© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Contents

Preface xvii
Prerequisites and Notation xvii
Uses for This Book xviii
What This Book Is Not xix

PART I PRELIMINARIES 1

CHAPTER 1 Introduction 3
1.1 How This Book Informs the Social Sciences 5
1.2 How This Book Informs the Digital Humanities 8
1.3 How This Book Informs Data Science in Industry and Government 9
1.4 A Guide to This Book 10
1.5 Conclusion 11

CHAPTER 2 Social Science Research and Text Analysis 13
2.1 Discovery 15
2.2 Measurement 16
2.3 Inference 17
2.4 Social Science as an Iterative and Cumulative Process 17
2.5 An Agnostic Approach to Text Analysis 18
2.6 Discovery, Measurement, and Causal Inference: How the Chinese Government Censors Social Media 20
2.7 Six Principles of Text Analysis 22
2.7.1 Social Science Theories and Substantive Knowledge are Essential for Research Design 22
2.7.2 Text Analysis does not Replace Humans—It Augments Them 24
2.7.3 Building, Refining, and Testing Social Science Theories Requires Iteration and Cumulation 26
2.7.4 Text Analysis Methods Distill Generalizations from Language 28
2.7.5 The Best Method Depends on the Task 29

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

2.7.6 Validations are Essential and Depend on the Theory and the Task 30

2.8 Conclusion: Text Data and Social Science 32

PART II SELECTION AND REPRESENTATION 33

CHAPTER 3 Principles of Selection and Representation 35

3.1 Principle 1: Question-Specific Corpus Construction 35

3.2 Principle 2: No Values-Free Corpus Construction 36

3.3 Principle 3: No Right Way to Represent Text 37

3.4 Principle 4: Validation 38

3.5 State of the Union Addresses 38

3.6 The Authorship of the Federalist Papers 39

3.7 Conclusion 40

CHAPTER 4 Selecting Documents 41

4.1 Populations and Quantities of Interest 42

4.2 Four Types of Bias 43

4.2.1 Resource Bias 43

4.2.2 Incentive Bias 44

4.2.3 Medium Bias 44

4.2.4 Retrieval Bias 45

4.3 Considerations of "Found Data" 46

4.4 Conclusion 46

CHAPTER 5 Bag of Words 48

5.1 The Bag of Words Model 48

5.2 Choose the Unit of Analysis 49

5.3 Tokenize 50

5.4 Reduce Complexity 52

5.4.1 Lowercase 52

5.4.2 Remove Punctuation 52

5.4.3 Remove Stop Words 53

5.4.4 Create Equivalence Classes (Lemmatize/Stem) 54

5.4.5 Filter by Frequency 55

5.5 Construct Document-Feature Matrix 55

5.6 Rethinking the Defaults 57

5.6.1 Authorship of the Federalist Papers 57

5.6.2 The Scale Argument against Preprocessing 58

5.7 Conclusion 59

CHAPTER 6 The Multinomial Language Model 60

6.1 Multinomial Distribution 61

6.2 Basic Language Modeling 63

6.3 Regularization and Smoothing 66

6.4 The Dirichlet Distribution 66

6.5 Conclusion 69

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

CHAPTER 7 The Vector Space Model and Similarity Metrics 70
7.1 Similarity Metrics 70
7.2 Distance Metrics 73
7.3 tf-idf Weighting 75
7.4 Conclusion 77

CHAPTER 8 Distributed Representations of Words 78
8.1 Why Word Embeddings 79
8.2 Estimating Word Embeddings 81
8.2.1 The Self-Supervision Insight 81
8.2.2 Design Choices in Word Embeddings 81
8.2.3 Latent Semantic Analysis 82
8.2.4 Neural Word Embeddings 82
8.2.5 Pretrained Embeddings 84
8.2.6 Rare Words 84
8.2.7 An Illustration 85
8.3 Aggregating Word Embeddings to the Document Level 86
8.4 Validation 87
8.5 Contextualized Word Embeddings 88
8.6 Conclusion 89

CHAPTER 9 Representations from Language Sequences 90
9.1 Text Reuse 90
9.2 Parts of Speech Tagging 91
9.2.1 Using Phrases to Improve Visualization 92
9.3 Named-Entity Recognition 94
9.4 Dependency Parsing 95
9.5 Broader Information Extraction Tasks 96
9.6 Conclusion 97

PART III DISCOVERY 99

CHAPTER 10 Principles of Discovery 103
10.1 Principle 1: Context Relevance 103
10.2 Principle 2: No Ground Truth 104
10.3 Principle 3: Judge the Concept, Not the Method 105
10.4 Principle 4: Separate Data Is Best 106
10.5 Conceptualizing the US Congress 106
10.6 Conclusion 109

CHAPTER 11 Discriminating Words 111
11.1 Mutual Information 112
11.2 Fightin' Words 115
11.3 Fictitious Prediction Problems 117
11.3.1 Standardized Test Statistics as Measures of Separation 118

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

11.3.2 χ^2 Test Statistics 118
11.3.3 Multinomial Inverse Regression 121
11.4 Conclusion 121

CHAPTER 12 Clustering 123
12.1 An Initial Example Using k-Means Clustering 124
12.2 Representations for Clustering 127
12.3 Approaches to Clustering 127
12.3.1 Components of a Clustering Method 128
12.3.2 Styles of Clustering Methods 130
12.3.3 Probabilistic Clustering Models 132
12.3.4 Algorithmic Clustering Models 134
12.3.5 Connections between Probabilistic and Algorithmic Clustering 137
12.4 Making Choices 137
12.4.1 Model Selection 137
12.4.2 Careful Reading 140
12.4.3 Choosing the Number of Clusters 140
12.5 The Human Side of Clustering 144
12.5.1 Interpretation 144
12.5.2 Interactive Clustering 144
12.6 Conclusion 145

CHAPTER 13 Topic Models 147
13.1 Latent Dirichlet Allocation 147
13.1.1 Inference 149
13.1.2 Example: Discovering Credit Claiming for Fire Grants in Congressional Press Releases 149
13.2 Interpreting the Output of Topic Models 151
13.3 Incorporating Structure into LDA 153
13.3.1 Structure with Upstream, Known Prevalence Covariates 154
13.3.2 Structure with Upstream, Known Content Covariates 154
13.3.3 Structure with Downstream, Known Covariates 156
13.3.4 Additional Sources of Structure 157
13.4 Structural Topic Models 157
13.4.1 Example: Discovering the Components of Radical Discourse 159
13.5 Labeling Topic Models 159
13.6 Conclusion 160

CHAPTER 14 Low-Dimensional Document Embeddings 162
14.1 Principal Component Analysis 162
14.1.1 Automated Methods for Labeling Principal Components 163
14.1.2 Manual Methods for Labeling Principal Components 164

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

14.1.3 Principal Component Analysis of Senate Press Releases 164
14.1.4 Choosing the Number of Principal Components 165
14.2 Classical Multidimensional Scaling 167
14.2.1 Extensions of Classical MDS 168
14.2.2 Applying Classical MDS to Senate Press Releases 168
14.3 Conclusion 169

PART IV MEASUREMENT 171

CHAPTER 15 Principles of Measurement 173
15.1 From Concept to Measurement 174
15.2 What Makes a Good Measurement 174
15.2.1 Principle 1: Measures should have Clear Goals 175
15.2.2 Principle 2: Source Material should Always be Identified and Ideally Made Public 175
15.2.3 Principle 3: The Coding Process should be Explainable and Reproducible 175
15.2.4 Principle 4: The Measure should be Validated 175
15.2.5 Principle 5: Limitations should be Explored, Documented and Communicated to the Audience 176
15.3 Balancing Discovery and Measurement with Sample Splits 176

