Many research projects involve analyzing sets of texts from the social web or elsewhere to get insights into issues, opinions, interests, news discussions, or communication styles. For example, many studies have investigated reactions to Covid-19 social distancing restrictions, conspiracy theories, and anti-vaccine sentiment on social media. This book describes word association thematic analysis, a mixed methods strategy to identify themes within a collection of social web or other texts. It identifies these themes in the differences between subsets of the texts, including female vs. male vs. nonbinary, older vs. newer, country A vs. country B, positive vs. negative sentiment, high scoring vs. low scoring, or subtopic A vs. subtopic B. It can also be used to identify the differences between a topic-focused collection of texts and a reference collection. The method starts by automatically finding words that are statistically significantly more common in one subset than another, then identifies the context of these words and groups them into themes. It is supported by the free Windows-based software Mozdeh for data collection or importing and for the quantitative analysis stages. This book explains the word association thematic analysis method, with examples, and gives practical advice for using it. It is primarily intended for social media researchers and students, although the method is applicable to any collection of short texts.
Word Association Thematic Analysis
A Social Media Text Exploration Strategy
Synthesis Lectures on Information Concepts, Retrieval, and Services

Editor
Gary Marchionini, University of North Carolina at Chapel Hill

Synthesis Lectures on Information Concepts, Retrieval, and Services publishes short books on topics pertaining to information science and applications of technology to information discovery, production, distribution, and management. Potential topics include: data models, indexing theory and algorithms, classification, information architecture, information economics, privacy and identity, scholarly communication, bibliometrics and webometrics, personal information management, human information behavior, digital libraries, archives and preservation, cultural informatics, information retrieval evaluation, data fusion, relevance feedback, recommendation systems, question answering, natural language processing for retrieval, text summarization, multimedia retrieval, multilingual retrieval, and exploratory search.

Word Association Thematic Analysis: A Social Media Text Exploration Strategy
Michael Thelwall

Simulating Information Retrieval Test Collections
David Hawking, Bodo Billerbeck, Paul Thomas, and Nick Craswell

Automatic Disambiguation of Author Names in Bibliographic Repositories
Anderson A. Ferreira, Marcos André Gonçalves, and Alberto H. F. Laender

Compatibility Modeling: Data and Knowledge Applications for Clothing Matching
Xuemeng Song, Liqiang Nie, and Yinglong Wang

Video Structure Meaning
Brian C. O'Connor and Richard L. Anderson

Interactive IR User Study Design, Evaluation, and Reporting
Jiqun Liu and Chirag Shah

The Practice of Crowdsourcing
Omar Alonso

Predicting Information Retrieval Performance
Robert M. Losee
Framing Privacy in Digital Collections with Ethical Decision Making
Virginia Dressler

Mobile Search Behaviors: An In-depth Analysis Based on Contexts, APPs, and Devices
Dan Wu and Shaobo Liang

Images in Social Media: Categorization and Organization of Images and Their Collections
Susanne Ørnager and Haakon Lund

Exploring Context in Information Behavior: Seeker, Situation, Surroundings, and Shared Identities
Naresh Kumar Agarwal

Researching Serendipity in Digital Information Environments
Lori McCay-Peet and Elaine G. Toms

Social Monitoring for Public Health
Michael J. Paul and Mark Dredze

Digital Libraries for Cultural Heritage: Development, Outcomes, and Challenges from European Perspectives
Tatjana Aparac-Jelušić

iRODS Primer 2: Integrated Rule-Oriented Data System
Hao Xu, Terrell Russell, Jason Coposky, Arcot Rajasekar, Reagan Moore, Antoine de Torcy, Michael Wan, Wayne Shroeder, and Sheau-Yen Chen

Information Architecture: The Design and Integration of Information Spaces, Second Edition
Wei Ding, Xia Lin, and Michael Zarro

Fuzzy Information Retrieval
Donald H. Kraft and Erin Colvin

Quantifying Research Integrity
Michael Seadle

Incidental Exposure to Online News
Borchuluun Yadamsuren and Sanda Erdelez

Web Indicators for Research Evaluation: A Practical Guide
Michael Thelwall

Trustworthy Policies for Distributed Repositories
Reagan W. Moore, Hao Xu, Mike Conway, Arcot Rajasekar, Jon Crabtree, and Helen Tibbo
The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?
Tefko Saracevic

Dynamic Information Retrieval Modeling
Grace Hui Yang, Marc Sloan, and Jun Wang

Learning from Multiple Social Networks
Liqiang Nie, Xuemeng Song, and Tat-Seng Chua

Scholarly Collaboration on the Academic Social Web
Daqing He and Wei Jeng

Scalability Challenges in Web Search Engines
B. Barla Cambazoglu and Ricardo Baeza-Yates

