

THIRD EDITION

Mining the Social Web

Matthew A. Russell and Mikhail Klassen

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY[®]

Table of Contents

Preface..... xi

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 5

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 11

1.3.3 Exploring Trending Topics 16

1.3.4 Searching for Tweets 20

1.4 Analyzing the 140 (or More) Characters 26

1.4.1 Extracting Tweet Entities 28

1.4.2 Analyzing Tweets and Tweet Entities with Frequency Analysis 30

1.4.3 Computing the Lexical Diversity of Tweets 33

1.4.4 Examining Patterns in Retweets 35

1.4.5 Visualizing Frequency Data with Histograms 37

1.5 Closing Remarks 42

1.6 Recommended Exercises 43

1.7 Online Resources 44

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

2.1 Overview 46

2.2 Exploring Facebook’s Graph API	46
2.2.1 Understanding the Graph API	48
2.2.2 Understanding the Open Graph Protocol	52
2.3 Analyzing Social Graph Connections	59
2.3.1 Analyzing Facebook Pages	63
2.3.2 Manipulating Data Using pandas	74
2.4 Closing Remarks	83
2.5 Recommended Exercises	84
2.6 Online Resources	85
3. Mining Instagram: Computer Vision, Neural Networks, Object Recognition, and Face Detection.	87
3.1 Overview	88
3.2 Exploring the Instagram API	89
3.2.1 Making Instagram API Requests	89
3.2.2 Retrieving Your Own Instagram Feed	92
3.2.3 Retrieving Media by Hashtag	93
3.3 Anatomy of an Instagram Post	94
3.4 Crash Course on Artificial Neural Networks	97
3.4.1 Training a Neural Network to “Look” at Pictures	99
3.4.2 Recognizing Handwritten Digits	100
3.4.3 Object Recognition Within Photos Using Pretrained Neural Networks	106
3.5 Applying Neural Networks to Instagram Posts	110
3.5.1 Tagging the Contents of an Image	110
3.5.2 Detecting Faces in Images	111
3.6 Closing Remarks	114
3.7 Recommended Exercises	114
3.8 Online Resources	115
4. Mining LinkedIn: Faceting Job Titles, Clustering Colleagues, and More.	117
4.1 Overview	118
4.2 Exploring the LinkedIn API	119
4.2.1 Making LinkedIn API Requests	119
4.2.2 Downloading LinkedIn Connections as a CSV File	123
4.3 Crash Course on Clustering Data	124
4.3.1 Normalizing Data to Enable Analysis	127
4.3.2 Measuring Similarity	139
4.3.3 Clustering Algorithms	141
4.4 Closing Remarks	157
4.5 Recommended Exercises	158
4.6 Online Resources	159

5. Mining Text Files: Computing Document Similarity, Extracting Collocations, and More.	161
.....	161
5.1 Overview	162
5.2 Text Files	162
5.3 A Whiz-Bang Introduction to TF-IDF	164
5.3.1 Term Frequency	165
5.3.2 Inverse Document Frequency	167
5.3.3 TF-IDF	168
5.4 Querying Human Language Data with TF-IDF	172
5.4.1 Introducing the Natural Language Toolkit	172
5.4.2 Applying TF-IDF to Human Language	175
5.4.3 Finding Similar Documents	177
5.4.4 Analyzing Bigrams in Human Language	185
5.4.5 Reflections on Analyzing Human Language Data	195
5.5 Closing Remarks	196
5.6 Recommended Exercises	197
5.7 Online Resources	198
6. Mining Web Pages: Using Natural Language Processing to Understand Human Language, Summarize Blog Posts, and More.	199
.....	199
6.1 Overview	200
6.2 Scraping, Parsing, and Crawling the Web	201
6.2.1 Breadth-First Search in Web Crawling	204
6.3 Discovering Semantics by Decoding Syntax	208
6.3.1 Natural Language Processing Illustrated Step-by-Step	210
6.3.2 Sentence Detection in Human Language Data	214
6.3.3 Document Summarization	218
6.4 Entity-Centric Analysis: A Paradigm Shift	228
6.4.1 Gisting Human Language Data	232
6.5 Quality of Analytics for Processing Human Language Data	238
6.6 Closing Remarks	240
6.7 Recommended Exercises	241
6.8 Online Resources	242
7. Mining Mailboxes: Analyzing Who’s Talking to Whom About What, How Often, and More.	245
.....	245
7.1 Overview	246
7.2 Obtaining and Processing a Mail Corpus	247
7.2.1 A Primer on Unix Mailboxes	247
7.2.2 Getting the Enron Data	252
7.2.3 Converting a Mail Corpus to a Unix Mailbox	254
7.2.4 Converting Unix Mailboxes to pandas DataFrames	256

