GABOR SZABO
GUNGOR POLATKAN
OSCAR BOYKIN
ANTONIOS CHALKIOPOULOS

SOCIAL MEDIA DATA MINING AND ANALYTICS

WILEY

Social Media Data Mining and Analytics

Gabor Szabo
Gungor Polatkan
Oscar Boykin
Antonios Chalkiopoulos

WILEY

Social Media Data Mining and Analytics

Published by
John Wiley & Sons, Inc.
10475 Crosspoint Boulevard
Indianapolis, IN 46256
www.wiley.com

Copyright © 2019 by John Wiley & Sons, Inc., Indianapolis, Indiana Published simultaneously in Canada

ISBN: 978-1-118-82485-6
ISBN: 978-1-118-82490-0 (ebk)
ISBN: 978-1-118-82489-4 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permissions.

Limit of Liability/Disclaimer of Warranty: The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that Internet websites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at http://booksupport.wiley.com. For more information about Wiley products, visit www.wiley.com.

Library of Congress Control Number: 2018956702

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

About the Authors

Gabor Szabo works on large-scale data analysis and modeling problems in social networks, self-organized online ecosystems, transportation systems, and autonomous driving. Previously, his research focus was on the description of randomly organized networks in online communities and biological systems at Harvard Medical School, the University of Notre Dame, and HP Labs. After that he built distributed algorithms to understand and predict user behavior at Twitter. He has created models for resource allocation in Lyft's ride-sharing network, and most recently he led a team at Tesla's Autopilot.

Gungor Polatkan is a machine learning expert and engineering leader with experience in building massive-scale distributed data pipelines serving personalized content at LinkedIn and Twitter. Most recently, he led the design and implementation of the AI backend for LinkedIn Learning and ramped the recommendation engine from scratch to hyper-personalized models learning billions of coefficients for 500M+ users. He deployed some of the first deep ranking models for search verticals at LinkedIn improving Talent Search. He enjoys leading teams, mentoring engineers, and fostering a culture of technical rigor and craftsmanship while iterating fast. He has worked in several notable applied research groups in Twitter, Princeton, Google, MERL and UC Berkeley before joining LinkedIn. He published and refereed papers at top-tier ML & AI venues such as UAI, ICML, and PAMI.

Oscar Boykin works on machine learning infrastructure at Stripe, building systems to predict fraud at scale. Prior to Stripe, Oscar spent more than 4 years at Twitter, first working on modeling and prediction for ads, and later on data infrastructure systems. At Twitter, Oscar co-developed many open-source scala libraries including Scalding, Algebird, Summingbird, and Chill. Before

Twitter, Oscar was an assistant professor of electrical and computer engineering at the University of Florida. Oscar has a Ph.D. in physics from the University of California, Los Angeles and is the coauthor of dozens of academic papers in top journals and conferences.

Antonios Chalkiopoulos is a fast/big data distributed system specialist with experience in delivering production-grade data pipelines in the media, IoT, retail, and finance industries. Antonios is a published author in big data, an open source contributor, and the co-founder and CEO of Landoop LTD. Landoop LTD builds the innovative and award winning Lenses platform for data in motion, which provides visibility and control over streaming data, data discovery via an intuitive web interface, and is a comprehensive SQL experience for data in motion, monitoring, alerting, data governance, multi-tenancy, and security. Lenses is a complete user experience for building and managing real-time data pipelines and micro-services.

About the Technical Editors

Sriram Krishnan is a senior director of the Einstein Platform team at Salesforce, where he is responsible for the foundational services that bring machine learning capabilities to Salesforce. Prior to Salesforce, Sriram was head of the Data Platform team at Twitter, and a tech lead on the Big Data Platform team at Twitter. He holds a Ph.D. in Computer Science from Indiana University, and spent several years as a researcher and group lead at the San Diego Supercomputer Center enabling scientific applications to use grid and cloud technologies. Sriram has co-authored more than 50 publications in the area of data, grid, and cloud computing, and his work has been cited more than 1700 times. Sriram has contributed to several influential open source projects that are being used widely in industry and academia.

Ben Peirce is director of XR Analytics at Samsung, which he joined on the acquisition of Vrtigo, a virtual reality analytics startup he co-founded. Previously, Ben built analytics systems at early stage startups in healthcare and advertising technology for over a decade. He holds a Ph.D. from Harvard, where he studied control systems and robotics.

Dashun Wang is an associate professor of management and organizations at the Kellogg School of Management, (by courtesy) industrial engineering and management sciences at the McCormick School of Engineering, and a core faculty at NICO, the Northwestern Institute on Complex Systems. Dashun received his Ph.D. in physics in 2013 from Northeastern University, where he was a member of the Center for Complex Network Research. From 2009 to 2013, he had also held an affiliation with Dana-Farber Cancer Institute, Harvard University as a research associate. He is a recipient of the AFOSR Young Investigator Award (2016).