Social Informatics Evolving
Pnina Fichman, Madelyn R. Sanfilippo, and Howard Rosenbaum

Mark S. Manasse

Building a Better World with Our Information: The Future of Personal Information Management, Part 3
William Jones

Click Models for Web Search
Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke

Information Communication
Feicheng Ma

Social Media and Library Services
Lorraine Mon

Analysis and Visualization of Citation Networks
Dangzhi Zhao and Andreas Strotmann

The Taxobook: Applications, Implementation, and Integration in Search, Part 3
Marjorie M. K. Hlava

The Taxobook: Principles and Practices of Building Taxonomies, Part 2
Marjorie M. K. Hlava
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measuring User Engagement</td>
<td>Mounia Lalmas, Heather O’Brien, and Elad Yom-Tov</td>
</tr>
<tr>
<td>Children's Internet Search: Using Roles to Understand Children's Search Behavior</td>
<td>Elizabeth Foss and Allison Druin</td>
</tr>
<tr>
<td>Digital Library Technologies: Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security</td>
<td>Edward A. Fox and Ricardo da Silva Torres</td>
</tr>
<tr>
<td>Information and Human Values</td>
<td>Kenneth R. Fleischmann</td>
</tr>
<tr>
<td>Multiculturalism and Information and Communication Technology</td>
<td>Pnina Fichman and Madelyn R. Sanfilippo</td>
</tr>
<tr>
<td>Transforming Technologies to Manage Our Information: The Future of Personal Information Management, Part II</td>
<td>William Jones</td>
</tr>
<tr>
<td>Designing for Digital Reading</td>
<td>Jennifer Pearson, George Buchanan, and Harold Thimbleby</td>
</tr>
<tr>
<td>Information Retrieval Models: Foundations and Relationships</td>
<td>Thomas Roelleke</td>
</tr>
<tr>
<td>Visual Information Retrieval Using Java and LIRE</td>
<td>Mathias Lux and Oge Marques</td>
</tr>
<tr>
<td>On the Efficient Determination of Most Near Neighbors: Horseshoes, Hand Grenades, Web Search and Other Situations When Close is Close Enough</td>
<td>Mark S. Manasse</td>
</tr>
<tr>
<td>The Answer Machine</td>
<td>Susan E. Feldman</td>
</tr>
</tbody>
</table>
Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach
Edward A. Fox, Marcos André Gonçalves, and Rao Shen

The Future of Personal Information Management, Part I: Our Information, Always and Forever
William Jones

Search User Interface Design
Max L. Wilson

Information Retrieval Evaluation
Donna Harman

Knowledge Management (KM) Processes in Organizations: Theoretical Foundations and Practice
Claire R. McInerney and Michael E. D. Koenig

Search-Based Applications: At the Confluence of Search and Database Technologies
Gregory Grefenstette and Laura Wilber

Information Concepts: From Books to Cyberspace Identities
Gary Marchionini

Estimating the Query Difficulty for Information Retrieval
David Carmel and Elad Yom-Tov

iRODS Primer: Integrated Rule-Oriented Data System

Meredith Ringel Morris and Jaime Teevan

Multimedia Information Retrieval
Stefan Rüger

Online Multiplayer Games
William Sims Bainbridge

Information Architecture: The Design and Integration of Information Spaces
Wei Ding and Xia Lin

Reading and Writing the Electronic Book
Catherine C. Marshall
Hypermedia Genes: An Evolutionary Perspective on Concepts, Models, and Architectures
Nuno M. Guimarães and Luís M. Carrico

Understanding User-Web Interactions via Web Analytics
Bernard J. (Jim) Jansen

XML Retrieval
Mounia Lalmas

Faceted Search
Daniel Tunkelang

Introduction to Webometrics: Quantitative Web Research for the Social Sciences
Michael Thelwall

Exploratory Search: Beyond the Query-Response Paradigm
Ryen W. White and Resa A. Roth

New Concepts in Digital Reference
R. David Lankes

Automated Metadata in Multimedia Information Systems: Creation, Refinement, Use in Surrogates, and Evaluation
Michael G. Christel
Word Association Thematic Analysis

A Social Media Text Exploration Strategy

Michael Thelwall
University of Wolverhampton

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #72
ABSTRACT
Many research projects involve analyzing sets of texts from the social web or elsewhere to get insights into issues, opinions, interests, news discussions, or communication styles. For example, many studies have investigated reactions to Covid-19 social distancing restrictions, conspiracy theories, and anti-vaccine sentiment on social media. This book describes word association thematic analysis, a mixed methods strategy to identify themes within a collection of social web or other texts. It identifies these themes in the differences between subsets of the texts, including female vs. male vs. nonbinary, older vs. newer, country A vs. country B, positive vs. negative sentiment, high scoring vs. low scoring, or subtopic A vs. subtopic B. It can also be used to identify the differences between a topic-focused collection of texts and a reference collection. The method starts by automatically finding words that are statistically significantly more common in one subset than another, then identifies the context of these words and groups them into themes. It is supported by the free Windows-based software Mozdeh for data collection or importing and for the quantitative analysis stages. This book explains the word association thematic analysis method, with examples, and gives practical advice for using it. It is primarily intended for social media researchers and students, although the method is applicable to any collection of short texts.