CHAPTER 16 Word Counting 178
16.1 Keyword Counting 178
16.2 Dictionary Methods 180
16.3 Limitations and Validations of Dictionary Methods 181
16.3.1 Moving Beyond Dictionaries: Wordscores 182
16.4 Conclusion 183

CHAPTER 17 An Overview of Supervised Classification 184
17.1 Example: Discursive Governance 185
17.2 Create a Training Set 186
17.3 Classify Documents with Supervised Learning 186
17.4 Check Performance 187
17.5 Using the Measure 187
17.6 Conclusion 188

CHAPTER 18 Coding a Training Set 189
18.1 Characteristics of a Good Training Set 190
18.2 Hand Coding 190
18.2.1 1: Decide on a Codebook 191
18.2.2 2: Select Coders 191
18.2.3 3: Select Documents to Code 191
18.2.4 4: Manage Coders 192
18.2.5 5: Check Reliability 192

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

18.2.6 Managing Drift 192
18.2.7 Example: Making the News 192
18.3 Crowdsourcing 193
18.4 Supervision with Found Data 195
18.5 Conclusion 196

CHAPTER 19 Classifying Documents with Supervised Learning 197
19.1 Naive Bayes 198
19.1.1 The Assumptions in Naive Bayes are Almost Certainly Wrong 200
19.1.2 Naive Bayes is a Generative Model 200
19.1.3 Naive Bayes is a Linear Classifier 201
19.2 Machine Learning 202
19.2.1 Fixed Basis Functions 203
19.2.2 Adaptive Basis Functions 205
19.2.3 Quantification 206
19.2.4 Concluding Thoughts on Supervised Learning with Random Samples 207
19.3 Example: Estimating Jihad Scores 207
19.4 Conclusion 210

CHAPTER 20 Checking Performance 211
20.1 Validation with Gold-Standard Data 211
20.1.1 Validation Set 212
20.1.2 Cross-Validation 213
20.1.3 The Importance of Gold-Standard Data 213
20.1.4 Ongoing Evaluations 214
20.2 Validation without Gold-Standard Data 214
20.2.1 Surrogate Labels 214
20.2.2 Partial Category Replication 215
20.2.3 Nonexpert Human Evaluation 215
20.2.4 Correspondence to External Information 215
20.3 Example: Validating Jihad Scores 216
20.4 Conclusion 217

CHAPTER 21 Repurposing Discovery Methods 219
21.1 Unsupervised Methods Tend to Measure Subject Better than Subtleties 219
21.2 Example: Scaling via Differential Word Rates 220
21.3 A Workflow for Repurposing Unsupervised Methods for Measurement 221
21.3.1 1: Split the Data 223
21.3.2 2: Fit the Model 223
21.3.3 3: Validate the Model 223
21.3.4 4: Fit to the Test Data and Revalidate 225
21.4 Concerns in Repurposing Unsupervised Methods for Measurement 225
21.4.1 Concern 1: The Method Always Returns a Result 226

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

21.4.2 Concern 2: Opaque Differences in Estimation Strategies 226
21.4.3 Concern 3: Sensitivity to Unintuitive Hyperparameters 227
21.4.4 Concern 4: Instability in results 227
21.4.5 Rethinking Stability 228
21.5 Conclusion 229

PART V INFERENCE 231

CHAPTER 22 Principles of Inference 233
22.1 Prediction 233
22.2 Causal Inference 234
22.2.1 Causal Inference Places Identification First 235
22.2.2 Prediction Is about Outcomes That Will Happen, Causal Inference is about Outcomes from Interventions 235
22.2.3 Prediction and Causal Inference Require Different Validations 236
22.2.4 Prediction and Causal Inference Use Features Differently 237
22.3 Comparing Prediction and Causal Inference 238
22.4 Partial and General Equilibrium in Prediction and Causal Inference 238
22.5 Conclusion 240

CHAPTER 23 Prediction 241
23.1 The Basic Task of Prediction 242
23.2 Similarities and Differences between Prediction and Measurement 243
23.3 Five Principles of Prediction 244
23.3.1 Predictive Features do not have to Cause the Outcome 244
23.3.2 Cross-Validation is not Always a Good Measure of Predictive Power 244
23.3.3 It's Not Always Better to be More Accurate on Average 246
23.3.4 There can be Practical Value in Interpreting Models for Prediction 247
23.3.5 It can be Difficult to Apply Prediction to Policymaking 247
23.4 Using Text as Data for Prediction: Examples 249
23.4.1 Source Prediction 249
23.4.2 Linguistic Prediction 253
23.4.3 Social Forecasting 254
23.4.4 Nowcasting 256
23.5 Conclusion 257

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

CHAPTER 24 Causal Inference 259
24.1 Introduction to Causal Inference 260
24.2 Similarities and Differences between Prediction and Measurement, and Causal Inference 263
24.3 Key Principles of Causal Inference with Text 263
24.3.1 The Core Problems of Causal Inference Remain, even when Working with Text 263
24.3.2 Our Conceptualization of the Treatment and Outcome Remains a Critical Component of Causal Inference with Text 264
24.3.3 The Challenges of Making Causal Inferences with Text Underscore the Need for Sequential Science 264
24.4 The Mapping Function 266
24.4.1 Causal Inference with g 267
24.4.2 Identification and Overfitting 268
24.5 Workflows for Making Causal Inferences with Text 269
24.5.1 Define g before Looking at the Documents 269
24.5.2 Use a Train/Test Split 269
24.5.3 Run Sequential Experiments 271
24.6 Conclusion 271

CHAPTER 25 Text as Outcome 272
25.1 An Experiment on Immigration 272
25.2 The Effect of Presidential Public Appeals 275
25.3 Conclusion 276

CHAPTER 26 Text as Treatment 277
26.1 An Experiment Using Trump's Tweets 279
26.2 A Candidate Biography Experiment 281
26.3 Conclusion 284

CHAPTER 27 Text as Confounder 285
27.1 Regression Adjustments for Text Confounders 287
27.2 Matching Adjustments for Text 290
27.3 Conclusion 292

PART VI CONCLUSION 295

CHAPTER 28 Conclusion 297
28.1 How to Use Text as Data in the Social Sciences 298
28.1.1 The Focus on Social Science Tasks 298
28.1.2 Iterative and Sequential Nature of the Social Sciences 298
28.1.3 Model Skepticism and the Application of Machine Learning to the Social Sciences 299

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

28.2 Applying Our Principles beyond Text Data **299**
28.3 Avoiding the Cycle of Creation and Destruction in Social Science Methodology **300**

Acknowledgments 303

Bibliography 307

Index 331

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

CHAPTER 1

Introduction

This is a book about the use of texts and language to make inferences about human behavior. Our framework for using text as data is aimed at a wide variety of audiences—from informing social science research, offering guidance for researchers in the digital humanities, providing solutions to problems in industry, and addressing issues faced in government. This book is relevant to such a wide range of scholars and practitioners because language is an important component of social interaction—it is how laws are recorded, religious beliefs articulated, and historical events reported. Language is also how individuals voice complaints to representatives, organizers appeal to their fellow citizens to join in protest, and advertisers persuade consumers to buy their product. And yet, quantitative social science research has made surprisingly little use of texts—until recently.

Texts were used sparingly because they were cumbersome to work with at scale. It was difficult to acquire documents because there was no clear way to collect and transcribe all the things people had written and said. Even if the texts could be acquired, it was impossibly time consuming to read collections of documents filled with billions of words. And even if the reading were possible, it was often perceived to be an impossible task to organize the texts into relevant categories, or to measure the presence of concepts of interest. Not surprisingly, texts did not play a central role in the evidence base of the social sciences. And when texts were used, the usage was either in small datasets or as the product of massive, well-funded teams of researchers.