Dr. Jian Wu is an assistant professor in the Department of Computer Science at the Old Dominion University. Dr. Wu obtained his Ph.D. in 2011 from Pennsylvania State University and then worked with Dr. C. Lee Giles on the CiteSeerX project as a tech leader. Dr. Wu's research interest is text mining and knowledge extraction on scholarly big data using machine learning, deep learning, and natural language processing. He has published nearly 30 peer-reviewed papers in ACM, IEEE, and AAAI conferences and magazines with best papers and nominations. He was the best reviewer in the ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2018. As a tech leader, Dr. Wu made critical improvements to the architecture, web crawling, and extraction modules of CiteSeerX, increasing the collection to 10 million by 2017.

Credits

Project Editor
Tom Dinse

Technical Editors
Sriram Krishnan
Ben Peirce
Dashun Wang
Dr. Jian Wu

Production Editor
Athiyappan Lalith Kumar

Copy Editor
San Dee Phillips

Production Manager
Kathleen Wisor

Content Enablement and Operations Manager
Pete Gaughan

Marketing Manager
Christie Hilbrich

Associate Publisher
Jim Minatel

Project Coordinator, Cover
Brent Savage

Proofreader
Evelyn Wellborn

Indexer
Johnna VanHoose Dinse

Cover Designer
Wiley

To our families who supported us even though
we missed a lot of time from them to write this book.

Acknowledgments

We would like to send our gratitude to our friends and colleagues at Twitter. With them invaluable discussions and collaborations have opened new perspectives for us to be able to look at social media data in unexpected ways, and allowed us to work on tools and approaches that let us expand our understanding of social media users. Their open-minded support throughout has always been greatly appreciated.

A very special thank you to Prof. David Blei, who provided the innovative research on topic modeling and a proper methodology for teaching machine learning through his Princeton class "Interacting with Data." In this book we followed his examples to cover the topics on representation learning and the applications in recommendations problems.

We would like to thank Jonathan Chang, the author of the R LDA package, for providing a machine learning tool for efficient and easy-to-use topic modeling techniques.

We would also like to thank Tom Dinse, Robert Elliott, and Jim Minatel, our editors at Wiley, who have been leading us down the path of publishing this book since the beginning for their great project management and editorial review of the content, as well as our team of technical editors for their review and insightful suggestions throughout the process. Moreover we would like to thank all the people who worked behind the scenes to help get this book together.

As for the rest of the authors, we would like to thank one of us, Gabor Szabo, who patiently shepherded the entire book writing process while we were working on it.

Contents

Introduction

This book is about using data to understand how social media services are used. Since the advent of Web 2.0, sites and services that give their users the power to actively change and contribute to the services' content have exploded in popularity. Social media finds its roots in early social networking and community communication services, including the bulletin board systems (BBS) of the 1980s, then the Usenet newsgroups, and Geocities in the '90s, whose communities organized around topical interests and provided their users with either email or chat room communications. The worldwide information communication network known as the Internet gave rise to a higher-level networking: a global web of connections among like-minded individuals and groups. Although the basic idea of connecting people across the globe has changed little since then, the scope and influence of social media services have attained never-before seen proportions. Although it's natural that a large part of the conversation is still happening in the "real world," the shift toward electronic information exchange on the level of human interactions has been getting stronger. The proliferation of mobile devices and connectivity puts the "Internet in our pockets," and with it the possibility to get in touch with our friends, families, and preferred businesses, anytime, anywhere.

No wonder that a myriad of services has popped up and started serving our needs for communication and sharing, which led to a transformation of public and private life. Through these services, we can immediately know what others think about politics, brands, products, and each other. By sharing their ideas privately or anonymously, people have the choice to speak their minds more freely than they would in traditional media. Everybody can be heard if they choose, so it's also become the responsibility of these services to find the needle in the haystack of people's contributions, so to speak, in delivering relevant and interesting content to us.

What's common to all these services? They are dependent on us, as they're only the mediators between humans. This means that in a way the mathematical regularities that we may discover through analyzing their usage data reflect our own behavior, so we can expect to see similar insights and challenges when we work with these datasets. The purpose of this book is to highlight these regularities and the technical approaches that lead to an understanding of how users of these services attend to them through the lens of the data these services collect.

Human Interactions Measured

Social media, as its name suggests, is driven by social interactions around the content that the online service provides. Social networking, for instance, makes it easy for individuals to connect with each other and share pictures and multimedia, news articles, Web content, and various other bits of information. In the most common usage scenario of these services, people go to Facebook to get updates about their friends, relatives, and acquaintances, and to share something about their lives with them. For example, on Twitter, because the follow relationship doesn't have to be reciprocated, users can learn about what any other user thinks, shares, or communicates with others. With LinkedIn, a professional social network, the goal is to connect like-minded professionals to each other through its network and its groups, and to serve as an interface between job seekers and companies looking to hire.

There are other social media services where the networking aspect of social interactions is used more as a facilitator rather than an end to co-create or enjoy shared content (for instance on Wikipedia, YouTube, or Instagram). Although the connections among users can be present, their purpose there is to make content discovery manageable for the users and to make the creation of content—for instance, Wikipedia articles—more efficient.

