O'REILLY®

第2版

Mining the Social Web

挖掘社交网络（影印版）

东南大學出版社

Matthew A. Russell 著

第2版

挖掘社交网络（影印版）

Mining the Social Web

Matthew A. Russell 著

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

O'Reilly Media, Inc.授权东南大学出版社出版

南京 东南大学出版社

图书在版编目（CIP）数据

挖掘社交网络:第2版:英文/(美)罗塞尔(Russell, M.A.)著.—影印本.—南京：东南大学出版社，2014.10

书名原文：Mining the Social Web, 2E

ISBN 978-7-5641-5005-1

I. ①挖… II. ①罗… III. ①互联网络－应用－人际关系学－英文 IV. ①C912.1-39

中国版本图书馆CIP数据核字（2014）第115561号

江苏省版权局著作权合同登记
图字：10-2013-373号

挖掘社交网络 第2版（影印版）

出版发行：东南大学出版社
地　　址：南京四牌楼2号　　邮编：210096
出 版 人：江建中
网　　址：http://www.seupress.com
电子邮件：press@seupress.com
印　　刷：常州市武进第三印刷有限公司
开　　本：787毫米×980毫米　16开本
印　　张：28
字　　数：548千字
版　　次：2014年10月第1版
印　　次：2014年10月第1次印刷
书　　号：ISBN 978-7-5641-5005-1
定　　价：78.00元

本社图书若有印装质量问题，请直接与营销部联系。电话（传真）：025-83791830

挖掘社交网络（影印版）

Mining the Social Web

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

—Ecclesiastes 10:10

Preface

The Web is more a social creation
than a technical one.
I designed it for a social effect—to
help people work together—and not as a
technical toy. The ultimate goal of the Web
is to support and improve our weblike existence
in the world. We clump into families, associations,
and companies. We develop trust across the miles
and distrust around the corner.

—Tim Berners-Lee, *Weaving the Web* (Harper)

README.1st

This book has been carefully designed to provide an incredible learning experience for a particular target audience, and in order to avoid any unnecessary confusion about its scope or purpose by way of disgruntled emails, bad book reviews, or other misunderstandings that can come up, the remainder of this preface tries to help you determine whether you are part of that target audience. As a very busy professional, I consider my time my most valuable asset, and I want you to know right from the beginning that I believe that the same is true of you. Although I often fail, I really do try to honor my neighbor above myself as I walk out this life, and this preface is my attempt to honor you, the reader, by making it clear whether or not this book can meet your expectations.

Managing Your Expectations

Some of the most basic assumptions this book makes about you as a reader is that you want to learn how to mine data from popular social web properties, avoid technology hassles when running sample code, and have *lots* of fun along the way. Although you could read this book solely for the purpose of learning what is possible, you should know up front that it has been written in such a way that you really could follow along with the many exercises and become a data miner once you've completed the few simple steps to set up a development environment. If you've done some programming before, you should find that it's relatively painless to get up and running with the code examples. Even if you've never programmed before but consider yourself the least bit tech-savvy, I daresay that you could use this book as a starting point to a remarkable journey that will stretch your mind in ways that you probably haven't even imagined yet.

To fully enjoy this book and all that it has to offer, you need to be interested in the vast possibilities for mining the rich data tucked away in popular social websites such as Twitter, Facebook, LinkedIn, and Google+, and you need to be motivated enough to download a virtual machine and follow along with the book's example code in IPython Notebook, a fantastic web-based tool that features all of the examples for every chapter. Executing the examples is usually as easy as pressing a few keys, since all of the code is presented to you in a friendly user interface. This book will teach you a few things that you'll be thankful to learn and will add a few indispensable tools to your toolbox, but perhaps even more importantly, it will tell you a story and entertain you along the way. It's a story about data science involving social websites, the data that's tucked away inside of them, and some of the intriguing possibilities of what you (or anyone else) could do with this data.

If you were to read this book from cover to cover, you'd notice that this story unfolds on a chapter-by-chapter basis. While each chapter roughly follows a predictable template that introduces a social website, teaches you how to use its API to fetch data, and introduces some techniques for data analysis, the broader story the book tells crescendos in complexity. Earlier chapters in the book take a little more time to introduce fundamental concepts, while later chapters systematically build upon the foundation from earlier chapters and gradually introduce a broad array of tools and techniques for mining the social web that you can take with you into other aspects of your life as a data scientist, analyst, visionary thinker, or curious reader.