Recently, there has been a dramatic change in the cost of analyzing large collections of text. Social scientists, digital humanities scholars, and industry professionals are now routinely making use of document collections. It has become common to see papers that use millions of social media messages, billions of words, and collections of books larger than the world's largest physical libraries. Part of this change has been technological. With the rapid expansion of the internet, texts became much easier to acquire. At the same time, computational power increased—laptop computers could handle computations that previously would require servers. And part of the change was also methodological. A burgeoning literature—first in computer science and computational linguistics, and later in the social sciences and digital humanities—developed tools, models, and software that facilitated the analysis and organization of texts at scale.

Almost all of the applications of large-scale text analysis in the social sciences use algorithms either first developed in computer science or built closely on those developments. For example, numerous papers within political science—including many of our

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

own—build on topic models (Blei, Ng, and Jordan, 2003; Quinn et al., 2010; Grimmer, 2010; Roberts et al., 2013) or use supervised learning algorithms for document classification (Joachims, 1998; Jones, Wilkerson, and Baumgartner, 2009; Stewart and Zhukov, 2009; Pan and Chen, 2018; Barberá et al., 2021). Social scientists have also made methodological contributions themselves, and in this book we will showcase many of these new models designed to accomplish new types of tasks. Many of these contributions have even flowed from the social sciences to computer science. Statistical models used to analyze roll call votes, such as Item Response Theory models, are now used in several computer science articles (Clinton, Jackman, and Rivers, 2004; Gerrish and Blei, 2011; Nguyen et al., 2015). Social scientists have broadly adapted the tools and techniques of computer scientists to social science questions.

However, the knowledge transfer from computer science and related fields has created confusion in how text as data models are applied, how they are validated, and how their output is interpreted. This confusion emerges because tasks in academic computer science are different than the tasks in social science, the digital humanities, and even parts of industry. While computer scientists are often (but not exclusively!) interested in information retrieval, recommendation systems, and benchmark linguistic tasks, a different community is interested in using "text as data" to learn about previously studied phenomena such as in social science, literature, and history. Despite these differences of purpose, text as data practitioners have tended to reflexively adopt the guidance from the computer science literature when doing their own work. This blind importing of the default methods and practices used to select, evaluate, and validate models from the computer science literature can lead to unintended consequences.

This book will demonstrate how to treat "text as data" for *social science tasks* and *social science problems*. We think this perspective can be useful beyond just the social sciences in the digital humanities, industry, and even mainstream computer science. We organize our argument around the core tasks of social science research: *discovery*, *measurement*, *prediction*, and *causal inference*. Discovery is the process of creating new conceptualizations or ways to organize the world. Measurement is the process where concepts are connected to data, allowing us to describe the prevalence of those concepts in the real world. These measures are then used to make a causal inference about the effect of some intervention or to predict values in the future. These tasks are sometimes related to computer science tasks that define the usual way to organize machine learning books. But as we will see, the usual distinctions made between particular types of algorithms—such as supervised and unsupervised—can obscure the ways these tools are employed to accomplish social science tasks.

Building on our experience developing and applying text as data methods in the social sciences, we emphasize a sequential, iterative, and inductive approach to research. Our experience has been that we learn the most in social science when we refine our concepts and measurements iteratively, improving our own understanding of definitions as we are exposed to new data. We also learn the most when we consider our evidence sequentially, confirming the results of prior work, then testing new hypotheses, and, finally, generating hypotheses for future work. Future studies continue the pattern, confirming the findings from prior studies, testing prior speculations, and generating new hypotheses. At the end of the process, the evidence is aggregated to summarize the results and to clarify what was learned. Importantly, this process doesn't happen within the context of a single article or book, but across a community of collaborators.

This inductive method provides a principled way to approach research that places a strong emphasis on an evolving understanding of the process under study. We call this

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

understanding theory—explanations of the systematic facets of social process. This is an intentionally broad definition encompassing formal theory, political/sociological theory, and general subject-area expertise. At the core of this book is an argument that scholars can learn a great deal about human behavior from texts but that to do so requires an engagement with the context in which those texts are produced. A deep understanding of the social science context will enable researchers to ask more important and impactful questions, ensure that the measures they extract are valid, and be more attentive to the practical and ethical implications of their work.

We write this book now because the use of text data is at a critical point. As more scholars adopt text as data methods for their research, a guide is essential to explain how text as data work in the social sciences differs from its work in computer science. Without such a guide, researchers outside of computer science solving problems run the risk of applying the wrong algorithms, validating the wrong quantities, and ultimately making inferences not justified by the evidence they have acquired.

We also focus on texts because they are an excellent vehicle for learning about recent advances in machine learning. The argument that we make in this book about how to organize social science research applies beyond texts. Indeed, we view our approach as useful for social science generally, but particularly in any application where researchers are using large-scale data to discover new categories, measure their prevalence, and then to assess their relationships in the world.

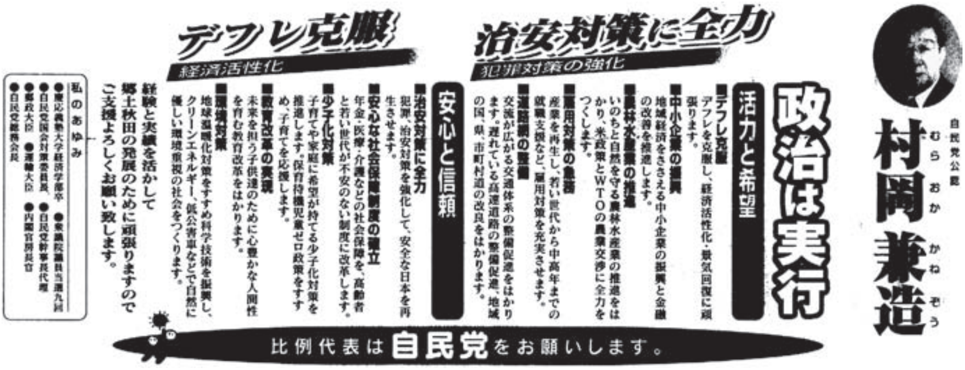
1.1 How This Book Informs the Social Sciences

A central argument of this book is that the goal of text as data research differs from the goals of computer science work. Fortunately, this difference is not so great that many of the tools and ideas first developed in other fields cannot be applied to text as data problems. It does imply, however, that we have to think more carefully about what we learn from applying those models.

To help us make our case, consider the use of texts by political scientist Amy Catalinac (Catalinac, 2016*a*)—a path-breaking demonstration of how electoral district structure affects political candidates' behavior. We focus on this book because the texts are used clearly, precisely, and effectively to make a social science point, even though the algorithm used to conduct the analysis comes from a different discipline. And importantly, the method for validation used is distinctively social scientific and thorough.

Catalinac's work begins with a puzzle: why have Japanese politicians allocated so much more attention to national security and foreign policy after 1997, despite significant social, political, and government constraints on the use of military and foreign policy discussions put in place after World War II? Catalinac (2016*a*) argues that a 1994 reform in how Japanese legislators are elected explains the change because it fundamentally altered the *incentives* that politicians face. Before the 1994 reform, Japanese legislators were elected through a system where each district was represented by multiple candidates and each party would run several candidates in each district trying to get the majority of the seats. Because multiple candidates from the same party couldn't effectively compete with their co-partisans on ideological issues, representatives tried to secure votes by delivering the most pork—spending that has only local impact, such as for building a bridge—to the district as possible. The new post-1994 reform system eliminated multi-member districts and replaced them with a parallel system: single-member districts—where voters cast their ballot for a candidate—and representatives

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.



