

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	16
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	30
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	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 Social Graph API	46
2.2.1. Understanding the Social Graph API	48
2.2.2. Understanding the Open Graph Protocol	54

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

5.2. Scraping, Parsing, and Crawling the Web	183
5.2.1. Breadth-First Search in Web Crawling	186
5.3. Discovering Semantics by Decoding Syntax	190
5.3.1. Natural Language Processing Illustrated Step-by-Step	192
5.3.2. Sentence Detection in Human Language Data	196
5.3.3. Document Summarization	200
5.4. Entity-Centric Analysis: A Paradigm Shift	209
5.4.1. Gisting Human Language Data	213
5.5. Quality of Analytics for Processing Human Language Data	219
5.6. Closing Remarks	222
5.7. Recommended Exercises	222
5.8. Online Resources	223
6. Mining Mailboxes: Analyzing Who's Talking to Whom About What, How Often, and More	225
.....	
6.1. Overview	226
6.2. Obtaining and Processing a Mail Corpus	227
6.2.1. A Primer on Unix Mailboxes	227
6.2.2. Getting the Enron Data	232
6.2.3. Converting a Mail Corpus to a Unix Mailbox	235
6.2.4. Converting Unix Mailboxes to JSON	236
6.2.5. Importing a JSONified Mail Corpus into MongoDB	240
6.2.6. Programmatically Accessing MongoDB with Python	244
6.3. Analyzing the Enron Corpus	247
6.3.1. Querying by Date/Time Range	248
6.3.2. Analyzing Patterns in Sender/Recipient Communications	250
6.3.3. Writing Advanced Queries	255
6.3.4. Searching Emails by Keywords	260
6.4. Discovering and Visualizing Time-Series Trends	264
6.5. Analyzing Your Own Mail Data	268
6.5.1. Accessing Your Gmail with OAuth	269
6.5.2. Fetching and Parsing Email Messages with IMAP	271
6.5.3. Visualizing Patterns in Gmail with the "Graph Your Inbox" Chrome Extension	273
6.6. Closing Remarks	274
6.7. Recommended Exercises	275
6.8. Online Resources	276
7. Mining GitHub: Inspecting Software Collaboration Habits, Building Interest Graphs, and More.....	279
7.1. Overview	280
7.2. Exploring GitHub's API	281

7.2.1. Creating a GitHub API Connection	282
7.2.2. Making GitHub API Requests	286
7.3. Modeling Data with Property Graphs	288
7.4. Analyzing GitHub Interest Graphs	292
7.4.1. Seeding an Interest Graph	292
7.4.2. Computing Graph Centrality Measures	296
7.4.3. Extending the Interest Graph with “Follows” Edges for Users	299
7.4.4. Using Nodes as Pivots for More Efficient Queries	311
7.4.5. Visualizing Interest Graphs	316
7.5. Closing Remarks	318
7.6. Recommended Exercises	318
7.7. Online Resources	320
8. Mining the Semantically Marked-Up Web: Extracting Microformats, Inferencing over RDF, and More.....	321
8.1. Overview	322
8.2. Microformats: Easy-to-Implement Metadata	322
8.2.1. Geocoordinates: A Common Thread for Just About Anything	325
8.2.2. Using Recipe Data to Improve Online Matchmaking	331
8.2.3. Accessing LinkedIn’s 200 Million Online Résumés	336
8.3. From Semantic Markup to Semantic Web: A Brief Interlude	338
8.4. The Semantic Web: An Evolutionary Revolution	339
8.4.1. Man Cannot Live on Facts Alone	340
8.4.2. Inferencing About an Open World	342
8.5. Closing Remarks	345
8.6. Recommended Exercises	346
8.7. Online Resources	347

Part II. Twitter Cookbook

9. Twitter Cookbook.....	351
9.1. Accessing Twitter’s API for Development Purposes	352
9.2. Doing the OAuth Dance to Access Twitter’s API for Production Purposes	353
9.3. Discovering the Trending Topics	358
9.4. Searching for Tweets	359
9.5. Constructing Convenient Function Calls	361
9.6. Saving and Restoring JSON Data with Text Files	362
9.7. Saving and Accessing JSON Data with MongoDB	363
9.8. Sampling the Twitter Firehose with the Streaming API	365
9.9. Collecting Time-Series Data	367
9.10. Extracting Tweet Entities	368

9.11. Finding the Most Popular Tweets in a Collection of Tweets	370
9.12. Finding the Most Popular Tweet Entities in a Collection of Tweets	372
9.13. Tabulating Frequency Analysis	373
9.14. Finding Users Who Have Retweeted a Status	374
9.15. Extracting a Retweet's Attribution	376
9.16. Making Robust Twitter Requests	378
9.17. Resolving User Profile Information	380
9.18. Extracting Tweet Entities from Arbitrary Text	381
9.19. Getting All Friends or Followers for a User	382
9.20. Analyzing a User's Friends and Followers	384
9.21. Harvesting a User's Tweets	386
9.22. Crawling a Friendship Graph	388
9.23. Analyzing Tweet Content	390
9.24. Summarizing Link Targets	391
9.25. Analyzing a User's Favorite Tweets	395
9.26. Closing Remarks	396
9.27. Recommended Exercises	397
9.28. Online Resources	398

Part III. Appendixes

A. Information About This Book's Virtual Machine Experience.	401
B. OAuth Primer.	403
C. Python and IPython Notebook Tips & Tricks.	409
Index.	411