Of course, there are many other social media sites and services, usually targeting a specific interest or domain (art, music, photography, academic institutions, geographical locations, religions, hobbies, and the list could go on), which just shows that online users have the deepest desire to connect to people based on their shared interests or commonalities.

One thing among these services, their vastly different areas of focus notwithstanding, is common: They exist only because their *users* and *audience* are there. This is what makes them different from "pre-created" or static Internet locations such as traditional media news sites, company home pages, directories, and just about any Web resource that is created centrally by a relatively small group of authorized content creators ("small" at least in comparison to the crowds of people that use social media services with numbers generally in the millions). The result of the collective dynamics of these millions of social media users is what we can observe when we dig deep into the usage patterns of these services, and this is what we're interested in understanding in this book.

Online Behavior Through Data Collection

When we collect usage log data from social media services, we have a glimpse into the statistical behavior of many human beings coming together who have similar motivations or expectations or act toward the same goal. Naturally, the way the given service is organized and how it highlights its content has a great influence on what we'll see in the logs about the users' activities. The access and usage logs are stored in the databases of the service, and, therefore, the statistical patterns by which we all interact with others and the content the service hosts are bound to show up in these traces. (Provided there *are* such patterns, and we don't just carry out our daily activities in a completely inconsistent and random way! We'll see that—as perhaps expected by common sense—statistical regularities are abundant everywhere.)

Fortunately, the services (in most cases) don't differ so radically from each other in their designs that they would give rise to completely different user behavior characteristics. What do we mean by this? Let's say, for example, that we want to measure a simple thing: how frequently users come back to our service within a week and take part in some activity. This would be just a number, ranging from 0 to (in theory) infinity, for every user. Of course, we won't see anyone undertake an infinite number of actions on our service within a limited amount of time, but it may still be a large number. So, having set our mind to measuring the number of activities, can we expect to have different *statistical* results for two different systems: users posting videos to their YouTube channels and users uploading photos to their Flickr accounts?

The answer, obviously, is a resounding yes. If we looked at the distributions of the number of times people used either YouTube or Flickr, respectively, we would, of course, see that the fraction of YouTube users who upload *one* video per week *is* different from the fraction of Flickr users who upload *one* image per week. This is natural, as these two different services attract different demographics with different usage scenarios, so the exact distributions, consequently, will be different. However, what is not perhaps straightforward is that in most online systems that researchers have looked at we find *a similar qualitative statistical behavior* for these distributions.

By "qualitative" we mean that although the exact parameters of the usage model may be different for the two respective services, the model itself, through which we can best describe user behavior in both systems, is still the same or very similar between the services (with perhaps slight variations).

The good news about this is that we can be reasonably confident that what we're measuring with the data in the activity logs is indeed the underlying human behavior that drives the content creation, diffusion, sharing, and more, on these sites. The other piece of good news is that from this we can extrapolate and if we encounter a new service operating on user-generated content, we can make educated guesses about what we can measure in it. Therefore, if we see something unexpected in the graphs that's different from the general pattern we have seen before, we should look for a service-specific reason for it that we can be inclined to explore further.

So, in a way, the methods and the results that we highlight in this book may well apply to a completely new service if it's also governed by the same underlying human behavior. With few exceptions, this is true of the social media services for which we're aware research exists, and therefore we like to think of these systems as providing insight into human behavior. The opportunity, then, to observe and describe many people acting loosely together is unprecedented; this is because of the digital footprints they leave behind in the services' logs. (Privacy issues are, of course, a valid practical concern, but here we're interested only in the large picture and not how specific individuals behave.) The next sections look at what kinds of data can be of interest in various social media services and which public datasets we'll be using for examples in this book.

What Types of Data Are Essential to Collect?

The questions we would like to ultimately answer with data determine the types of data you need to collect, but in general, the more data you have at your disposal, the better you can answer those and future questions as well. You never know when you want to refine or expand the data analysis, so if you design a service, it's better to think ahead and log all or almost all the interactions users have with the service and each other. These days, storage is inexpensive, so it's wise to cater to as many future data needs as possible by not trying to optimize too early for storage space. Naturally, as the service evolves and it becomes clear what the focus areas are, it's possible to trim the data collection back and refactor the existing data sources, if necessary.

To better understand the user activity data we generally require, let's look at some typical questions around social media usage that we could be interested in answering:

- Who are the most active/inactive users? How many of them do we have?
- How does usage evolve over time? Can we predict usage per user segment (by geography, demographics, type of usage) ahead of time?
- How do we match users to content? Users to users? How do we surface content of interest to the user in a timely manner?
- What do users' networks look like? Do more engaged users form different kinds of networks?

- Why do people leave the service, if they do (*churn*)? Are there precursors to this churn, and can we predict it?
- What brings new users to the service to join? Do they like it, and if not, what makes happy users different from dissatisfied ones?
- Are there users who exploit our service in any way? Is there any spamming, unscrupulous usage, and deceptive behavior going on among the users?
- What are the most "interesting" or "trending" pieces of content at any given time? Who are attending to it from among our users, how can we find it, and what is it about?
- Can we find specific content of interest to us among the sheer amounts of streaming or historical data that the users produce? For instance, can we find users who mentioned a specific word or subject recently?
- What pieces of content are "popular" among the users? Are there big differences among their popularities, and if so, how big?