自民党公認

村岡兼造

むらおか かねぞう

政治は実行

活力と希望

■デフレ克服
デフレを克服し、経済活性化・景気回復に頑張ります。

■中小企業の振興
地域経済をささえる中小企業の振興と金融の改善を推進します。

■農林水産業の推進
いのちと自然を守る農林水産業の推進をはかり、米政策とＷＴＯの農業交渉に全力をつくします。

■雇用対策の急務
産業を再生し、若い世代から中高年までの就職支援など、雇用対策を充実させます。

■道路網の整備
交流が広がる交通体系の整備促進をはかります。遅れている高速道路の整備促進、地域の国、県、市町村道の改良をはかります。

治安対策に全力

犯罪対策の強化

安心と信頼

■治安対策に全力
犯罪、治安対策を強化して、安全な日本を再生させます。

■安心な社会保障制度の確立
年金・医療・介護などの社会保障を、高齢者と若い世代が不安のない制度に改革します。

■少子化対策
子育てや家庭に希望が持てる少子化対策を推進します。保育待機児童ゼロ政策をすすめ、子育てを応援します。

■教育改革の実現
未来を担う子供達のために心豊かな人間性を育む教育改革をはかります。

■環境対策
地球温暖化対策をすすめ科学技術を振興し、クリーンエネルギー、低公害車などで自然に優しい環境重視の社会をつくります。

デフレ克服！

経済活性化

経験と実績を活かして
郷土秋田の発展のために頑張りますので
ご支援よろしくお願い致します。

私のあゆみ
●慶応義塾大学経済学部卒 ●衆議院議員当選九回
●自民党国会対策委員長、●自民党幹事長代理
●郵政大臣 ●運輸大臣 ●内閣官房長官
●自民党総務会長

比例代表は自民党をお願いします。

Figure 1.1. An example of a candidate manifesto of Kanezo Muraoka from 2003, Figure 3.7 from Catalinac (2016*a*).

for the whole country—where voters cast their ballot for a party and the elected officials are chosen from the party's list. This new system allowed the parties to impose stricter ideological discipline on their members and the choices of voters became less about the individual personalities and more about party platforms. Thus, the argument goes, the reform changed the legislators' *incentives*. Focusing on local issues like pork was now less advantageous than focusing on national issues like foreign policy.

The argument proceeds through iteration and induction. To begin understanding the effect of the change in electoral rules on electoral strategy, Catalinac collected an original dataset of 7,497 Japanese Diet candidate manifestos. The manifestos are nearly ideal data for her study: they are important to candidates and voters, under the control of candidates, and available for all candidates for all elections for a period before and after the shift in electoral rules. We discuss the principles for data collection in Chapter 4, but Catalinac's exemplary work shows that working with text data does not mean that we must opt for the most convenient data. Rather, Catalinac engaged in a painstaking data collection process to find the manifestos through archival visits and digitize them through manual transcription. This process alone took years.

With the data in hand, Catalinac uses an inductive approach to learn the categories in her data she needs to investigate her empirical puzzle: what elected officials are discussing when they run for office. Catalinac uses a well-known statistical model, *Latent Dirichlet Allocation* (LDA)—which we return to in Chapter 13—to discover an underlying set of topics and to measure the proportion of each manifesto that belongs to each topic. As Catalinac describes,

> Typically, the model is fit iteratively. The researcher sets some number of topics; runs the model; ascertains the nature of the topics outputted by reading the words and documents identified as having high probabilities of belonging to each of the topics; and decides whether or not those topics are substantively meaningful.… My approach was also iterative and guided by my hypotheses. (Catalinac, 2016*a*, p. 84)

As we describe in Chapter 4, discovery with text data does not mean that we begin with a blank slate. Catalinac's prior work, qualitative interviews, and expertise in Japenese politics helped to shape the discoveries she made in the text. We can bring

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

this prior knowledge to bear in discovery; theory and hunches play a role in defining our categories, but so too does the data itself.

Catalinac uses the model fit from LDA to measure the prevalence of candidates' discussions of pork, policy, and other categories of interest. To establish which topics capture these categories, Catalinac engages in extensive validation. Importantly, her validations are not the validations most commonly conducted in computer science, where LDA originated. Those validations tend to focus on how LDA functions as a language model—that is, how well it is able to predict unseen words in a document. For Catalinac's purposes, it isn't important that the model can predict unseen words—she has all the words! Instead, her validations are designed to demonstrate that her model has uncovered an organization that is interesting and useful for her particular social scientific task: assessing how a change in the structure of districts affected the behavior of candidates and elected officials. Catalinac engages in two broad kinds of validation. First, she does an in-depth analysis of the particular topics that the model automatically discovers, reading both the high probability words the model assigns to the topic and the manifestos the model indicates are most aligned with each topic. This analysis assures the reader that her labels and interpretations of the computer-discovered topics are both valid and helpful for her social scientific task. Second, she shows that her measures align with well-known facts about Japanese politics. This step ensures that the measures that come from the manifestos are not idiosyncratic or reflecting a wildly different process than that studied in other work. It also provides further evidence that the labels Catalinac assigns to texts are valid reflections of the content of those texts.

Of course, Catalinac is not interested in just categorizing the texts for their own sake—she wants to use the categories assigned to the texts as a source of data to learn about the world. In particular, she wants to estimate the causal effect of the 1994 electoral reform on the shift in issues discussed by candidates when they are running. To do this, she uses her validated model and careful research design to pursue her claim that the electoral reform causes average candidates to shift from a focus on pork to a focus on national security. This is a particularly challenging setting for causal inference, because the reform changes across all districts at the same time. After showing that, in practice, there is a substantial increase in the discussion of national security following the 1994 reforms, Catalinac moves to rule out alternative explanations. She shows that there is no sudden influx of candidates that we would expect to discuss national security. Nor, she argues, does this increase in the importance of national security merely reflect an ideological shift in the parties. And she argues that there is no evidence that voters suddenly want candidates who prioritize national security.

Our brief examination of Catalinac (2016*a*) reveals how sequence, iteration, and induction can lead to substantively interesting and theoretically important research. Further, Catalinac illustrates a point that we will return to throughout the book, that validations for text as data research are necessary and look quite different from validations in computer science. Rather than a focus on prediction, text as data researchers are much more interested in how well their models provide insights into concepts of interest, how well measurement tools sort documents according to those rules, and how well the assumptions needed for accurate causal inference or prediction are met. These points travel well beyond political science, to other social scientists studying human behavior including sociology (DiMaggio, 2015; Evans and Aceves, 2016; Franzosi, 2004), economics (Gentzkow, Kelly, and Taddy, 2019), psychology (Schwartz et al., 2013), and law (Livermore and Rockmore, 2019).

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

1.2 How This Book Informs the Digital Humanities

Our view of how to apply text as data methods was developed and refined through our experience with social science research. But we will argue that our approach to text as data can provide useful insights into other fields as well. In parallel to the meteoric rise of text as data methods within the social sciences, there has been rapidly growing interest in using computational tools to study literature, history, and the humanities more generally. This burgeoning field, termed *Digital Humanities*, shares much in common with text as data in the social sciences in that it draws on computational tools to answer classic questions in the field.

The use of text as data methods has drawn considerable funding and has already made impressive contributions to the study of literature (Jockers, 2013; Piper, 2018; Underwood, 2019). Computational tools have been used to study the nature of genres (Rybicki and Eder, 2011), poems (Long and So, 2016), the contours of ideas (Berry and Fagerjord, 2017), and many other things (Moretti, 2013). To reach their conclusions, scholars working in this area follow many of the same procedures and use similar tools to those in the social sciences. They represent their texts using numbers and then apply models or algorithms that originate in other fields to reach substantive conclusions.