Some of the most popular social websites have transitioned from fad to mainstream to household names over recent years, changing the way we live our lives on and off the Web and enabling technology to bring out the best (and sometimes the worst) in us. Generally speaking, each chapter of this book interlaces slivers of the social web along with data mining, analysis, and visualization techniques to explore data and answer the following representative questions:

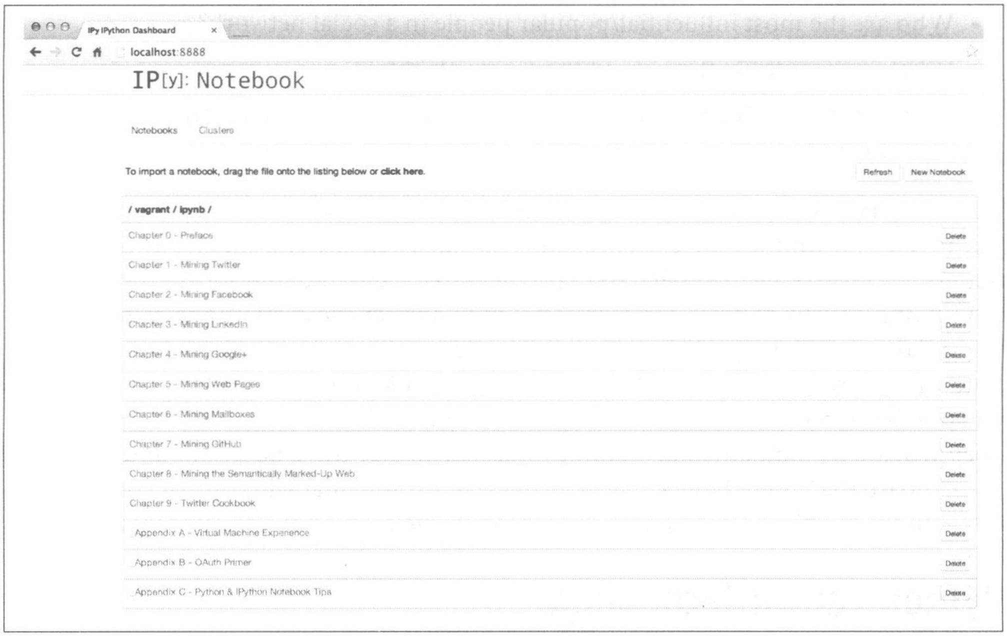
- Who knows whom, and which people are common to their social networks?
- How frequently are particular people communicating with one another?
- Which social network connections generate the most value for a particular niche?
- How does geography affect your social connections in an online world?

- Who are the most influential/popular people in a social network?
- What are people chatting about (and is it valuable)?
- What are people interested in based upon the human language that they use in a digital world?

The answers to these basic kinds of questions often yield valuable insight and present lucrative opportunities for entrepreneurs, social scientists, and other curious practitioners who are trying to understand a problem space and find solutions. Activities such as building a turnkey killer app from scratch to answer these questions, venturing far beyond the typical usage of visualization libraries, and constructing just about anything state-of-the-art are not within the scope of this book. You'll be really disappointed if you purchase this book because you want to do one of those things. However, this book does provide the fundamental building blocks to answer these questions and provide a springboard that might be exactly what you need to build that killer app or conduct that research study. Skim a few chapters and see for yourself. This book covers a lot of ground.

Python-Centric Technology

This book intentionally takes advantage of the Python programming language for all of its example code. Python's intuitive syntax, amazing ecosystem of packages that trivialize API access and data manipulation, and core data structures that are practically JSON (*http://bit.ly/1a1kFaF*) make it an excellent teaching tool that's powerful yet also very easy to get up and running. As if that weren't enough to make Python both a great pedagogical choice and a very pragmatic choice for mining the social web, there's IPython Notebook (*http://bit.ly/1a1kFr4*), a powerful, interactive Python interpreter that provides a notebook-like user experience from within your web browser and combines code execution, code output, text, mathematical typesetting, plots, and more. It's difficult to imagine a better user experience for a learning environment, because it trivializes the problem of delivering sample code that you as the reader can follow along with and execute with no hassles. Figure P-1 provides an illustration of the IPython Notebook experience, demonstrating the dashboard of notebooks for each chapter of the book. Figure P-2 shows a view of one notebook.

