
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 | 101 |
| 3.4.3 Object Recognition Within Photos Using Pretrained Neural Networks | 107 |
| 3.5 Applying Neural Networks to Instagram Posts | 111 |
| 3.5.1 Tagging the Contents of an Image | 111 |
| 3.5.2 Detecting Faces in Images | 112 |
| 3.6 Closing Remarks | 115 |
| 3.7 Recommended Exercises | 115 |
| 3.8 Online Resources | 116 |
| | |
| 4. Mining LinkedIn: Faceting Job Titles, Clustering Colleagues, and More. | 119 |
| 4.1 Overview | 120 |
| 4.2 Exploring the LinkedIn API | 121 |
| 4.2.1 Making LinkedIn API Requests | 121 |
| 4.2.2 Downloading LinkedIn Connections as a CSV File | 125 |
| 4.3 Crash Course on Clustering Data | 126 |
| 4.3.1 Normalizing Data to Enable Analysis | 129 |
| 4.3.2 Measuring Similarity | 141 |
| 4.3.3 Clustering Algorithms | 143 |
| 4.4 Closing Remarks | 159 |
| 4.5 Recommended Exercises | 160 |
| 4.6 Online Resources | 161 |

| | |
|---|------------|
| 5. Mining Text Files: Computing Document Similarity, Extracting Collocations, and More. | 163 |
| 5.1 Overview | 164 |
| 5.2 Text Files | 164 |
| 5.3 A Whiz-Bang Introduction to TF-IDF | 166 |
| 5.3.1 Term Frequency | 167 |
| 5.3.2 Inverse Document Frequency | 169 |
| 5.3.3 TF-IDF | 170 |
| 5.4 Querying Human Language Data with TF-IDF | 174 |
| 5.4.1 Introducing the Natural Language Toolkit | 174 |
| 5.4.2 Applying TF-IDF to Human Language | 177 |
| 5.4.3 Finding Similar Documents | 179 |
| 5.4.4 Analyzing Bigrams in Human Language | 187 |
| 5.4.5 Reflections on Analyzing Human Language Data | 197 |
| 5.5 Closing Remarks | 198 |
| 5.6 Recommended Exercises | 199 |
| 5.7 Online Resources | 200 |
| | |
| 6. Mining Web Pages: Using Natural Language Processing to Understand Human Language, Summarize Blog Posts, and More. | 201 |
| 6.1 Overview | 202 |
| 6.2 Scraping, Parsing, and Crawling the Web | 203 |
| 6.2.1 Breadth-First Search in Web Crawling | 206 |
| 6.3 Discovering Semantics by Decoding Syntax | 210 |
| 6.3.1 Natural Language Processing Illustrated Step-by-Step | 212 |
| 6.3.2 Sentence Detection in Human Language Data | 216 |
| 6.3.3 Document Summarization | 220 |
| 6.4 Entity-Centric Analysis: A Paradigm Shift | 230 |
| 6.4.1 Gisting Human Language Data | 234 |
| 6.5 Quality of Analytics for Processing Human Language Data | 240 |
| 6.6 Closing Remarks | 242 |
| 6.7 Recommended Exercises | 243 |
| 6.8 Online Resources | 244 |
| | |
| 7. Mining Mailboxes: Analyzing Who’s Talking to Whom About What, How Often, and More. | 247 |
| 7.1 Overview | 248 |
| 7.2 Obtaining and Processing a Mail Corpus | 249 |
| 7.2.1 A Primer on Unix Mailboxes | 249 |
| 7.2.2 Getting the Enron Data | 254 |
| 7.2.3 Converting a Mail Corpus to a Unix Mailbox | 256 |
| 7.2.4 Converting Unix Mailboxes to pandas DataFrames | 258 |

| | |
|--|------------|
| 7.3 Analyzing the Enron Corpus | 261 |
| 7.3.1 Querying by Date/Time Range | 262 |
| 7.3.2 Analyzing Patterns in Sender/Recipient Communications | 266 |
| 7.3.3 Searching Emails by Keywords | 269 |
| 7.4 Analyzing Your Own Mail Data | 271 |
| 7.4.1 Accessing Your Gmail with OAuth | 273 |
| 7.4.2 Fetching and Parsing Email Messages | 275 |
| 7.4.3 Visualizing Patterns in Email with Immersion | 278 |
| 7.5 Closing Remarks | 278 |
| 7.6 Recommended Exercises | 279 |
| 7.7 Online Resources | 280 |
| 8. Mining GitHub: Inspecting Software Collaboration Habits, Building Interest Graphs, and More. | 283 |
| 8.1 Overview | 284 |
| 8.2 Exploring GitHub's API | 285 |
| 8.2.1 Creating a GitHub API Connection | 286 |
| 8.2.2 Making GitHub API Requests | 290 |
| 8.3 Modeling Data with Property Graphs | 292 |
| 8.4 Analyzing GitHub Interest Graphs | 296 |
| 8.4.1 Seeding an Interest Graph | 296 |
| 8.4.2 Computing Graph Centrality Measures | 300 |
| 8.4.3 Extending the Interest Graph with "Follows" Edges for Users | 303 |
| 8.4.4 Using Nodes as Pivots for More Efficient Queries | 315 |
| 8.4.5 Visualizing Interest Graphs | 320 |
| 8.5 Closing Remarks | 322 |
| 8.6 Recommended Exercises | 323 |
| 8.7 Online Resources | 324 |
| <hr/> | |
| Part II. Twitter Cookbook | |
| 9. Twitter Cookbook | 329 |
| 9.1 Accessing Twitter's API for Development Purposes | 330 |
| 9.2 Doing the OAuth Dance to Access Twitter's API for Production Purposes | 332 |
| 9.3 Discovering the Trending Topics | 336 |
| 9.4 Searching for Tweets | 337 |
| 9.5 Constructing Convenient Function Calls | 339 |
| 9.6 Saving and Restoring JSON Data with Text Files | 340 |
| 9.7 Saving and Accessing JSON Data with MongoDB | 341 |
| 9.8 Sampling the Twitter Firehose with the Streaming API | 344 |
| 9.9 Collecting Time-Series Data | 346 |

| | |
|--|-----|
| 9.10 Extracting Tweet Entities | 347 |
| 9.11 Finding the Most Popular Tweets in a Collection of Tweets | 349 |
| 9.12 Finding the Most Popular Tweet Entities in a Collection of Tweets | 351 |
| 9.13 Tabulating Frequency Analysis | 352 |
| 9.14 Finding Users Who Have Retweeted a Status | 353 |
| 9.15 Extracting a Retweet's Attribution | 355 |
| 9.16 Making Robust Twitter Requests | 357 |
| 9.17 Resolving User Profile Information | 359 |
| 9.18 Extracting Tweet Entities from Arbitrary Text | 361 |
| 9.19 Getting All Friends or Followers for a User | 361 |
| 9.20 Analyzing a User's Friends and Followers | 364 |
| 9.21 Harvesting a User's Tweets | 365 |
| 9.22 Crawling a Friendship Graph | 367 |
| 9.23 Analyzing Tweet Content | 369 |
| 9.24 Summarizing Link Targets | 371 |
| 9.25 Analyzing a User's Favorite Tweets | 374 |
| 9.26 Closing Remarks | 375 |
| 9.27 Recommended Exercises | 376 |
| 9.28 Online Resources | 377 |

Part III. Appendixes

| | |
|--|-----|
| A. Information About This Book's Virtual Machine Experience..... | 381 |
| B. OAuth Primer..... | 383 |
| C. Python and Jupyter Notebook Tips and Tricks..... | 389 |
| Index..... | 391 |