Even though scholars in the Digital Humanities (DH) come from a humanistic tradition, we will show how the goals of their analysis fit well within the framework of our book. And as a result, our argument about how to use text as data methods to make valid inferences will cover many of the applications of computational tools in the humanistic fields. A major difference between DH and the social sciences is that digital humanists are often interested in inferences about the particular text that is being studied, rather than the text as an indicator of some other, larger process. As a result, digital humanities have thus far tended to focus on the discovery and measurement steps of the research process, while devoting less attention to making causal inferences or predictions. Digital humanists use their large corpora to make new and important discoveries about organizations in their texts. They then use tools to measure the prevalence of those quantities, to describe how the prevalence of the characteristics has changed over time, or to measure how well defined a category is over time.

As with any field that rises so suddenly, there has been considerable dissent about the prospect of the digital humanities. Some of this dissent lies well outside of the scope of our book and focuses on the political and epistemological consequences of opening up the humanities to computational tools. Instead we will engage with other critiques of digital humanities that stipulate to the “rules” laid out in computational papers. These critics argue that the digital humanities is not capable of achieving the inferential goals it lays out and therefore the analysis is doomed from the start. A recent and prominent objection comes from Da (2019), who summarizes her own argument as,

> In a nutshell the problem with computational literary analysis as it stands is that what is robust is obvious (in the empirical sense) and what is not obvious is not robust, a situation not easily overcome given the nature of literary data and the nature of statistical inquiry. (Da, 2019, p. 45)

Da (2019)’s critique goes to the heart of how results are evaluated and relies heavily on procedures and best practices imported from computer science (as does, it is worth noting, much of the work she is critiquing). As we have argued above, directly

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

importing rules from other fields to studying texts in new domains can be suboptimal. When we directly import the recommendations from computer science and statistics to text-based inferences in the humanities or social sciences we might make problematic inferences, recommendations that are misguided, or misplaced assessments about the feasibility of computational analysis for a field.

Yet Da's critique is a useful foil for illuminating a key feature of our approach that departs from much of the work in the digital humanities. In Chapter 2, we offer six core principles which reflect a broader "radically agnostic" view of text as data methods. We reject the idea that models of text should be optimized to recover one true underlying, inherent organization in the texts—because, we argue, no one such organization exists. In much of the digital humanities, and Da's critique, there is an implicit assumption that the statistical models or algorithms are uncovering an ideal categorization of the data that exists outside of the research question asked and the models estimated. This approach is in tension with much of the theoretical work in the humanities, but seemingly arises because this is a motivating assumption in much of computer science and statistics, where it provides a convenient fiction for evaluating model performance.

On our account, organizations are useful if they help us to uncover a categorization of the data that is useful for answering a research question. If two models disagree on how to categorize texts, there is no sense in determining which one is any more "right" than the other. We would not, for example, want to argue that an organization of texts based on the expression of positive or negative emotion is more right than an organization based on the topic of the text. Rather, we will argue that some organizations are more useful than others for addressing a particular question. For example, we might argue that a model is particularly useful for studying genre, because it provides an organization that leads the researcher to an insight about the trajectory of books that would have been impossible otherwise. Once you have an organization, you can find the best *measurement* of that particular categorization. You can then test the measurement with extensive *validation*. But because there is a multiplicity of useful and valid organizations, a method that does not provide a "robust" answer to how texts should be organized will be less concerning than critics argue. What becomes important is the credibility of the validations once an organization has been selected and its utility in answering the research question.

We also will emphasize throughout our book that text as data methods should not displace the careful and thoughtful humanist. And there is no sense in which inferences should be made in the field of digital humanities without the reader directly involved. This emphasis on using computational methods to improve inferences will help allay some concerns about the role of digital humanities scholarship. The computational tools should not replace traditional modes of scholarship. When used well, computational tools should help provide broader context for scholars, illuminate patterns that are otherwise impossible to identify manually, and generally amplify—rather than replace—the human efforts of the scholars using them.

1.3 How This Book Informs Data Science in Industry and Government

Computational tools have also revolutionized how companies use text as data in their products and how government uses text to represent the views of constituents. The applications of these tools are nearly endless in industry. Companies use messages that users post on their website to better target advertisements, to make suggestions about

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

new content, or to help individuals connect with elected officials. In government, there is the chance to use text as data methods to better represent the views of constituents publicly commenting on proposed rule changes at bureaucratic agencies or expressing their views to elected officials.

The stakes are high when applying text as data methods to industrial-scale problems. Perhaps the most politically sensitive application of text as data methods is content moderation: the attempt by social media companies (and sometimes governments) to regulate the content that is distributed on their platform. In the wake of the Russian misinformation campaign in the 2016 US election, social media companies faced increased pressure to identify and remove misinformation from their sites, to report on the effect of misinformation that occurred during the campaign, and to demonstrate that new procedures were fair and did not disproportionately target particular ideologies. The tools used to identify this content will appear throughout this book and will draw on a similar set of computational resources that we introduce.

Beyond the questions of political sensitivity, the application of text as data methods will also be high stakes because of the large amounts of money that will be spent based on the recommendations of the systems. For example, trading firms now use computational tools to guide their investments or to quickly learn about content from central bankers. Text as data methods also help drive advertising decisions that represent a massive share of the economy. Getting these decisions "right," then, is important for many business practices.

Our book is useful for data scientists, because these tasks are inherently social science tasks. Moderating content to suppress misinformation or hate speech is fundamentally a measurement task. When companies decide which ads will cause the largest increase in sales for their clients, they are engaged in causal inference. And when traders make decisions based on the content of documents or statements from officials, they are engaged in prediction. Recognizing the omnipresence of social science within industry is essential, because many data scientists receive their professional training outside of the social sciences. These fields do an excellent job of providing the computational tools necessary for working with the massive datasets that companies create, but often fail to expose researchers to core design principles behind the tasks those tools are built for.

This book, and indeed its very organizational structure, is designed to remove focus from the individual models and computational tools and refocus on the differences between tasks like discovery and measurement or prediction and causal inference. Identifying these differences is essential, because the different tasks imply that different models should be used, different information sets should be conditioned upon, and different assumptions are needed to justify conclusions.

1.4 A Guide to This Book

Our book spans fields within the social sciences, digital humanities, computer science, industry, and government. To convey our view on how to work with text as data in these disparate fields, we provide a different organization of our book. While most computational social science books organize the manuscript around algorithms, in this book we organize the book around tasks. We focus on tasks to emphasize what is different when social scientists approach text as data research. This also enables us to explain how the same algorithm can be used to accomplish different tasks and how validations for an algorithm might differ, depending on the goal at hand when applying that algorithm.

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

We organize our book around five key tasks: representation, discovery, measurement, prediction, and causal inference. Underlying this task-based focus is a set of principles of text analysis that we outline in Chapter 2. There, we explain our *radically agnostic* approach to text as data inference. We generally reject the view that there is an underlying structure that statistical models applied to text are recovering. Rather, we view statistical models as useful (and incomplete) summaries of the text documents. This view provides us with important insights into how to validate models, how to assess models that provide different organizations, and the role of humans within the research process.

In Part 2 we discuss selection and representation: the process of acquiring texts and then representing the content quantitatively. When selecting texts, basic principles of sample selection matter a great deal, even though there is a temptation to select content that is most conveniently available. When representing texts, we explain how different representations provide different useful insights into the texts and set the stage for future models in the book.

Part 3 introduces a series of models for discovery. By discovery we mean the use of models to uncover and refine conceptualizations, or organizations of the world. We show how a wide array of models can help suggest different organizations that can help researchers gain new insights into the world. We begin with methods used to uncover words that are indicative of differences between how two groups speak. These methods can be used to compare groups of documents—for example, legislators from two different political parties—or to help label categorizations inferred from other inductive methods. We then discuss some computer-assisted techniques for discovery, including models for partitioning data that exhaustively assign each observation to a single category. We then explain how clustering methods can be extended to admixture models, which represent each document as proportionally assigned to different categories. Finally, we describe methods for embedding documents into lower-dimensional spaces, which can shed light on underlying continuous variables in the data.