7.3 Analyzing the Enron Corpus	259
7.3.1 Querying by Date/Time Range	260
7.3.2 Analyzing Patterns in Sender/Recipient Communications	264
7.3.3 Searching Emails by Keywords	267
7.4 Analyzing Your Own Mail Data	269
7.4.1 Accessing Your Gmail with OAuth	271
7.4.2 Fetching and Parsing Email Messages	273
7.4.3 Visualizing Patterns in Email with Immersion	275
7.5 Closing Remarks	276
7.6 Recommended Exercises	277
7.7 Online Resources	278

8. Mining GitHub: Inspecting Software Collaboration Habits, Building Interest Graphs, and More.	279
8.1 Overview	280
8.2 Exploring GitHub’s API	281
8.2.1 Creating a GitHub API Connection	282
8.2.2 Making GitHub API Requests	286
8.3 Modeling Data with Property Graphs	288
8.4 Analyzing GitHub Interest Graphs	292
8.4.1 Seeding an Interest Graph	292
8.4.2 Computing Graph Centrality Measures	296
8.4.3 Extending the Interest Graph with “Follows” Edges for Users	299
8.4.4 Using Nodes as Pivots for More Efficient Queries	311
8.4.5 Visualizing Interest Graphs	316
8.5 Closing Remarks	318
8.6 Recommended Exercises	319
8.7 Online Resources	320

Part II. Twitter Cookbook

9. Twitter Cookbook.	325
9.1 Accessing Twitter’s API for Development Purposes	326
9.2 Doing the OAuth Dance to Access Twitter’s API for Production Purposes	328
9.3 Discovering the Trending Topics	332
9.4 Searching for Tweets	333
9.5 Constructing Convenient Function Calls	335
9.6 Saving and Restoring JSON Data with Text Files	336
9.7 Saving and Accessing JSON Data with MongoDB	337
9.8 Sampling the Twitter Firehose with the Streaming API	340
9.9 Collecting Time-Series Data	342

9.10 Extracting Tweet Entities	343
9.11 Finding the Most Popular Tweets in a Collection of Tweets	345
9.12 Finding the Most Popular Tweet Entities in a Collection of Tweets	347
9.13 Tabulating Frequency Analysis	348
9.14 Finding Users Who Have Retweeted a Status	349
9.15 Extracting a Retweet's Attribution	351
9.16 Making Robust Twitter Requests	353
9.17 Resolving User Profile Information	355
9.18 Extracting Tweet Entities from Arbitrary Text	357
9.19 Getting All Friends or Followers for a User	357
9.20 Analyzing a User's Friends and Followers	360
9.21 Harvesting a User's Tweets	361
9.22 Crawling a Friendship Graph	363
9.23 Analyzing Tweet Content	365
9.24 Summarizing Link Targets	367
9.25 Analyzing a User's Favorite Tweets	370
9.26 Closing Remarks	371
9.27 Recommended Exercises	372
9.28 Online Resources	373

Part III. Appendixes

A. Information About This Book's Virtual Machine Experience.....	377
B. OAuth Primer.....	379
C. Python and Jupyter Notebook Tips and Tricks.....	385
Index.....	387