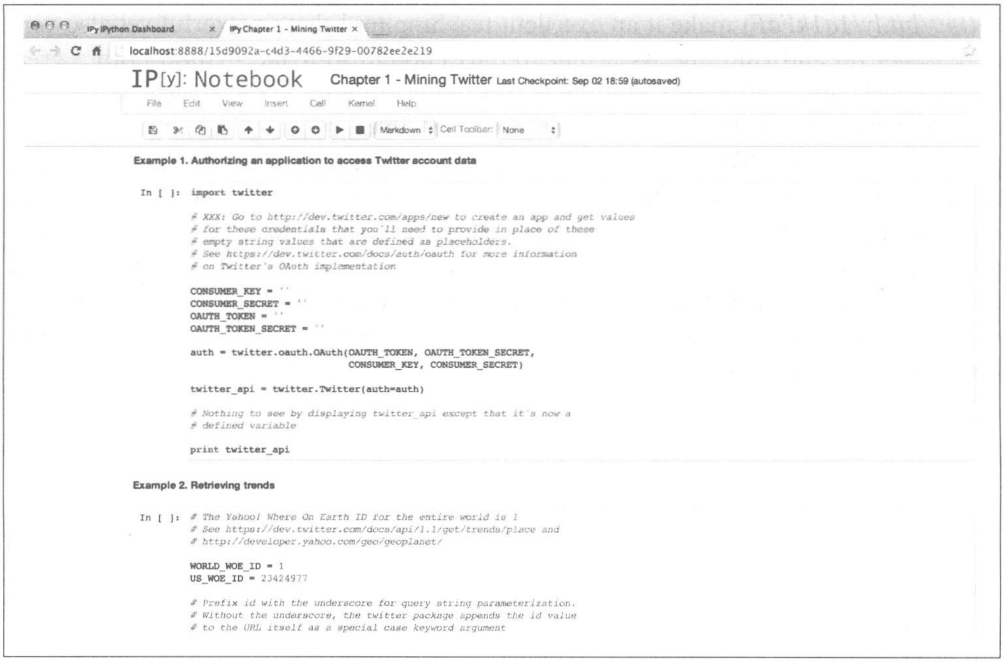


To import a notebook, drag the file onto the listing below or **click here.**

/ vagrant / ipynb /

Chapter 0 - Preface
Chapter 1 - Mining Twitter
Chapter 2 - Mining Facebook
Chapter 3 - Mining LinkedIn
Chapter 4 - Mining Google+
Chapter 5 - Mining Web Pages
Chapter 6 - Mining Mailboxes
Chapter 7 - Mining GitHub
Chapter 8 - Mining the Semantically Marked-Up Web
Chapter 9 - Twitter Cookbook
Appendix A - Virtual Machine Experience
Appendix B - OAuth Primer
Appendix C - Python & IPython Notebook Tips

Figure P-1. Overview of IPython Notebook; a dashboard of notebooks


IPy IPython Dashboard
IPy Chapter 1 - Mining Twitter
localhost:8888/15d9092a-c4d3-4466-9f29-00782ee2e219
IP[y]: Notebook
Chapter 1 - Mining Twitter
Last Checkpoint: Sep 02 18:59 (autosaved)
File
Edit
View
Insert
Cell
Kernel
Help
Markdown
Cell Toolbar:
None

Example 1. Authorizing an application to access Twitter account data

```
In [ ]: import twitter

        # XXX: Go to http://dev.twitter.com/apps/new to create an app and get values
        # for these credentials that you'll need to provide in place of these
        # empty string values that are defined as placeholders.
        # See https://dev.twitter.com/docs/auth/oauth for more information
        # on Twitter's OAuth implementation

        CONSUMER_KEY = ''
        CONSUMER_SECRET = ''
        OAUTH_TOKEN = ''
        OAUTH_TOKEN_SECRET = ''

        auth = twitter.oauth.OAuth(OAUTH_TOKEN, OAUTH_TOKEN_SECRET,
                                   CONSUMER_KEY, CONSUMER_SECRET)

        twitter_api = twitter.Twitter(auth=auth)

        # Nothing to see by displaying twitter_api except that it's now a
        # defined variable

        print twitter_api
```

Example 2. Retrieving trends

```
In [ ]: # The Yahoo! Where On Earth ID for the entire world is 1
        # See https://dev.twitter.com/docs/api/1.1/get/trends/place and
        # http://developer.yahoo.com/geo/geoplanet/

        WORLD_WOE_ID = 1
        US_WOE_ID = 23424977

        # Prefix id with the underscore for query string parameterization.
        # Without the underscore, the twitter package appends the id value
        # to the URL itself as a special case keyword argument
```

Figure P-2. Overview of IPython Notebook; the "Chapter 1-Mining Twitter" notebook

Every chapter in this book has a corresponding IPython Notebook with example code that makes it a pleasure to study the code, tinker around with it, and customize it for your own purposes. If you've done some programming but have never seen Python syntax, skimming ahead a few pages should hopefully be all the confirmation that you need. Excellent documentation is available online, and the official Python tutorial (*http://bit.ly/1a1kDj8*) is a good place to start if you're looking for a solid introduction to Python as a programming language. This book's Python source code is written in Python 2.7, the latest release of the 2.x line. (Although perhaps not entirely trivial, it's not too difficult to imagine using some of the automated tools to up-convert it to Python 3 for anyone who is interested in helping to make that happen.)