KEYWORDS
word association, social media, thematic analysis, text analysis, statistics, data collection
# Contents

Acknowledgments .......................................................... xv

1 Introduction ........................................................................ 1
  1.1 Overview: Word Association Detection, Contextualization, and
       Thematic Analysis ..................................................... 3
  1.2 WAA and WATA Examples ............................................. 5
  1.3 Research Philosophy: Mixed Methods ............................. 7
  1.4 Comparison with Traditional Social Research Methods ........... 8
  1.5 Software: Mozdeh ...................................................... 10
  1.6 Language and International Considerations ...................... 11
  1.7 Using This Book ....................................................... 11

2 Data Collection with Mozdeh ........................................... 13
  2.1 Sample Size ............................................................ 13
  2.2 Data Collection Methods ............................................ 13
    2.2.1 Query-Based Post Collection (Twitter) ....................... 13
    2.2.2 User-Based Post Collection (Twitter, YouTube) .......... 15
    2.2.3 User Profile Bio Collection (Twitter) ....................... 16
    2.2.4 Other Methods to Identify a Set of Texts for Analysis .... 16
  2.3 Mozdeh Data Collection Procedures ............................... 16
    2.3.1 Query-Based Post Collection in Mozdeh .................... 17
    2.3.2 User-Based Post Collection with Mozdeh .................. 19
    2.3.3 User Profile Bio Collection with Mozdeh .................. 19
    2.3.4 Importing Text Collections (Academic Publications, Other)
       into Mozdeh ....................................................... 19
  2.4 Filtering the Initial Dataset ......................................... 21
  2.5 Summary ................................................................ 21

3 Word Association Detection: Term Identification ................. 23
  3.1 Word Association Detection: Subset vs. the Rest ............... 24
    3.1.1 Choice of Comparator Set .................................... 25
    3.1.2 Word Association Detection with Mozdeh ................. 26
  3.2 Word Association Detection: Subset A vs. Subset B ............ 30
3.3 Statistical Details .............................................. 33
  3.3.1 Theoretical Assumptions .................................. 35
  3.3.2 Effect Sizes ............................................ 36
3.4 Protection Against Accidentally Finding Irrelevant Words .............. 37
3.5 Sample Size Requirements ..................................... 37
3.6 Limitations ................................................... 37
3.7 Summary .................................................... 38

4  Word Association Contextualization: Term Meaning and Context ........ 41
4.1 Word Association Contextualization from a Random Sample of Relevant Texts ................................................ 43
  4.1.1 Word Association Contextualization Overview ................ 43
  4.1.2 Creating the WAC Text Sample with Mozdeh ................ 44
  4.1.3 Detecting Typical Meanings for Polysemous Words and Words in Expressions .......................................... 47
  4.1.4 Identifying Word Contexts ................................ 48
  4.1.5 Difficult Cases: Pronouns and Stylistic Words ................. 49
  4.1.6 Random Sample Summary ................................ 50
4.2 Advanced Word Association Contextualization .......................... 52
  4.2.1 Follow-Up Word Association Detection for the Term ........... 52
  4.2.2 Follow-Up Word Association Detection Within Texts Containing the Term ............................................ 54
4.3 Word Association Contextualization Test Assumptions ................ 56
4.4 Causes of Word Associations: News, Viral Sharing, or Societal Differences? ...................................................... 57
4.5 Using Multiple Coders ......................................... 58
4.6 Reporting WAA ............................................... 59
4.7 Summary .................................................... 59

5  Word Association Thematic Analysis: Theme Detection .................. 61
5.1 Assigning Themes to Words ...................................... 61
5.2 WATA Procedure .............................................. 62
5.3 WATA Visual and Performing Arts Example ........................ 64
5.4 Using Multiple Coders ......................................... 69
5.5 Reporting WATA ............................................... 69
  5.5.1 Theme Prevalence Information ............................. 70
  5.5.2 Discussing the Results ................................... 72
5.6 Summary .................................................... 72
Acknowledgments

Thank you to Professor Stephanie W. Haas of the University of North Carolina School of Information and Library Science and Professor Han Woo Park of the Department of Media and Communication at YeungNam University for very helpful comments on an earlier version of this book.
CHAPTER 1

Introduction

Many social science researchers analyze sets of texts to detect themes relevant to their research goals. In politics, they might identify the main topics discussed by the supporters of a new Spanish left-wing political party, international environmental activist group, or U.S.-based vaccine conspiracy theory. Health scientists might investigate how patients with asthma in Nigeria share information about it, or whether misinformation about asthma in India is shared online. In media and communication studies, the objective might be to investigate reactions to an environmental disaster or to identify gendered communication styles. In marketing, a project might identify the aspects of tourist attractions in India evaluated by visitors in online TripAdvisor reviews. While there are many existing research methods that can help these projects to find themes, the method introduced in this book explores gender, nationality, sentiment, popularity, or topic differences within the texts to identify themes in a way that is supported by statistical tests, as conducted automatically by the supporting software, and gives more fine-grained results than common existing methods.