The chapters in this book address some of these questions and offer answers for specific services. As may be apparent, some of these can be best answered by doing active experiments with our users, in particular A/B testing experiments. (In an A/B testing experiment we show one feature or use one algorithm for one set of users *A*, and another for another set of users *B*. By measuring the differences in user activities between the *A* and *B* groups we can decide what influence the change in the feature had on users.) However, because we focus more on analyzing data that has been collected previously and learning as much about it as possible, we won't cover this powerful technique generally used to optimize the user experience on the service.

What kind of data should we collect either from the service we run or from other social media services we have access to, then? Guided by the previous questions, a few aspects of log data should be required for our analysis:

1. As users come to our service, they carry out specific actions: reading articles, viewing pictures, tagging photos, and sharing status updates. The (anonymized) identity of the users is what we want to know when we ask ourselves about what they are doing, along with a description of the actions.
2. We also need to know *when* they are taking the actions. Sub-second resolution for data collection (milli- or microseconds) usually suffices.
3. Obviously, for each action there could be a multitude of different kinds of metadata pieces that go along with it. If, for instance, the user favors or likes a post, we obviously want to store the unique identifier of that post together with the action.

As any of the users may have many actions over a period, the raw data logged in such a way may ultimately take a large amount of backend storage to save. This could take a long time to process for even simple questions; also, we don't always need *all* the information for the most common questions. Therefore, we normally create snapshots of *aggregated* data through automated ETL (extract, transform, load) processes in a production environment, for instance about the current state of the social graph with all the relationships among the users, the number of Tweets, posts, and photos that they have created or shared, and so on. When we want to analyze the data to gain certain insights, these aggregations are frequently the first source of information to turn to.

Although we need to think about how to best store all this data in appropriate databases, the design and implementation of such schemas is a science and is beyond the scope of this book. Also, we would like to rather focus on the way insights can be derived from the data and will use publicly available data from social media services to illustrate how we proceed with the different types of analyses.

Asking and Answering Questions with Data

Our goal is to expose you to several common situations you will encounter while making sense of data generated by social media services. The usual way of studying empirical phenomena (not necessarily just related to social media) has been following the centuries-long tradition of the scientific method:

1. Asking the question comes first, in generic terms. This doesn't yet have to involve any further assumptions about the data; we're just formalizing what we'd like to know about a specific behavior. For instance, "What are the temporal dynamics of users coming back to the service so that we can predict how long their session on the service will last?"
2. Optionally formulate a hypothesis about the expected outcome. This is useful for verifying whether your preconceptions make sense. Also, if you have a model in mind that you think best describes the quantitative outcome, you can check this. After you have formulated a hypothesis, predict what the result should be if the hypothesis holds. This step is optional because, if you don't want to build a model around the question and your goal is to use the result only to gain insights, you can skip this step. A hypothesis to the question in step 1, for instance, can be that "users come back to the service in a random manner, independently of whether they used it recently." (Whether this actual hypothesis is true in real services, you'll see later in Chapter 3.)
3. Determine the procedure to follow and what input data to collect to answer the question asked in step 1. Although the procedure is usually

straightforward given the computational tools and existing techniques you have, you usually have a lot of freedom in social media to select the test data set. Do you want to take samples from among the users or use everyone? What date range will you use? Do you filter out certain actions you consider undesirable? You obviously want to be thorough and explore as much about the data as possible to gain confidence about the results, for instance by taking different periods for the dataset or looking at different user cohorts. For the question you want to answer (see step 1), you may want to take the timestamps of any action generated by the users for a given month, for instance, and then take time differences between subsequent timestamps and analyze their temporal correlations.

4. Perform the data analysis! Ideally, the data collection has been already done by you or for you so that you don't have to wait for that. If your goal is to test your hypothesis, you also want to perform statistical testing. If you just want to gain insights, your numerical results are the answer to the question you asked.

The Datasets Used in This Book

To elucidate the processes and regularities that you can observe in social media due to human interactions, you naturally want to use some existing data coming from such systems, downloadable from various places on the Internet. Although most of the social media services keep their data private (privacy concerns being the paramount reason but also because these datasets can become huge), some services, most notably Wikipedia, make *all* their data available to the public. In other cases, academic researchers have collected data from these services through crawling or data sharing. The following sections list the data sources that we used throughout the book. We encourage you to try (and expand on) the examples for which having these datasets at hand is a prerequisite.

We selected a few services that have public, widely available, and easily obtainable datasets about their users and their content, to show what results we can expect in actual social media services for the questions we'll be asking. The names of these services should be familiar, and we also wanted to ensure that the datasets are at least medium-sized for users and the time range they span, and thus are amenable to analysis to draw meaningful conclusions. Follow the practical examples showcased throughout the book; to this end, the following sections describe the datasets used. As a summary, Table I.1 provides short descriptions of the example datasets we use.