IPython Notebook is great, but if you're new to the Python programming world, advising you to just follow the instructions online to configure your development environment would be a bit counterproductive (and possibly even rude). To make your experience with this book as enjoyable as possible, a turnkey virtual machine is available that has IPython Notebook and all of the other dependencies that you'll need to follow along with the examples from this book preinstalled and ready to go. All that you have to do is follow a few simple steps, and in about 15 minutes, you'll be off to the races. If you have a programming background, you'll be able to configure your own development environment, but my hope is that I'll convince you that the virtual machine experience is a better starting point.

See Appendix A for more detailed information on the virtual machine experience for this book. Appendix C is also worth your attention: it presents some IPython Notebook tips and common Python programming idioms that are used throughout this book's source code.

Whether you're a Python novice or a guru, the book's latest bug-fixed source code and accompanying scripts for building the virtual machine are available on GitHub (*http://bit.ly/1a1kFHM*), a social Git (*http://bit.ly/16mhOep*) repository that will always reflect the most up-to-date example code available. The hope is that social coding will enhance collaboration between like-minded folks who want to work together to extend the examples and hack away at fascinating problems. Hopefully, you'll fork, extend, and improve the source—and maybe even make some new friends or acquaintances along the way.

The official GitHub repository containing the latest and greatest bug-fixed source code for this book is available at *http://bit.ly/MiningTheSocialWeb2E*.

Improvements Specific to the Second Edition

When I began working on this second edition of *Mining the Social Web*, I don't think I quite realized what I was getting myself into. What started out as a "substantial update" is now what I'd consider almost a rewrite of the first edition. I've extensively updated each chapter, I've strategically added new content, and I really do believe that this second edition is superior to the first in almost every way. My earnest hope is that it's going to be able to reach a much wider audience than the first edition and invigorate a broad community of interest with tools, techniques, and practical advice to implement ideas that depend on munging and analyzing data from social websites. If I am successful in this endeavor, we'll see a broader awareness of what it is possible to do with data from social websites and more budding entrepreneurs and enthusiastic hobbyists putting social web data to work.

A book is a product, and first editions of any product can be vastly improved upon, aren't always what customers ideally would have wanted, and can have great potential if appropriate feedback is humbly accepted and adjustments are made. This book is no exception, and the feedback and learning experience from interacting with readers and consumers of this book's sample code over the past few years have been incredibly important in shaping this book to be far beyond anything I could have designed if left to my own devices. I've incorporated as much of that feedback as possible, and it mostly boils down to the theme of *simplifying the learning experience for readers*.

Simplification presents itself in this second edition in a variety of ways. Perhaps most notably, one of the biggest differences between this book and the previous edition is that the technology toolchain is vastly simplified, and I've employed configuration management by way of an amazing virtualization technology called Vagrant (*http://bit.ly/1a1kGeH*). The previous edition involved a variety of databases for storage, various visualization toolkits, and assumed that readers could just figure out most of the installation and configuration by reading the online instructions.

This edition, on the other hand, goes to great lengths to introduce as few disparate technology dependencies as possible and presents them all with a virtual machine experience that abstracts away the complexities of software installation and configuration, which are sometimes considerably more challenging than they might initially seem. From a certain vantage point, the core toolbox is just IPython Notebook and some third-party package dependencies (all of which are versioned so that updates to open source software don't cause code breakage) that come preinstalled on a virtual machine. Inline visualizations are even baked into the IPython Notebooks, rendering from within IPython Notebook itself, and are consolidated down to a single JavaScript toolkit (D3.js (*http://d3js.org*)) that maintains visually consistent aesthetics across the chapters.

Continuing with the theme of simplification, spending less time introducing disparate technology in the book affords the opportunity to spend more time engaging in fundamental exercises in analysis. One of the recurring critiques from readers of the first edition's content was that more time should have been spent analyzing and discussing the implications of the exercises (a fair criticism indeed). My hope is that this second edition delivers on that wonderful suggestion by augmenting existing content with additional explanations in some of the void that was left behind. In a sense, this second edition does "more with less," and it delivers significantly more value to you as the reader because of it.