Part 4 describes our approach to measurement: assessing the prevalence of documents within a set of categories or assessing their location along a predetermined spectrum. We explain how to combine human judgment with machine learning methods to extend human annotations coded in a training set to a much larger dataset. When performing measurement, we explain how a discovery method can be repurposed to measure a category of interest. We include an extensive discussion of how to validate each of these measures, no matter what method produced them.

Building on the concepts and measures we have described, Part 5 explains how to apply the methods for prediction and causal inference. First, we describe how to use text as data methods to make predictions about how the world will be in the future. We discuss different types of predictive tasks and highlight how the threats to inference may vary with the setting. Next, we describe how to use the measures from texts as either the outcome or the intervention variable to make causal inferences. We explain the particular concerns that can emerge when text methods are used and provide a set of tools for assessing when a stringent set of assumptions is met.

1.5 Conclusion

There is immense promise with text as data research. With large amounts of data, complicated models, and custom measures, there is also the possibility of using these methods and getting the research wrong. Text is complicated and meaning is often

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

subtle. The risk is that if scholars overclaim on what text methods can do, they will undermine the case for using text methods.

Our book is intended as a guide for researchers about what is feasible with text as data methods and what is infeasible. We want to help readers learn about the immense set of tasks that text as data methods can help them accomplish. At the same time, we also want to help our readers to recognize the limits of text methods. We start out on this goal in the next chapter, where we articulate the basic principles that will guide our approach to text as data research.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Index

Abello, James, 45
accuracy: confusion matrices, 212; cross-validation, 213, 245; Google Flu Trends, 256; paradox, 246; training set, 186; validation set, 212
Achen, Christopher H., 13
affinity propagation, 135
AFGP. *See* Assistance to Firefighters Grant Program (AFGP)
Aggarwal, Charu C., 207
agnostic approach, 11, 18–20; text as data methods, 30; text analysis methods, 29
Ahmed, Amr, 260
AIC. *See* Akaike Information Criterion (AIC)
Airoldi, Edoardo M., 158, 160, 260
Akaike Information Criterion (AIC), 141
à la carte embeddings, 85
Aletras, Nikolaos, 254
annotations, 11, 123, 172n2, 193, 194
APIs. *See* Application Programming Interfaces (APIs)
Application Programming Interfaces (APIs), 46
Arora, Sanjeev, 85n1, 227
artisanal data, 25
Assistance to Firefighters Grant Program (AFGP), 151
ATE. *See* average treatment effect (ATE)
Athey, Susan, 248
automatic generalization, 80
average treatment effect (ATE), 268, 272, 277, 278; causal inference, 228, 238, 267; linear regression, 262; STM model, 292; unbiased estimation of, 286

bag of words model, 48–49; default procedure, 57–59; document-feature matrix, 55–57; language creation with, 61; reduce complexity, 52–55; tokenize, 50–52; unit of analysis, 49–50
Bail, Christopher, 91, 169
Banerjee, Arindam, 133
Barberá, Pablo, 190, 191
basis functions, 204; adaptive, 205–206; fixed, 203–205
Baumgartner, Frank, 16
Bayesian hierarchical model, 66, 147
Bayesian inference, 67, 68; cluster data, 141; with Dirichlet prior, 115
Bayesian Information Criterion (BIC), 141
Bearman, Peter S., 38, 93, 169
Beauchamp, Nicholas, 183, 254–256
Belloni, Alexandre, 288
benchmarks, 255
Bengio, Yoshua, 82, 206
Berliner, Daniel, 44
Berry, Christopher R., 108
bet on sparsity principle, 205
betweenness centrality, 170
bias: incentive, 44; sample selection, 42; types of, 43–46
BIC. *See* Bayesian Information Criterion (BIC)
big data, 24, 298, 301
bigrams, 51
Bischof, Jonathan, 160
Bishop, Christopher, 207
Blaydes, Lisa, 25, 42, 157
Blei, David M., 60, 153, 155, 156, 157n5, 287
Bojanowski, Piotr, 84
Boydstun, Amber E., 192, 193, 269
Brachman, Jarret, 217
Broadwell, Peter, 45

Carlson, David, 195
Casas, Andreu, 228
Catalinac, Amy, 5–7, 41, 153, 224, 266
categorical distribution, 61
causal inferences, 15–17, 20–22, 234–238, 260–263; assumptions, 106; mapping function, 266–269; *vs.* prediction, 263; principles of, 263–266; workflows for, 269–271
CBOW. *See* continuous bag of words (CBOW) model
Chancellor, Stevie, 248
Chang, Charles, 206
Chang, Jonathan, 139, 142
Chen, Kai, 80
Chernozhukov, Victor, 288
Chinese government censors, social media, 20–22
Christian societies, 42
Chuang, Jason, 93
Classical Multidimensional Scaling (MDS), 167–169; extensions of, 168; to senate press releases, 168–169

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

classifiers, 184, 200; linear, 201–202; precision of, 247; softmax, 201
clustering algorithms, 12, 73, 123, 124, 131, 134–135, 137; affinity propagation, 135; approaches to, 127–143; document dissimilarity, 128–129; hierarchical clustering, 135–136; human side of, 144–145; *k*-means clustering, 124–127; partition quality, 129–130; probabilistic clustering models, 132–134; representations for, 127; styles of, 130–132
coders, 172, 190–192
Cohen, Mark A., 272, 273
Cointet, Jean-Philippe, 38, 93, 169
Colaresi, Michael, 115, 116, 181
community detection algorithms, 169
computational abductive analysis, 159
conditional ignorability, 286
confusion matrix, 212
congressional record, 185
conjugate prior, 67
conservative coalition, 106
contextual integrity, 37
continuous bag of words (CBOW) model, 83
convergent validity, 214
Corrado, Greg S., 80
cosine distance, 72–74
Courville, Aaron, 206
cross-validation, 213, 245

Da, Nan Z., 8–9
Danforth, Christopher M., 43, 180
Dean, Jeff, 80
deductive approach, 27
deep learning revolution, 206
Deerwester, Scott, 82
default procedure, 57; authorship of the federalist papers, 57–58; scale argument against preprocessing, 58–59
Denny, Matthew J., 19, 228
dependency parsing, 95–96
Devlin, Jacob, 88
DH. *See* digital humanities (DH)
Dhillon, Inderjit, 133
dictionary methods, 180–183; limitations and validations of, 181–182
Diermeier, Daniel, 181
digital humanities (DH), 8–9
DiMaggio, Paul, 153
Dirichlet distribution, 66–69
discovery principles: context relevance, 103–104; judge the concept, not the method, 105; no ground truth, 104–105; separate data is best, 106
discursive governance, 185
distributed representations, 79
distributional hypothesis, 78
doc2vec, 86
document-feature matrix, 55–57; down-weighting frequent words, 75; Latent Semantic Analysis (LSA), 82; Naive Bayes, 198
Dodds, Peter Sheridan, 43, 180
downstream covariate model, 156
dummy variable, 61
Duneier, Mitchell, 179, 180
Dynamic Weighted Nominal Three-Step Estimation (DW-NOMINATE) algorithm, 107, 108

