

# Social Media as Social Science Data

---

**Steven Lloyd Wilson**

Brandeis University



# Contents

<i>List of Figures</i>	page viii
<i>List of Tables</i>	ix
<i>List of Code Snippets</i>	x
<i>Acknowledgments</i>	xiii
<b>1. Why Social Media Matters to the Social Sciences</b>	<b>1</b>
1.1 The Theory of Why It Matters	2
1.2 Using Social Media Data	5
<b>2. Getting Started with Social Media Data</b>	<b>10</b>
2.1 What Is Twitter Data?	11
2.2 Accessing Twitter Data	12
2.2.1 The Search API	13
2.2.2 The Streaming API	14
2.2.3 Other Access	15
2.3 The Nuts and Bolts of the Data	16
2.3.1 Unique Ids	17
2.3.2 User Data	18
2.3.3 The Tweet Text Itself	18
2.3.4 Languages	19
2.3.5 Times and Time Zones	20
2.3.6 Entities	20
2.4 Downloading Twitter Data	21
2.4.1 A Framework for Data Collection	22
2.4.2 Getting Set Up	23
2.4.3 A Basic Downloader	25
2.4.4 A Basic Processor	28
2.4.5 Adding a Database Backend	30
2.4.6 Processing Multiple Files	34
2.4.7 Collecting a Worldwide Sample	36
2.4.8 Compressing Data	38
2.5 Conclusion	39

<b>3. Content Analysis of Social Media Data</b>	<b>40</b>
3.1 Text and Twitter	40
3.1.1 Downloading Tweets by Keyword	44
3.1.2 Downloading Tweets by Language	45
3.2 Other Content from Tweets	46
3.2.1 Hashtags	46
3.2.2 Mentions	49
3.2.3 Links and Their Content	51
3.2.4 Extracting and Downloading Images	55
3.2.5 Extracting and Downloading Videos	60
3.3 Computer Content Analysis of Text	63
3.3.1 Tools for Text	63
3.3.2 Sentiment Analysis	67
3.3.3 Topic Modeling	69
3.3.4 Supervised Learning Models	75
3.4 Key Takeaways	84
<b>4. Geospatial Analysis of Social Media Data</b>	<b>85</b>
4.1 Data Availability	85
4.2 Collecting Geocoded Data from Twitter	91
4.3 Processing Geocoded Tweets	95
4.3.1 Caching Coordinates	98
4.3.2 Guessing the Right Country	99
4.3.3 Fuzzy Matching for Coastlines	105
4.3.4 Simplifying the Search Further	107
4.3.5 Subnational Matching	110
4.3.6 Matching Places	111
4.4 Grid Cell Identification	113
4.5 Solving the Location Problem	115
4.6 Key Takeaways	121
<b>5. Network Analysis of Social Media Data</b>	<b>122</b>
5.1 Getting Started with Networks	122
5.2 Collecting Data by User	123
5.2.1 A One-Time Timeline Downloader	124
5.2.2 An Infrastructure for Timeline Downloading	126
5.2.3 Additional User Data	128
5.2.4 Sample Project: Tracking Congress	130
5.3 Networks of Friends and Followers	132
5.3.1 Friend Networks	134
5.3.2 Co-friend Networks	135
5.3.3 Networks with Entities as Nodes	137

---

5.4	Basics of Network Analysis	138
5.5	Solving the Bot Problem	142
5.6	Key Takeaways	145
6.	<b>The Ethics of Using Social Media Data</b>	146
6.1	Social Media and the IRB	146
6.2	Ethics in Web Scraping	151
6.3	Researcher Trauma	153
6.4	A Researcher's Perspective	157
	<i>References</i>	161
	<i>Index</i>	168