Table I.1: Descriptions and Locations of the Datasets Used in This Book

SERVICE	MAIN PAGE	DATASET
Wikipedia	wikipedia.org	Revision and page meta information, no actual text
Twitter	twitter.com	Tweets created
Stack Exchange	scifi.stackexchange.com	Questions and answers from Stack Exchange's Science Fiction & Fantasy category
LiveJournal	livejournal.com	Directed social network connections
Cora dataset		Scientific documents from an academic search engine
MovieLens	movielens.org	Sample of movie ratings
Amazon Fine Food Reviews		Historical reviews on Amazon for "Fine Foods"

NOTE **Wikipedia and Stack Exchange content are licensed under the Creative Commons Attribution-ShareAlike 3.0 License,** `https://creativecommons.org/licenses/by-sa/3.0/`**; Livejournal data collected are due to Mislove et al., "Measurement and Analysis of Online Social Networks," IMC 2007,** `http://social-networks.mpi-sws.org/data-imc2007.html`**; the MovieLens dataset is from GroupLens Research,** `http://grouplens.org/datasets/movielens/`**; and Cora appeared in McCallum et al., "Automating the Construction of Internet Portals with Machine Learning," Information Retrieval vol 3, issue 2, 2000.**

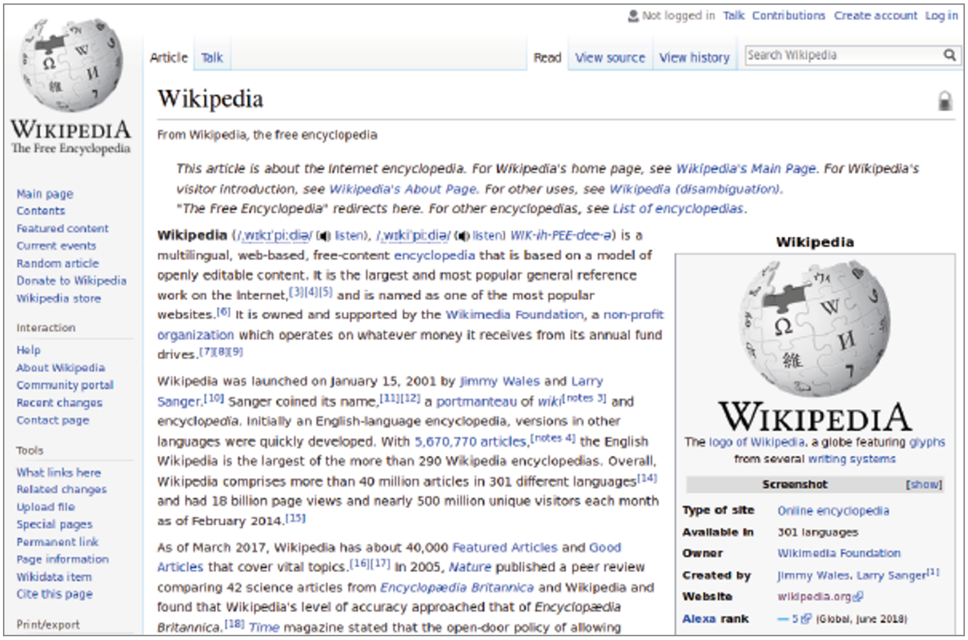
We made it easier for you to obtain these datasets: run `data/download_all.sh`, available from the book's downloads to get all the data files that the examples build on. (Note that due to the large size of the datasets, especially the Wikipedia dataset, at 50-60 GB the downloads take some time to complete). The location of the source code is given at the end of this Introduction.

Wikipedia

The biggest dataset we use is the English-language Wikipedia's revision histories of the several million articles it hosts. Wikipedia is a collaboratively edited encyclopedia, and the English version has approximately 5.7 million articles in 2018, with approximately 300,000 monthly active editors (`http://en.wikipedia.org/wiki/Wikipedia:Statistics`). A screen shot of the article "Wikipedia" can be seen in Figure I.1.

Twitter

On Twitter (Figure I.2), users can send out status updates of at most 140 characters in length (until 2017, when the service increased the maximum length of updates). Other users, who "follow" the sender, will receive these short messages



Wikipedia

From Wikipedia, the free encyclopedia

This article is about the Internet encyclopedia. For Wikipedia's home page, see Wikipedia's Main Page. For Wikipedia's visitor introduction, see Wikipedia's About Page. For other uses, see Wikipedia (disambiguation).
"The Free Encyclopedia" redirects here. For other encyclopedias, see List of encyclopedias.