This book describes two methods to identify themes in collections of texts from the social web or elsewhere. The texts might be sets of tweets about Covid-19, comments on YouTube videos posted by prominent fashion bloggers, posts to a sports discussion forum, journal article abstracts, news articles, TripAdvisor reviews, or another collection of short texts. The two methods described in this book are word association analysis (WAA) and word association thematic analysis (WATA). The standard WAA method identifies a set of words indicating differences in the set of texts. For example, it might find that the word cosplay was more tweeted by females than males in a set of manga tweets, indicating a gender difference in this aspect of manga fandom. WATA includes a follow-up identification of themes in WAA results when there are too many words to report individually. For example, in a set of manga tweets, one theme might be that fighting-related words are more used by male manga fans on Twitter. WATA is a large sample extension of WAA because each theme consists of one or more words identified by the first method. These methods are supported by the free Windows-based software Mozdeh (http://mozeh.wlv.ac.uk).

The two methods both work by identifying differences between subsets of the texts, as specified by the researcher, including the following.

- **Gender:** Which issues tend to be discussed more by males, females, or nonbinary people? (Nonbinary genders are currently only identifiable from tweets for technical reasons.)
1. INTRODUCTION

- **Country**: Which issues tend to be discussed more in one country than another (sharing a common language)?

- **Popularity**: Which issues are the most or least successful (e.g., as reflected by retweet counts)?

- **Sentiment**: Which issues generate the most positivity or negativity?

- **Time**: Which issues were discussed most during a given period, such as in the first or last documents?

- **Subtopics**: Which issues were discussed more within a given subtopic (e.g., the care home subtopic within Covid-19 tweets)?

- **Topic/reference set**: Which issues were discussed more within the topic than in a reference set (e.g., what characterizes political tweets compared to general tweets)?

The word association methods center on a comparison between two sets of texts or between two parts of a text collection, but they are also suitable for text-based research projects that do not center on comparisons. This is because they can give general insights into the topics discussed using the last method above: comparison with a reference set. For example, an investigation into how people with Attention Deficit Hyperactivity Disorder (ADHD) discussed their condition on Twitter compared their tweets to the tweets of people discussing other conditions to identify ADHD-specific themes. This study gave additional insights compared to a standard thematic analysis of the same ADHD tweets that did not contrast them with a reference set (Thelwall et al., 2021).

This book is for researchers and social media students that wish to analyze issues in the social web or other short text document collections. The goal might be to analyze an issue indirectly through the available social web or other documents or to analyze the social web itself directly through posts. Using the ADHD example above, in the first (indirect evidence) case the research question might be, “What do people with ADHD consider important about their condition?” and would be answered indirectly through social web evidence, assuming that the social web reflected the offline world to some extent. In the second (direct evidence) case, the question would focus on the social website, “How do people with ADHD tweet about their condition?” and get direct evidence from Twitter.
1.1 OVERVIEW: WORD ASSOCIATION DETECTION, CONTEXTUALIZATION, AND THEMATIC ANALYSIS

The first part of either type of project, whether WAA or WATA, is Word Association Detection (WAD). This uses the software Mozdeh to process a set of texts to find words that are statistically significantly more common in one subset of the texts than another. Here are two illustrative examples.

- Suppose that you are analyzing a set of 50,000 tweets with the hashtag #BlackLivesMatter and are investigating gender differences in tweeting about this issue. Mozdeh would produce lists of words that are more prevalent in male, female or nonbinary tweets. The result might be: Female: *structural, inequalities*; Male: *Trump, Democrat*; Nonbinary: *protest* (the lists are usually much longer).

- Suppose that you are seeking international UK vs. U.S. differences in comments posted to fashion influencer YouTube channels. For the WAD part, Mozdeh would separate the comments on U.S. YouTubers’ videos from the comments on UK YouTubers’ videos and produce a list of words that are more prevalent in the comments of one of either the UK or the U.S. The result might be: U.S.: *proms, Abercrombie*; UK: *Gents, leggings* (again, the lists are usually much longer).