Egami,Naoki, 260, 266–269, 272, 273, 276–278
Eisenstein, Jacob, 90, 207, 260
EM algorithm. *See* Expectation-Maximization (EM) algorithm
entropy, 112; conditional, 113
Erlich, Aaron, 44
Euclidean distance, 74
Expectation-Maximization (EM) algorithm, 133, 134
expressed ideology, 208
extrinsic evaluations, 87

face validity, 31
Faruqui, Manaal, 84
fastText embeddings, 84
Federalist papers, 39–41, 57–58; document-feature matrix of, 75; extraction tasks, 96; prediction, 249; stop words and common words in, 76; three-word vocabulary, 64, 71; word counts for, 64
Federal Open Market Committees (FOMC), 44
feedback loops, 248
Felter, Joseph, 217
feminist movements, 112
Fenno, Richard F., 108
fictitious prediction problems, 112, 116–118
Fightin' words, 111, 115–116
fishing, 268, 299
flat *vs.* hierarchical clustering method, 132
FOMC. *See* Federal Open Market Committees (FOMC)
Fong, Christian, 209, 260, 269, 278, 279, 281–283
Foreign Affairs, 38
Foulds, James, 225
found data, 46, 195
Fowler, Anthony, 27, 108
FPCILV. *See* Fundamental Problem of Causal Inference with Latent Variables (FPCILV)
Franco, Annie, 272, 275–276
Freedman, Michael, 104, 159
Freedom of Information Act, 43–44
frequency inverse document frequency weighting, 75. *See also* tf-idf weighting
FREX, 160
Friedman, Jerome, 205
Fundamental Problem of Causal Inference with Latent Variables (FPCILV), 268
fuzzy clustering, 131

g, 266–269
game-theoretic model, 13

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Ganesh, Ananya, 88
Garg, Nikhil, 86
Garry, John, 182, 183
generalization error, 243
Gentzkow, Matthew, 117, 118
GFT algorithm. *See* Google Flu Trends (GFT) algorithm
ghetto, 179, 180
Ghosh, Joydeep, 133
ghost stories, 45
Gill, Michael, 43, 260
Gillion, Daniel Q., 185–188, 266
Ginsberg, Jeremy, 256
GloVe, 83
Goertz, Gary, 214
gold-standard data, without validation: correspondence to external information, 215; nonexpert human evaluation, 215; partial category replication, 215; surrogate labels, 214–215
gold-standard data, with validation, 211–212; cross-validation, 213; importance of, 213–214; ongoing evaluations, 214; validation set, 212–213
Goodfellow, Ian, 206
Google Flu Trends (GFT) algorithm, 256
Google Ngrams corpus, 43
government censorship, 20
GPUs. *See* Graphics Processing Units (GPUs)
graph-cutting algorithms, 136
Graphics Processing Units (GPUs), 88
gray zone, 214
grounded theory, 14, 100, 104, 159

Hall, Andrew B, 43, 260
hand coding: check reliability, 192; codebook, 191; document selection, 191–192; drift, types of, 192; *Making the News,* 192–193; manage coders, 192; select coders, 191
Hansen, Christian, 288
harmful language, 36
Hastie, Trevor, 205
Healy, Andrew J., 27, 28
Heer, Jeffrey, 93
Holland, Paul W., 236
Hopkins, Daniel J., 24, 206, 207
Hu, Yuening, 157
Hvitfeldt, Emil, 206
hypothesis validity, 31, 215

ideal point, 107
ideological dominance assumption, 221
industrial-scale problems, 10
information extraction tasks, 96–97
International Relations (IR), 287
intrinsic evaluations, 87
IR. *See* International Relations (IR)
IRT. *See* item response theory (IRT)
Isaac, William, 248
item response theory (IRT), 220

Jaros, Kyle, 94, 95
Jarvis, Stephen, 254
Jerzak, Connor T., 207
Jones, Bryan, 16
Jordan, Michael, 201
Jursfsky, Dan, 90

Kang, Jun Seok, 254
Karell, Daniel, 104, 159
Kaufman, Aaron Russell, 300
Keith, Katherine, 96
Kelly, Bryan, 117
Keohane, R. O., 171, 176
Khodak, Mikhail, 85
King, Gary, 20, 21, 24, 124, 127, 139, 145, 171, 176, 206–207, 219
k-means clustering, 124–127
Knox, Dean, 300
Kraft, Peter, 300
Krippendorff, Klaus, 190
Kristensen, Evald Tang, 45

label fidelity, 32
labeling topic models, 159–161
Lafferty, John D., 155, 157n5
language model, 7, 63–66. *See also* multinomial language model
laplace smoothing, 166
Latent Dirichlet Allocation (LDA), 6, 7, 18, 147–153; content covariates, 154–156; downstream covariate model, 156; example, 149–151; incorporating structure into, 153–154; interpretation, 151–153; prevalence covariate topic models, 154; structure, additional sources of, 157; supervised LDA (sLDA), 156
latent semantic analysis, 82
Laver, Michael, 182, 183
LDA. *See* Latent Dirichlet Allocation (LDA)
Le, Quoc, 86
lemmatizing, 54, 58, 78
Lim, Chloe, 272, 275–276
linear classifier, 201–202
linear regression, 60, 77, 119, 137, 197, 203, 226, 227, 235, 242, 262, 263, 285, 288
linguistic prediction, 242, 253–254
loss function, 197, 202–204, 242
Loughran, Tim, 181, 269
Lu, Ying, 207
Lucas, Christopher, 155, 300
Lum, Kristian, 248
Lundberg, Ian, 37

MacRae, Duncan, 106
Madison. Airoldi, 249
Maeda, Ko, 220
Malhotra, Neil, 27, 28
Maliniak, Daniel, 285, 287, 288, 292
Manning, Christopher D., 84, 93

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

mapping function, 266–269
Martin, James, 90
Masterson, Michael, 206
Mayhew, David, 124, 126
Mcauliffe, Jon D., 156
McCallum, Andrew, 88, 153, 260
McCants, William, 217
McDonald, Bill, 181, 269
McGhee, Eric, 99
McQueen, Alison, 24, 25, 42, 157
MDS. *See* Classical Multidimensional Scaling (MDS)
means *vs.* mediods clustering method, 131–132
measurement, 9, 16–17, 20–22; model, 4, 38
measurement principles: clear goals, 175; coding process, 175; from concept to, 174; discovery and, 176–177; good measurement, 174–175; limitations, 176; source material, 175; validation, 175
medium bias, 44–45
Messing, Solomon, 151
Mikolov, Tomas, 80, 86
Mimno, David, 138, 153, 260
Minkowski distance metrics, 74
MNIR. *See* Multinomial Inverse Regression (MNIR)
Mo, Cecilia Hyunjung, 27, 28
Monroe, Burt, 115, 116, 181, 220
Montagnes, B. Pablo, 27
Montgomery, Jacob M., 139, 195
Morgan, Stephen L., 260
Mosteller, Frederick, 39–41, 64, 65, 75, 249
Mozer, Reagan, 287
multinomial inverse regression (MNIR), 121
multinomial language model: Dirichlet distribution, 66–69; language model, 63–66; multinomial distribution, 61–63; regularization and smoothing, 66
multinomials clustering model, 133
Murphy, Kevin P., 207
Muslim societies, 42
mutual information, 112–114

Nag, Manish, 153
naive assumption, 198
Naive Bayes (NB), 198–200; adaptive basis functions, 205–206; assumptions in, 200; fixed basis functions, 203–205; generative model, 200–201; linear classifier, 201–202; machine learning, 202–203; quantification, 206–207
named-entity recognition (NER), 94–95
National Football League (NFL) games, 27–28
natural language processing, 30, 34, 82, 87, 88, 94, 97, 254
Nay, John J., 254
NB. *See* Naive Bayes (NB)
negative advertisement, 23
Nelson, Laura K., 104, 112, 113, 115, 116, 119–122
NER. *See* named-entity recognition (NER)
Neuendorf, Kimberly A., 190, 192
neural word embeddings, 82–84
Ng, Andrew, 201
n-grams approach, 51
Nielsen, Richard A., 196, 207–210, 216–218, 260, 266, 285, 287, 290, 292
Nissenbaum, Helen, 37
no right way to represent text, 37
no values-free corpus construction, 36–37
nowcasting process, 231, 256–257

observational data, 21
OCR technology. *See* optical character recognition (OCR) technology
one-hot encoding, 61
O'Neil, Cathy, 248
optical character recognition (OCR) technology, 34, 50
optimization, 130
out of sample, 213, 235, 236, 239, 255, 298