Wikipedia (/ˌwɪkɪˈpiːdiə/ (listen), /ˌwɪkiˈpiːdiə/ (listen) *WIK-ih-PEE-dee-ə*) is a multilingual, web-based, free-content encyclopedia that is based on a model of openly editable content. It is the largest and most popular general reference work on the Internet,[3][4][5] and is named as one of the most popular websites.[6] It is owned and supported by the Wikimedia Foundation, a non-profit organization which operates on whatever money it receives from its annual fund drives.[7][8][9]

Wikipedia was launched on January 15, 2001 by Jimmy Wales and Larry Sanger.[10] Sanger coined its name,[11][12] a portmanteau of *wiki*[notes 3] and *encyclopedia*. Initially an English-language encyclopedia, versions in other languages were quickly developed. With 5,670,770 articles,[notes 4] the English Wikipedia is the largest of the more than 290 Wikipedia encyclopedias. Overall, Wikipedia comprises more than 40 million articles in 301 different languages[14] and had 18 billion page views and nearly 500 million unique visitors each month as of February 2014.[15]

As of March 2017, Wikipedia has about 40,000 Featured Articles and Good Articles that cover vital topics.[16][17] In 2005, *Nature* published a peer review comparing 42 science articles from *Encyclopædia Britannica* and Wikipedia and found that Wikipedia's level of accuracy approached that of *Encyclopædia Britannica*.[18] *Time* magazine stated that the open-door policy of allowing

Wikipedia

The logo of Wikipedia, a globe featuring glyphs from several writing systems

Screenshot [show]

Type of site	Online encyclopedia
Available in	301 languages
Owner	Wikimedia Foundation
Created by	Jimmy Wales, Larry Sanger[1]
Website	wikipedia.org
Alexa rank	— 5 (Global, June 2018)

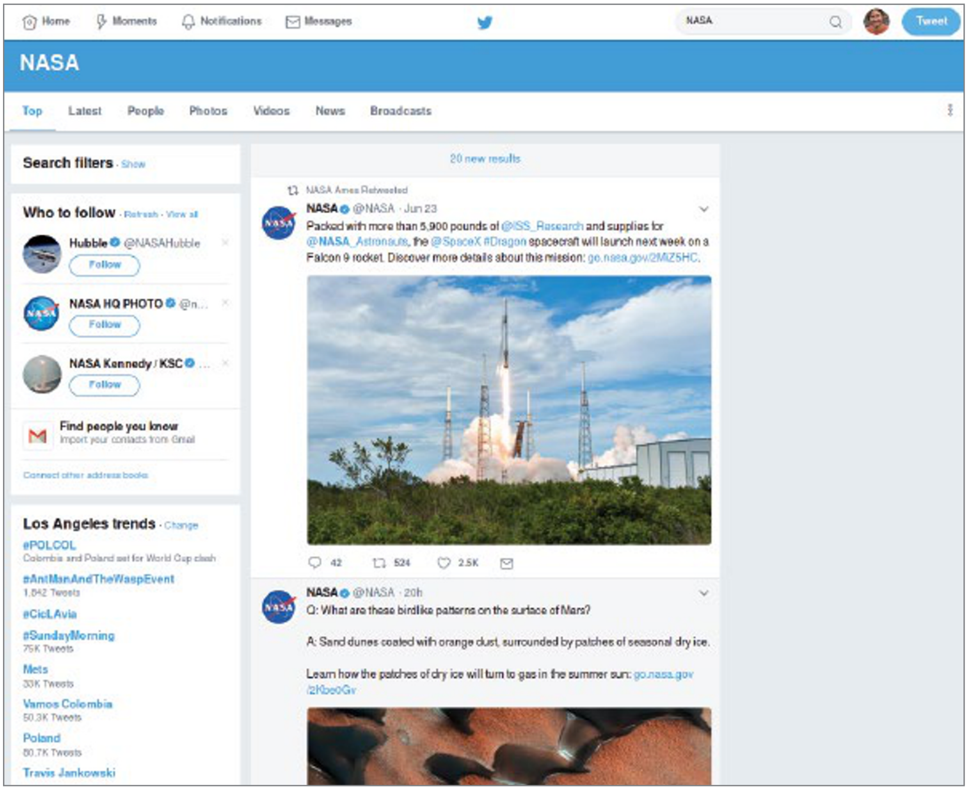
Figure I.1: An entry from the online encyclopedia Wikipedia about Wikipedia

in their so-called timeline. Pictures and short videos can also be attached to the status update. Many users follow news sources, celebrities, or their friends and family. Often, Twitter is considered an "information network" where users can follow anyone who they're interested in getting updates from, and those users do not have to follow them back.

We will collect Tweets using Twitter's API to analyze the activity of a sample of the users in Chapter 1.

Stack Exchange

Stack Exchange (Figure I.3) is a federated network of websites following the model of question answering, where users ask a question on a variety of different topics, and other users can answer these questions and vote both on questions and answers. This way high-quality content (at least in the eyes of the users) rises to the top. As of 2018, the Stack Exchange network consists of more than 350 sites covering different topics from software programming to astronomy to poker. The most well-known of these sites is the one that the network started with in 2008, Stack Overflow, focusing on various topics in computer programming. In Chapter 4, we take one of the topical Stack Exchange sites, the Science Fiction & Fantasy category, and look at the various properties of the posts that users submit there.



NASA @NASA · Jun 23
Packed with more than 5,900 pounds of @ISS_Research and supplies for @NASA_Astronauts, the @SpaceX #Dragon spacecraft will launch next week on a Falcon 9 rocket. Discover more details about this mission: go.nasa.gov/2MiZ5HC.