To illustrate this further, suppose also that one of the words reported by Mozdeh as being female-associated in the #BlackLivesMatter tweets is “structural”, with 2.3% of female-authored tweets containing the term compared to 1.1% of nonbinary or male-authored tweets. Although this is a small percentage difference, Mozdeh reports that it is statistically significant, so is unlikely to have occurred by chance. Lists of words and percentages are the output of the WAD stage.

The next stage is Word Association Contextualization (WAC), which involves reading texts containing each term discovered to identify its typical meaning (if it is polysemic) and context. For example, the female-associated word “structural” in the #BlackLivesMatter tweets has a dictionary definition of, “relating to the way in which parts of a system or object are arranged” (Cambridge Dictionary, 2020). In the #BlackLivesMatter tweets the context of this word is much more specific than given by the dictionary definition: it is almost always used by females to point out the structural inequalities in the system that disadvantage people because they are Black. This contextualization suggests that females are more likely to tweet about racist structural inequalities than are males or nonbinaries (male #BlackLivesMatter tweets on this topic in particular tend to be more party political). In this case, the word “inequalities” was also female associated (1.2% vs. 0.6%), for the same underlying reason that females using #BlackLivesMatter were more likely to tweet about structural inequalities than were males using #BlackLivesMatter.

WAA comprises WAD and WAC, producing lists of words that are more prevalent in one subset of the texts than another (e.g., males vs. females; nonbinary vs. male/female; earlier vs. later;
topic A vs. topic B; popular vs. unpopular; country A vs. country B; all texts vs. reference set) and short explanations of the contexts of each word:

\[ \text{WAA} = \text{WAD} + \text{WAC} \]

If the WAA stage produces many terms (more than 15–25) then a Thematic Analysis (TA) extra stage is needed to group the WAA terms into coherent themes containing one or more terms that have similar meanings or contexts. The purpose of this stage is to synthesize long lists into a much shorter list of themes to help with reporting the results and to help with drawing wider conclusions from them. If the TA stage is used, then the method is called WATA:

\[ \text{WATA} = \text{WAD} + \text{WAC} + \text{TA} \]

A WATA therefore has three main stages (Figure 1.1). The WAD stage is automated: After the texts are split into two or three groups (see later), the computer finds words that are more prevalent in Set A than Set B. The WAC stage is manual: a human reads appropriate texts in Set A to add context to the WAD words. The TA stage is also manual: a human organizes the contextualized words into coherent themes.
1.2 WAA AND WATA EXAMPLES

WAA studies have analyzed YouTube comments collected by Mozdeh or TripAdvisor reviews imported into Mozdeh (Table 1.1). They investigated an aspect of gender and included other types of analysis within the published article. All studies should be available free online to consult for detailed examples of the methods, although they do not name WAA explicitly.

Figure 1.1: The three main stages of word association thematic analysis.
Table 1.1: Studies using WAA

<table>
<thead>
<tr>
<th>Topic</th>
<th>Data</th>
<th>Comparison</th>
<th>Example Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gendered reactions to YouTube science videos</td>
<td>Comments on videos from 50 YouTube science channels</td>
<td>Female vs. male</td>
<td>Offensive gendered language is rare but directed at females (Thelwall and Mas-Bleda, 2018).</td>
</tr>
<tr>
<td>Dance</td>
<td>Comments on dance videos in YouTube</td>
<td>Female vs. male</td>
<td>Gendered and sentiment-associating terms for each dance style (Thelwall, 2018a).</td>
</tr>
<tr>
<td>Gender bias in sentiment analysis</td>
<td>TripAdvisor reviews</td>
<td>Female vs. male</td>
<td>Female hotel reviews express sentiment more explicitly (Tables 2 and 3 of Thelwall, 2018d).</td>
</tr>
</tbody>
</table>

WATA studies have analyzed social media (tweets, Twitter profile descriptions, YouTube comments, SteemIt, and Reddit), and academic journal article titles, abstracts, and keywords (Table 1.2). Most investigated an aspect of gender. The project about ADHD tweeting focused on a single topic, personal experiences of ADHD, rather than on any type of difference. Nevertheless, as a methodological choice, this paper compared tweets about ADHD with tweets about other disorders or diseases to discover features that were particularly prevalent for ADHD. A similar strategy was used in the paper about bullying discussions on YouTube, which contrasted them to comments on other topics. Other projects compared countries, time periods, or popularity levels. For the Twitter and YouTube projects, the program Mozdeh collected and analyzed the texts whereas for the other projects it imported them for analysis.