Pan, Jennifer, 20, 21, 94, 95, 97
PAP. *See* pre-analysis plan (PAP)
partition quality, 129–130
parts of speech (POS) tagging, 90, 91–94
PCA. *See* principal component analysis (PCA)
Pechenick, Eitan Adam, 43, 180
Pennington, Jeffrey, 84
policy agendas, 16
population of interest, 42–43
POS tagging. *See* parts of speech (POS) tagging
potential outcomes, 261–264, 267–269, 278, 286
Powers, Ryan, 285, 287, 288, 292
pragmatic approach, 29
pre-analysis plan (PAP), 265
precision: of classifier, 247; confusion matrices, 212
prediction, 233–234; *vs.* causal inference, 238–239
predictions, 17, 26–27; basic task of, 242–243; linguistic prediction, 253–254; and measurement, 243–244; principles of, 244–248; social forecasting, 254–256; task, 184
pretrained embeddings, 84
principal component analysis (PCA), 162–167; automated methods for, 163–164; manual methods for, 164; number of, 165–167; of senate press releases, 164–165
prior distribution, 66
probabilistic, soft, mean-based, and flat clustering method (mixture of multinomials clustering model), 133
probabilistic and algorithmic clustering connections: model selection, 137–140; number of clusters, 140–143; reading the documents, 140
probabilistic approaches, 131
probabilistic clustering models, 132–134
probabilistic *vs.* algorithmic clustering method, 131
Procter, Rob, 254
Proksch, Sven-Oliver, 220, 221

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

quantification, 206–207
quantities of interest, 42–43
question-specific corpus construction, 35–36
Quinn, Kevin, 115, 116, 144, 153, 181, 260

rare words, 84–85
ReadMe algorithm, 207
recall, confusion matrices, 212
reduce complexity: create equivalence classes, 54–55; filter by frequency, 55; lowercase, 52; remove punctuation, 52–53; remove stop words, 53–54
Reilly, Shauna, 254
representations: ethical issues, 36; lower-dimensional, 28; of text, 30
research design, 22–24; full-cycle, 14; KPR's research design, 22
resource bias, 43–44
retrieval bias, 45–46
retrofitting, 84
Richey, Sean, 254
Riffe, Daniel, 190
Rule, Alix, 38, 93, 169
rule-based algorithm, 91
Rust, Roland T., 272, 273

Salganik, Matthew, 37
Schnabel, Tobias, 87
Schonhardt-Bailey, Cheryl, 44
Schrodt, Philip A., 189
Sebeok, Thomas A., 174
semantic validity, 215
Sen, Maya, 300
sensitivity analysis, 236
Shapiro, Jesse M., 118
Silge, Julia, 206
similarity metrics, 70–73. *See also* vector space model
singular value decomposition (SVD), 82, 290
skip-gram, 86
Slapin, Jonathan B., 220, 221
sLDA. *See* supervised LDA (sLDA)
Smith, David, 91
Smyth, Padhraic, 225
Socher, Richard, 84
social forecasting, 254
social media, 10, 16; Chinese government censors, 20–22
social prediction, 242
social process: systematic facets of, 5
social sciences, 5–7; iterative and cumulative process, 17–18; theories, 29
softmax classifier, 201
soft *vs.* hard clustering method, 131
SOTU. *See* State of the Union (SOTU)
source prediction, 241, 249–253
Spirling, Arthur, 19, 77, 228
Sra, Suvrit, 133
State of the Union (SOTU), 38–39
Steen, Sara, 272, 273
stemming, 54, 55
STM. *See* structural topic models (STM)
stop words, 53, 54
Stramp, Nicholas, 91
Strezhnev, Anton, 207
Strubell, Emma, 88
structural approach, 19
structural topic models (STM), 157–159, 192
supervised classification: check performance, 187; classify documents with, 186; discursive governance, 185–186; measurement, 187–188; training set, 186
supervised LDA (sLDA), 156
supervised learning, 4, 21, 207–210; Jihad scores, 207–209; machine learning, 202–207; Naive bayes, 198–202
Sutskever, Ilya, 80
SVD. *See* singular value decomposition (SVD)

Taddy, Matt, 117, 121
Tangherlini, Timothy R., 45
task-based approach, 29, 30, 298
Taylor, J. Benjamin, 254
test data, 177
text analysis, principles of: best method depends on the task, 29–30; building, refining, and testing social science theories, 26–28; essential and depend on the theory, 30–32; social science theories and substantive knowledge, 22–24; text analysis, 24–26; text analysis methods, 28–29
text reuse, 90–91
tf-idf weighting, 75–77
Thurstone, L. L., 195
Tibshirani, Robert, 205
Tingley, Dustin, 160, 227
Tkachenko, Nataliya, 254
tokenization, 50–52
topic: content of, 153; different versions of the same, 153; prevalence of, 153
topic models: labeling topic models, 159–161; Latent Dirichlet Allocation (LDA), 147–153; structural topic models, 157–159
Torres, Michelle, 300
training, 86, 88, 255; cost implications of, 88; Naive Bayes, 199, 201
training data, 177, 213, 214–216; coders, 192; harmful language in, 36; machine learning algorithm, 91; sIBP process, 282; validation, 187, 225
training set, coding: characteristics of, 190; crowd-sourcing, 193–195; hand coding, 190–193; supervision with found data, 195–196
transfer learning, 79
trigrams, 51
Twitter data, 36
Tyler, Matthew, 209, 260, 269, 278, 279, 281–283

unigrams, 51, 92, 93, 124, 255
unit of analysis, 49–50

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

unsupervised methods, 79, 219–220; concerns in repurposing, 225–229; fit the model, 223; hand coding, 223–224; hypothesis validity, 224; model-based assessments, 224–225; split the data, 223; surrogate labels, 224; test data and revalidate, 225; validate the model, 223–225; workflow for repurposing, 221–225
upstream covariate models, 153
US Congress, 106–109

validation methods, 9, 31, 38; cross-validation, 213, 245; of dictionary methods, 181–182; measurement principles, 175; word embedding estimation, 87–88
validation set, 212–213
variational inference, 133
vector space model: distance metrics, 74–75; similarity metrics, 70–73; tf-idf weighting, 75–77
Veitch, Victor, 287
Verba, S., 171, 176
virgin texts, 183
vocabulary, 50–52, 54–56, 58, 61, 64, 71, 78–80, 82; distinctive words, 144; distribution of, 158; domain-specific vocabulary, 84
von Mises-Fisher model, 129, 133–134

Wallace, David L., 39–41, 64, 65, 75, 249
Wallach, Hanna M., 25, 225
Walter, Barbara F., 285, 287, 288, 292
Wang, Yixin, 287
Westwood, Sean J., 151
Wilkerson, John, 16, 91, 228
Winship, Christopher, 260
Wittgenstein, Ludwig, 79
word2vec, 83
word counting: dictionary methods, 180–183; keyword counting, 178–180
word embedding estimation: aggregation, 86–87; contextualize, 88–89; design choices in, 81–82; illustration, 85–86; latent semantic analysis, 82; neural word embeddings, 82–84; pretrained embeddings, 84; rare words, 84–85; self-supervision insight, 81; validation, 87–88
Wordfish, 221
wordscores algorithm, 182–183
word segmentation model, 50

Xing, Eric P., 260

Yih, Wen-tau, 80
Ying, Luwei, 139

Zeps, Valdis J., 174
Zhang, Han, 97
Zweig, Geoffrey, 80
$\chi 2$ test statistics, 119–121