NASA @NASA · 20h
Q: What are these birdlike patterns on the surface of Mars?

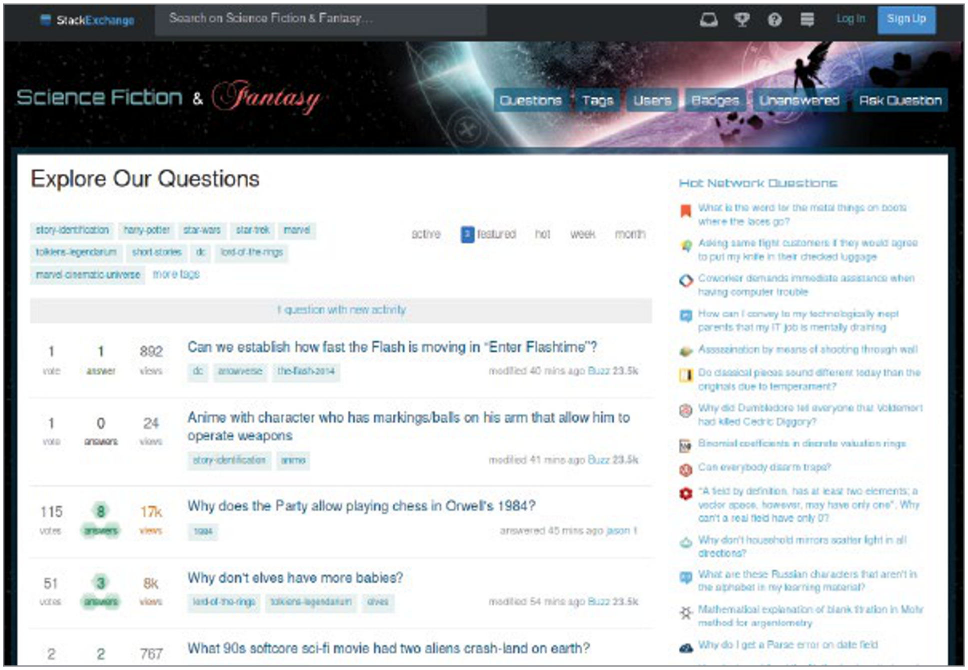
A: Sand dunes coated with orange dust, surrounded by patches of seasonal dry ice.

Learn how the patches of dry ice will turn to gas in the summer sun: go.nasa.gov/2KbeOGv

Figure I.2: A screen shot of a typical Twitter search timeline. Tweets appear in the main section, whereas trending topics and "who to follow" recommendations are shown on the side.

LiveJournal

LiveJournal (Figure I.4) is an online journal keeping and blogging service, in which users can make either mutual or unilateral connections to other users. Friends of users can read their protected entries, and conversely, the blog posts of friends show up on their "friends page." We'll use this dataset to study the directed connection structure of a social network in Chapter 2.



Explore Our Questions

1 question with new activity

1 vote 1 answer 892 views — Can we establish how fast the Flash is moving in "Enter Flashtime"? — dc arrowverse the-flash-2014 — modified 40 mins ago Buzz 23.5k

1 vote 0 answers 24 views — Anime with character who has markings/balls on his arm that allow him to operate weapons — story-identification anime — modified 41 mins ago Buzz 23.5k

115 votes 8 answers 17k views — Why does the Party allow playing chess in Orwell's 1984? — 1984 — answered 45 mins ago jason 1

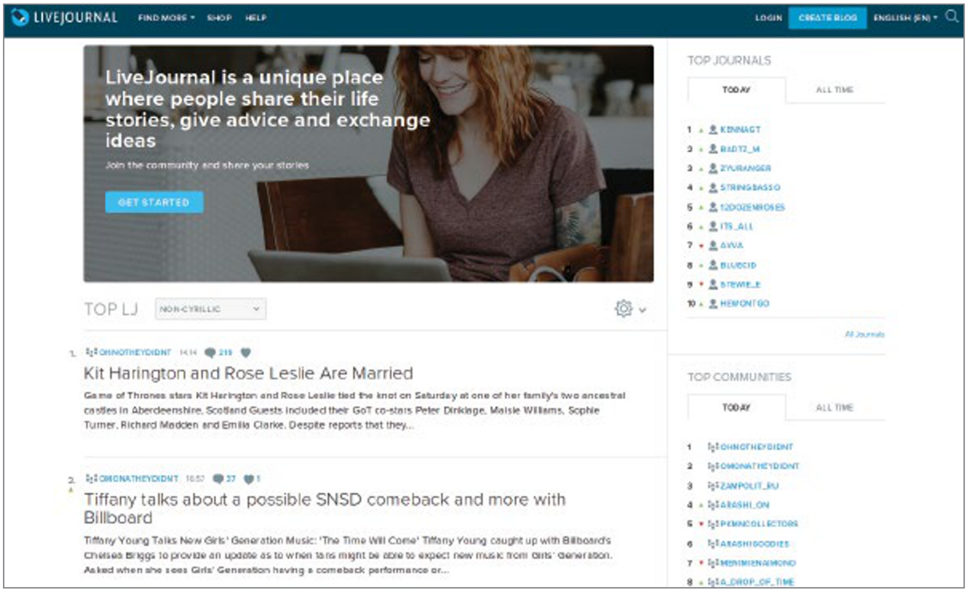
51 votes 3 answers 8k views — Why don't elves have more babies? — lord-of-the-rings tolkiens-legendarium elves — modified 54 mins ago Buzz 23.5k

2 2 767 — What 90s softcore sci-fi movie had two aliens crash-land on earth?