Table 1.2: Studies using WATA

<table>
<thead>
<tr>
<th>Topic</th>
<th>Data</th>
<th>Comparison</th>
<th>Example Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender differences in reactions to Covid-19</td>
<td>Tweets mentioning Covid-19</td>
<td>Female vs. male</td>
<td>Females tweet more about safety; males more about politics (Thelwall and Thelwall, 2020).</td>
</tr>
<tr>
<td>Personal experiences of ADHD</td>
<td>Tweets about “my ADHD”</td>
<td>ADHD vs. other disorders</td>
<td>The brain is discussed as if it is a separate entity (Thelwall, et al., submitted).</td>
</tr>
<tr>
<td><strong>Evolution of #BlackLivesMatter during Covid-19</strong></td>
<td>Covid-19 tweets about racism</td>
<td>Tweets in four different periods</td>
<td>The George Floyd killing led to tweets about systematic racism (Thelwall and Thelwall, submitted-a).</td>
</tr>
<tr>
<td>-------------------------------------------------</td>
<td>-----------------------------</td>
<td>--------------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Self-presentation on Twitter</strong></td>
<td>UK Twitter profiles</td>
<td>Female vs. male vs. nonbinary</td>
<td>Nonbinary profiles more likely to mention games and sexuality (Thelwall et al., 2021).</td>
</tr>
<tr>
<td><strong>Autism on Twitter</strong></td>
<td>U.S. autism tweets during Covid-19</td>
<td>Autism vs. others</td>
<td>Autistic tweeters do not have distinctive reactions to Covid-19 (Thelwall and Thelwall, submitted).</td>
</tr>
<tr>
<td><strong>Gender differences in museum interests</strong></td>
<td>Comments on YouTube museum videos</td>
<td>Female vs. male</td>
<td>Females are more explicitly positive about content (Thelwall, 2018c).</td>
</tr>
<tr>
<td><strong>Discussions of bullying in YouTube</strong></td>
<td>Comments on YouTube influencer videos</td>
<td>Bullying vs. others</td>
<td>(Thelwall and Cash, to appear)</td>
</tr>
<tr>
<td><strong>Interests on Reddit</strong></td>
<td>Reddit posts</td>
<td>Female vs. male</td>
<td>Females more likely to mention doctors in the science subreddit (Thelwall and Stuart, 2019).</td>
</tr>
<tr>
<td><strong>Factors associated with success in SteemIt</strong></td>
<td>Steemit (like Reddit) posts</td>
<td>Successful vs. unsuccessful posts</td>
<td>Financial news is less likely to be rewarded (Thelwall, 2018b).</td>
</tr>
<tr>
<td><strong>Nursing research</strong></td>
<td>Nursing journal articles*</td>
<td>U.S. vs. other countries</td>
<td>Nursing administration and management is not studied in some countries (Thelwall and Mas-Bleda, in press).</td>
</tr>
<tr>
<td><strong>U.S. research subjects</strong></td>
<td>U.S. journal articles*</td>
<td>Female vs. male</td>
<td>Lists of gendered research topics and styles (Thelwall, et al., 2019b).</td>
</tr>
<tr>
<td><strong>UK research subjects</strong></td>
<td>UK journal articles*</td>
<td>Female vs. male</td>
<td>Lists of gendered research topics and styles (Thelwall et al., 2020).</td>
</tr>
<tr>
<td><strong>Indian research subjects</strong></td>
<td>Indian journal articles*</td>
<td>Female vs. male</td>
<td>Lists of gendered research topics and styles (Thelwall, et al., 2019a).</td>
</tr>
</tbody>
</table>

*Article titles, abstracts, and keywords were analyzed but not full texts.
1.3 RESEARCH PHILOSOPHY: MIXED METHODS

Word association thematic analysis uses a mixed methods research paradigm (Johnson, Onwuegbuzie, and Turner, 2007) driven by pragmatism: the need to combine both quantitative and qualitative elements sequentially to extract meaningful information from word frequencies in a set of texts. More specifically, it uses a sequential mixed methods design (Schoonenboom and Johnson, 2017) with a quantitative method followed by qualitative methods. The purpose of the qualitative methods is to explain the quantitative results, so WATA could also be classed as a two-phase explanatory mixed methods research design (Creswell, Plano Clark, Gutmann, and Hanson, 2003).

The first part of WATA involves automatically identifying words that are more common in one set of texts than another. This is purely quantitative. The software used incorporates statistical significance testing to create a set of words of interest and requires no human involvement or interpretation. Outside WATA, this approach is sometimes used for illustrative purposes, such as to create a word cloud of common words. Simple lists (or clouds) of words would not give useful information about the social context of the data or meaningful interpretations of it. They put the onus on the reader to interpret the meaning of the words identified. Someone looking at a word cloud would have to guess why the words were included and try to identify any patterns.

The second part of WATA involves human subjective judgments to identify the contexts of the words identified from the quantitative first stage and to organize the results into themes. This converts that list of words into patterns in the data that incorporate some degree of its social context. This is a qualitative approach because it involves subjective judgements and theme building.

