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator: Robert Romano

Printing History:
February 2011: First Edition.

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: 978-1-449-30316-7

[LSI]

1296485191

Table of Contents

Preface

Introduction

This intentionally terse recipe collection provides you with 21 easily adaptable Twitter mining recipes and is a spin-off of Mining the Social Web (O'Reilly), a more comprehensive work that covers a much larger cross-section of the social web and related analysis. Think of this ebook as the jetpack that you can strap onto that great Twitter mining idea you've been noodling on—whether it's as simple as running some disposible scripts to crunch some numbers, or as extensive as creating a full-blown interactive web application.

All of the recipes in this book are written in Python, and if you are reasonably confident with any other programming language, you'll be able to quickly get up to speed and become productive with virtually no trouble at all. Beyond the Python language itself, you'll also want to be familiar with `easy_install` (*http://pypi.python.org/pypi/setuptools*) so that you can get third-party packages that we'll be using along the way. A great warmup for this ebook is Chapter 1 (Hacking on Twitter Data) from Mining the Social Web. It walks you through tools like `easy_install` and discusses specific environment issues that might be helpful—and the best news is that you can download a full resolution copy, absolutely free!

One other thing you should consider doing up front, if you haven't already, is quickly skimming through the official Twitter API documentation and related development documents linked on that page. Twitter has a very easy-to-use API with a lot of degrees of freedom, and `twitter` (*http://github.com/sixohsix/twitter*), a third-party package we'll use extensively, is a beautiful wrapper around the API. Once you know a little bit about the API, it'll quickly become obvious how to interact with it using `twitter`.

Finally—enjoy! And be sure to follow @SocialWebMining on Twitter or "like" the Mining the Social Web Facebook page to stay up to date with the latest updates, news, additional content, and more.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
: Indicates new terms, URLs, email addresses, filenames, and file extensions.

`Constant width`
: Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

`Constant width bold`
: Shows commands or other text that should be typed literally by the user.

`Constant width italic`
: Shows text that should be replaced with user-supplied values or by values determined by context.

This icon signifies a tip, suggestion, or general note.

This icon indicates a warning or caution.

Using Code Examples

This book is here to help you get your job done. In general, you may use the code in this book in your programs and documentation. You do not need to contact us for permission unless you're reproducing a significant portion of the code. For example, writing a program that uses several chunks of code from this book does not require permission. Selling or distributing a CD-ROM of examples from O'Reilly books does require permission. Answering a question by citing this book and quoting example code does not require permission. Incorporating a significant amount of example code from this book into your product's documentation does require permission.

We appreciate, but do not require, attribution. An attribution usually includes the title, author, publisher, and ISBN. For example: "*21 Recipes for Mining Twitter* by Matthew A. Russell (O'Reilly). Copyright 2011 Matthew A. Russell, 978-1-449-30316-7."

If you feel your use of code examples falls outside fair use or the permission given above, feel free to contact us at *permissions@oreilly.com*.

Safari® Books Online

Safari Books Online is an on-demand digital library that lets you easily search over 7,500 technology and creative reference books and videos to find the answers you need quickly.

With a subscription, you can read any page and watch any video from our library online. Read books on your cell phone and mobile devices. Access new titles before they are available for print, and get exclusive access to manuscripts in development and post feedback for the authors. Copy and paste code samples, organize your favorites, download chapters, bookmark key sections, create notes, print out pages, and benefit from tons of other time-saving features.

O'Reilly Media has uploaded this book to the Safari Books Online service. To have full digital access to this book and others on similar topics from O'Reilly and other publishers, sign up for free at *http://my.safaribooksonline.com*.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at:

http://oreilly.com/catalog/9781449303167

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*