Hot Network Questions

- What is the word for the metal things on boots where the laces go?
- Asking same flight customers if they would agree to put my knife in their checked luggage
- Coworker demands immediate assistance when having computer trouble
- How can I convey to my technologically inept parents that my IT job is mentally draining
- Assassination by means of shooting through wall
- Do classical pieces sound different today than the originals due to temperament?
- Why did Dumbledore tell everyone that Voldemort had killed Cedric Diggory?
- Binomial coefficients in discrete valuation rings
- Can everybody disarm traps?
- "A field by definition, has at least two elements; a vector space, however, may have only one". Why can't a real field have only 0?
- Why don't household mirrors scatter light in all directions?
- What are these Russian characters that aren't in the alphabet in my learning material?
- Mathematical explanation of blank titration in Mohr method for argentometry
- Why do I get a Parse error on date field

Figure I.3: Stack Exchange is a question answering service with a lot of topical sub-sites. We chose the Science Fiction & Fantasy category as it is not overly technical in nature (compared to computer-related categories or those focused on mathematics, for instance), yet has a decent number of users and amount of content.



LiveJournal is a unique place where people share their life stories, give advice and exchange ideas

Join the community and share your stories

1. OHNOTHEYDIDNT 14:14 219

Kit Harington and Rose Leslie Are Married

Game of Thrones stars Kit Harington and Rose Leslie tied the knot on Saturday at one of her family's two ancestral castles in Aberdeenshire, Scotland Guests included their GoT co-stars Peter Dinklage, Maisie Williams, Sophie Turner, Richard Madden and Emilia Clarke. Despite reports that they...

2. OMONATHEYDIDNT 16:57 37 1

Tiffany talks about a possible SNSD comeback and more with Billboard

Tiffany Young Talks New Girls' Generation Music: 'The Time Will Come' Tiffany Young caught up with Billboard's Chelsea Briggs to provide an update as to when fans might be able to expect new music from Girls' Generation. Asked when she sees Girls' Generation having a comeback performance or...

TOP JOURNALS

1 KENNAGT
2 RADT2_M
3 ZYURANGER
4 STRINGBASSO
5 12DOZENROSES
6 ITS_ALL
7 AVVA
8 BLUECID
9 STEWIE_E
10 HEMONTGO

TOP COMMUNITIES

1 OHNOTHEYDIDNT
2 OMONATHEYDIDNT
3 ZAMPOLIT_RU
4 ARASHI_ON
5 PKMNCOLLECTORS
6 ARASHIGOODIES
7 MENIMIENAIMONO
8 A_DROP_OF_TIME

Figure I.4: The main page of LiveJournal, a blogging platform that encourages the creation of communities as well

Scientific Documents from Cora

This is a smaller dataset, containing the texts of 2,410 scientific documents from the Cora search engine. (This search engine has been deprecated; it was a proof-of-concept search engine for academic publications in computer science.) We use this dataset to illustrate the topic modeling approach for natural language texts in Chapter 4. The dataset comes bundled with the `lda` R package; no additional download will be necessary.

Amazon Fine Food Reviews

This is a dataset containing "Fine Food" reviews from Amazon, including product review summaries, scores, and some user details. The dataset has data for 10 years, through October 2012. For more details, see `https://snap.stanford.edu/data/web-FineFoods.html`.

MovieLens Movie Ratings

This dataset contains movie ratings from the MovieLens service (`https://movielens.org/`) on a scale ranging from 1 through 5, left by 938 users on 1,682 movies. Chapter 6 uses this dataset in the examples to predict how users would likely rate a movie that they haven't seen yet, given how other users like them have rated the movie before.

The Languages and Frameworks Used in This Book

The examples in this book are predominantly written in three programming languages and frameworks: R, Python, and Scalding. We use R for its excellent capabilities in statistics, machine learning, and graphics; Python because preprocessing large datasets and interfacing with service APIs is easy and fast in this language; and Scalding because it's a flexible and robust framework for carrying out distributed computations on MapReduce.

In general, we also believe these tools are great to know for data mining; therefore, we assume that you are familiar with them, or at least can understand code written in them. They provide a rapid development path for prototyping algorithms and writing quick tests around data, and through the extensive community support available for them, answers to almost any common technical challenge are readily available on online forums.

The titles of the code examples in this book reference the example's source code file (unless the code snippet is very short). The source files are in the `src/`*`chapterX`* subfolder of the book's code repository, where *`X`* refers to the chapter where the code example appears.