Both the quantitative and qualitative parts of WATA are needed. Without the qualitative second part, the list of words is ambiguous and context free, giving little insight into the source of the texts. Without the quantitative first part, a qualitative analyses of words selected somehow from a set of texts would be meaningless. While other qualitative and quantitative methods can also be applied to sets of texts, this book makes the case that the WATA combination provides a useful new mixed methods approach.

1.4 COMPARISON WITH TRADITIONAL SOCIAL RESEARCH METHODS

Social research into attitudes, beliefs, and knowledge (called topics here for convenience) has traditionally used surveys, interviews, focus groups, or statistical data. Surveys and statistical data are ideal for obtaining information about large collections of people for pre-defined questions. In contrast, interviews and focus groups can be exploratory, allowing participants to introduce new ideas to the researchers. The alternative approach followed in this book is to harness texts discussing a topic from places where they already exist, such as the social web, and identify important themes. This strategy has both advantages and disadvantages compared to the others. Its main advantages
are data collection speed and large sample size. Its main disadvantages are the restriction to the topics discussed in the documents analyzed and the possibility that the people creating the documents may be unrepresentative. Although there are many social research data sources, five broad types are summarized below, for context.

- **Interviews**: The researcher identifies a sample of people relative to the research question, interviews them in person or remotely and analyzes the transcripts of the interviews using a text analysis method, such as thematic analysis (Braun and Clarke, 2006, 2013) or content analysis (Neuendorf, 2016).

- **Surveys**: The researcher designs a questionnaire with closed questions and (often) some open questions and then (usually) sends it to a large set (hundreds or thousands) of people in the target group. Closed questions can be analyzed with descriptive or inferential statistics. Open questions may be analyzed informally or with a text analysis method, such as content analysis or thematic analysis.

- **Statistical data**: The researcher obtains data about the target population from existing sources and analyzes them statistically to obtain conclusions about the relationships between variables. For example, the relationship between poverty and academic achievement might be analyzed with national government statistics about school performance, combined with a different set of national government data about poverty levels in the catchment areas of each school.

- **Social web texts**: The researcher gathers a set of topic-relevant texts from a social website, such as based on a set of queries and analyzes them with a text analysis method.

- **Documents**: The researcher obtains a sample or complete set of documents encapsulating the issue to be discussed and then analyzes them with a text analysis or discourse analysis method. This is common in corpus linguistics to identify patterns of language use from sets of novels, genres, or natural language transcripts. It is also used in the scientometrics field to analyze patterns of scientific investigation from the titles and/or abstracts of journal articles.

This book describes word association analysis methods to analyze texts from social websites or other document collections: the last two in the above list.

While the information from closed survey questions can be analyzed with statistics, qualitative methods, such as thematic analysis, discourse analysis, or content analysis, are needed to explore the text produced by open-ended survey questions, interviews and focus groups. All research methods have limitations (Table 1.3), with the nature of these limitations varying between methods. These methods are time consuming (for the investigator; also needing participant time), require
10 1. INTRODUCTION

ethical approval (also takes time), and the ability to generalize may be limited by small samples (interviews, focus groups) or unrepresentative samples (e.g., self-selection or sampling biases for most surveys).

<p>| Table 1.3: A comparison of factors affecting a selection of text-based social research data |</p>
<table>
<thead>
<tr>
<th>Factor</th>
<th>Interviews</th>
<th>Surveys</th>
<th>Statistical</th>
<th>Social web</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Labor intensive to organize and conduct</td>
<td>Labor intensive to design and sample well</td>
<td>Labor intensive to learn</td>
<td>Labor intensive to learn</td>
<td>Labor intensive to learn</td>
</tr>
<tr>
<td>Sample Bias</td>
<td>Small, impossible to be representative</td>
<td>Selection and self-selection biases common</td>
<td>Depends on sources</td>
<td>Social web user bias</td>
<td>Depends on sources</td>
</tr>
<tr>
<td>Ethics</td>
<td>Approval and safeguards needed</td>
<td>Approval needed</td>
<td>Approval often not needed</td>
<td>Approval often not needed if public</td>
<td>Approval not needed if public</td>
</tr>
<tr>
<td>IT</td>
<td>Recommended (e.g., Nvivo)</td>
<td>Statistical software</td>
<td>Statistical software</td>
<td>Specialist software</td>
<td>Specialist software</td>
</tr>
<tr>
<td>Topic Bias</td>
<td>Researcher questions and participant suggestions</td>
<td>Researcher questions and participant suggestions</td>
<td>Researcher ideas</td>
<td>Participant topics of interest</td>
<td>Scope of document collections</td>
</tr>
</tbody>
</table>

1.5 SOFTWARE: MOZDEH

This book draws upon methods implemented within the free Windows software Mozdeh (http://mozdeh.wlv.ac.uk). Researchers without access to a Windows computer (or a computer with Windows as part of a dual installation, such as through Boot Camp on a Mac) will need to borrow one for the first stage of the method.