In terms of structural reorganization, you may notice that a chapter on GitHub has been added to this second edition. GitHub is interesting for a variety of reasons, and as you'll observe from reviewing the chapter, it's not all just about "social coding" (although that's a big part of it). GitHub is a very social website that spans international boundaries, is rapidly becoming a general purpose collaboration hub that extends beyond coding, and can fairly be interpreted as an interest graph—a graph that connects people and the things that interest them. Interest graphs, whether derived from GitHub or elsewhere, are a very important concept in the unfolding saga that is the Web, and as someone interested in the social web, you won't want to overlook them.

In addition to a new chapter on GitHub, the two "advanced" chapters on Twitter from the first edition have been refactored and expanded into a collection of more easily adaptable Twitter recipes that are organized into Chapter 9. Whereas the opening chapter of the book starts off slowly and warms you up to the notion of social web APIs and data mining, the final chapter of the book comes back full circle with a battery of diverse building blocks that you can adapt and assemble in various ways to achieve a truly enormous set of possibilities. Finally, the chapter that was previously dedicated to microformats has been folded into what is now Chapter 8, which is designed to be more of a forward-looking kind of cocktail discussion about the "semantically marked-up web" than an extensive collection of programming exercises, like the chapters before it.

Constructive feedback is always welcome, and I'd enjoy hearing from you by way of a book review, tweet to @SocialWebMining (*http://bit.ly/1a1kHzq*), or comment on *Mining the Social Web*'s Facebook wall (*http://on.fb.me/1a1kHPQ*). The book's official website and blog that extends the book with longer-form content is at *http://MiningTheSocialWeb.com*.

Conventions Used in This Book

This book is *extensively* hyperlinked, which makes it ideal to read in an electronic format such as a DRM-free PDF that can be purchased directly from O'Reilly as an ebook. Purchasing it as an ebook through O'Reilly also guarantees that you will get automatic updates for the book as they become available. The links have been shortened using the *bit.ly* service for the benefit of customers with the printed version of the book. All hyperlinks have been vetted.

The following typographical conventions are used in this book:

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

`Constant width`
: Indicates program listings, and is used 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. Also occasionally used for emphasis in code listings.

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

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

This icon indicates a warning or caution.

Using Code Examples

The latest sample code for this book is maintained on GitHub at *http://bit.ly/MiningTheSocialWeb2E*, the official code repository for the book. You are encouraged to monitor this repository for the latest bug-fixed code as well as extended examples by the author and the rest of the social coding community. If you are reading a paper copy of this book, there is a possibility that the code examples in print may not be up to date, but so long as you are working from the book's GitHub repository, you will always have the latest bug-fixed example code. If you are taking advantage of this book's virtual machine experience, you'll already have the latest source code, but if you are opting to work on your own development environment, be sure to take advantage of the ability to download a source code archive directly from the GitHub repository.

Please log issues involving example code to the GitHub repository's issue tracker as opposed to the O'Reilly catalog's errata tracker. As issues are resolved in the source code at GitHub, updates are published back to the book's manuscript, which is then periodically provided to readers as an ebook update.

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 require attribution according to the OSS license under which the code is released. An attribution usually includes the title, author, publisher, and ISBN. For example: "*Mining the Social Web, 2nd Edition*, by Matthew A. Russell. Copyright 2014 Matthew A. Russell, 978-1-449-36761-9."

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 (*www.safaribooksonline.com*) is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of product mixes and pricing programs for organizations, government agencies, and individuals. Subscribers have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and dozens more. For more information about Safari Books Online, please visit us online.

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 *non-code-related errata* and additional information. You can access this page at:

http://bit.ly/mining_social_web_2e

Any errata related to the sample code should be submitted as a ticket through GitHub's issue tracker at:

http://github.com/ptwobrussell/Mining-the-Social-Web/issues

Readers can request general help from the author and publisher through GetSatisfaction at:

http://getsatisfaction.com/oreilly

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

bookquestions@oreilly.com

For more information about our books, conferences, Resource Centers, and the O'Reilly Network, see our website at:

http://www.oreilly.com

Acknowledgments for the Second Edition

I'll reiterate from my acknowledgments for the first edition that writing a book is a tremendous sacrifice. The time that you spend away from friends and family (which happens mostly during an extended period on nights and weekends) is quite costly and can't be recovered, and you really do need a certain amount of moral support to make it through to the other side with relationships intact. Thanks again to my very patient friends and family, who really shouldn't have tolerated me writing another book and probably think that I have some kind of chronic disorder that involves a strange addic-