

MINING SOCIAL MEDIA

**Finding Stories
in Internet Data**

by Lam Thuy Vo



**no starch
press**

San Francisco

BRIEF CONTENTS

Acknowledgments	xv
Introduction	xvii
PART I: DATA MINING	1
Chapter 1: The Programming Languages You'll Need to Know	3
Chapter 2: Where to Get Your Data.	27
Chapter 3: Getting Data with Code	43
Chapter 4: Scraping Your Own Facebook Data.	63
Chapter 5: Scraping a Live Site	77
PART II: DATA ANALYSIS	99
Chapter 6: Introduction to Data Analysis.	101
Chapter 7: Visualizing Your Data.	123
Chapter 8: Advanced Tools for Data Analysis	135
Chapter 9: Finding Trends in Reddit Data	151
Chapter 10: Measuring the Twitter Activity of Political Actors	163
Chapter 11: Where to Go from Here	177
Index	181

CONTENTS IN DETAIL

ACKNOWLEDGMENTS	xv
------------------------	-----------

INTRODUCTION	xvii
---------------------	-------------

What Is Data Analysis?xviii
Who Is This Book For?xviii
Conventions Used in This Bookxviii
What This Book Covers	xix
Part I: Data Mining	xix
Part II: Data Analysisxx
Downloading and Installing Pythonxx
Installing on Windowsxxi
Installing on macOSxxi
Getting Help When You're Stuckxxi
Summaryxxiv

PART I: DATA MINING	1
----------------------------	----------

1	
THE PROGRAMMING LANGUAGES YOU'LL NEED TO KNOW	3

Frontend Languages	4
How HTML Works.	4
How CSS Works.	6
How JavaScript Works	12
Backend Languages	14
Using Python	14
Getting Started with Python	14
Working with Numbers	16
Working with Strings.	17
Storing Values in Variables	17
Storing Multiple Values in Lists	19
Working with Functions	20
Creating Your Own Functions.	21
Using Loops	22
Using Conditionals	23
Summary	25

2		
WHERE TO GET YOUR DATA		27
What Is an API?		28
Using an API to Get Data		28
Getting a YouTube API Key		31
Retrieving JSON Objects Using Your Credentials		31
Answering a Research Question Using Data		37
Refining the Data That Your API Returns		41
Summary		41
3		
GETTING DATA WITH CODE		43
Writing Your First Script		44
Running a Script		44
Planning Out a Script		46
Libraries and pip		46
Creating a URL-based API Call		48
Storing Data in a Spreadsheet		49
Converting JSON into a Dictionary		51
Going Back to the Script		51
Running the Finished Script		53
Dealing with API Pagination		55
Templates: How to Make Your Code Reusable		57
Storing Values That Change in Variables		57
Storing Code in a Reusable Function		58
Summary		61
4		
SCRAPING YOUR OWN FACEBOOK DATA		63
Your Data Sources		64
Downloading Your Facebook Data		64
Reviewing the Data and Inspecting the Code		66
Structuring Information as Data		67
Scraping Automatically		68
Analyzing HTML Code to Recognize Patterns		70
Grabbing the Elements You Need		70
Extracting the Contents		71
Writing Data into a Spreadsheet		72
Building Your Rows List		72
Writing to Your .csv File		74
Running the Script		75
Summary		76
5		
SCRAPING A LIVE SITE		77
Messy Data		78
Ethical Considerations for Data Scraping		80
The Robots Exclusion Protocol		80

The Terms of Service	82
Technical Considerations for Data Scraping	82
Reasons for Scraping Data	82
Scraping from a Live Website	83
Analyzing the Page’s Contents	84
Storing the Page Content in Variables	88
Making the Script Reusable	92
Practicing Polite Scraping	94
Summary	98

PART II: DATA ANALYSIS 99

6 INTRODUCTION TO DATA ANALYSIS 101

The Process of Data Analysis	102
Bot Spotting	103
Getting Started with Google Sheets	104
Modifying and Formatting the Data	106
Aggregating the Data	110
Using Pivot Tables to Summarize Data	110
Using Formulas to Do Math	112
Sorting and Filtering the Data	114
Merging Data Sets	117
Other Ways to Use Google Sheets	121
Summary	122

7 VISUALIZING YOUR DATA 123

Understanding Our Bot Through Charts	124
Choosing a Chart	124
Specifying a Time Period	128
Making a Chart	128
Conditional Formatting	131
Single-Color Formatting	131
Color Scale Formatting	132
Summary	133

8 ADVANCED TOOLS FOR DATA ANALYSIS 135

Using Jupyter Notebook	136
Setting Up a Virtual Environment	136
Organizing the Notebook	138
Installing Jupyter and Creating Your First Notebook	139
Working with Cells	140
What Is pandas?	142
Working with Series and Data Frames	143
Reading and Exploring Large Data Files	145

Looking at the Data	146
Viewing Specific Columns and Rows	148
Summary	149

9 FINDING TRENDS IN REDDIT DATA 151

Clarifying Our Research Objective	152
Outlining a Method	152
Narrowing the Data’s Scope	153
Selecting Data from Specific Columns	153
Handling Null Values	154
Classifying the Data	156
Summarizing the Data	157
Sorting the Data	158
Describing the Data	160
Summary	162

10 MEASURING THE TWITTER ACTIVITY OF POLITICAL ACTORS 163

Getting Started	164
Setting Up Your Environment	164
Loading the Data into Your Notebook	165
Lambdas	168
Filtering the Data Set	169
Formatting the Data as datetimes	170
Resampling the Data	172
Plotting the Data	175
Summary	176

11 WHERE TO GO FROM HERE 177

Coding Styles	178
Statistical Analysis	179
Other Kinds of Analyses	179
Conclusion	180

INDEX 181