Mozdeh can harvest data from selected social network sites (Twitter and YouTube at the time of writing) and save it to the local computer for processing. It can also import texts from other sources. Its analysis interface has buttons for all the word association analyses described in this book. Conducting any type of word association analysis with Mozdeh therefore entails collecting/importing and processing with the software, but the investigator still needs to interpret the Mozdeh outputs.
1.6 LANGUAGE AND INTERNATIONAL CONSIDERATIONS

The methods described in this book apply, in theory, to any written language on the Internet, although the examples below are all for English-language texts. The start of the WATA method involves comparing the frequency of words across sets of text in the same language, which can be any single language. There is one caveat, however. Some written languages (e.g., Burmese, Chinese, Japanese, Khmer, Lao, Thai, and Vietnamese) do not use spaces or other markers between words, leaving the reader to group characters into words to understand the meaning of a text. This causes a problem for the methods in this book that work at the level of words. There are algorithms that can split texts into words for the above languages, however. The above-mentioned software Mozdeh incorporates a word segmenting tool of this kind for Japanese (atilika, 2014), but for other languages an external program would be needed (e.g., Manning et al., 2014), with the output fed into Mozdeh.

Many languages are extensively spoken in multiple countries and many countries have substantial numbers of native speakers of multiple languages. Since the methods in this book work within a single language, consideration must be given to (a) which language to focus on, (b) how to ensure that only people from the desired country or countries are included for that language, and (c) whether to run multiple studies, one per language, if there are multiple important languages. For example, a study of social media use in the U.S. might want to collect tweets in both English and Spanish for two parallel studies and consider how to ensure that the tweet authors largely exclude English and Spanish speakers from elsewhere in the world (Mozdeh can help with this for tweets but not for YouTube comments).

1.7 USING THIS BOOK

This book describes the core word association analysis and word association thematic analysis methods, outlining how to conduct them with Mozdeh, explaining the rationale behind the methods used and relevant considerations for researchers. It is important to try out the methods a few times before starting a full-scale research project so that key early decisions (particularly for data collection) do not lead to invalid research projects. It is therefore essential to experiment with the methods after each chapter. Pilot datasets for this should be as large as possible because the methods are more powerful and insightful for larger collections of texts, so a small sample is likely to give disappointing results.

The book is arranged in a linear sequence, where many chapters depend on previous chapters. It is therefore best read in sequential order except that some chapters can be skipped.

• **Data collection with Mozdeh**: Skip the subsections not relevant to your project.
1. INTRODUCTION

- **Word association detection statistical details**: This can be skipped if you are not interested in the statistical background to the methods.

- **Word association thematic analysis chapters**: These can be omitted if you expect a small sample of texts or otherwise do not need a thematic analysis.

- **Comparison with other methods**: This can be safely skipped but this information may be useful when justifying your research methods in your dissertation, thesis, or article.
CHAPTER 2

Data Collection with Mozdeh

This chapter gives a brief overview of the generic issues involved with collecting data, together with examples for Twitter and YouTube and an overview of the Mozdeh procedures. It finishes with a discussion about importing texts, with the example of academic publication metadata.

2.1 SAMPLE SIZE

For all sources of texts, a large sample is needed for a reasonable analysis. This is because there are unlikely to be statistically significant differences in small sets of texts. If a reference set of texts is being used, then the guideline can be halved because the reference set adds to the statistical power of the initial word frequency test. As a rough guideline, the following minimum numbers of texts are recommended.

- 10,000 is the minimum for any kind of analysis, or 5,000 if there is also a larger reference set. This will only be enough if they are tightly focused on a topic. More will be needed for loose topics where the texts cover many issues.

- 100,000 is the recommended minimum number of texts to analyze in normal circumstances, or 50,000 if there is also a larger reference set. If this is not possible then another analysis approach may be preferable.

2.2 DATA COLLECTION METHODS

2.2.1 QUERY-BASED POST COLLECTION (TWITTER)

A common way to start a social media analysis project with Mozdeh or any other analytics tool is to build a set of queries to match the topic investigated. In the easiest case, the query might be defined by the project itself. For example, an analysis of a Twitter campaign based around one or more hashtags would use those hashtags as the queries (Makita et al., 2020; Potts and Radford, 2019). In this case, the decision to use the hashtags is straightforward if most tweets within the campaign are expected to use them. A distinctive and almost unique word may also be a suitable query if it is reasonable to expect most relevant tweets to contain it. For example, the single query BRCA has been used to focus on the cancer community around the BRCA gene mutation (Vicari, 2020), variants of the name Intizar for a study of reactions in Turkey to her outing as LGBTQI+