Introduction

Preface

We are surrounded by text. It appears in government reports, policy documents, business communications, news media, and personal correspondence. Text is not merely a record of our world—it shapes it. Laws are made through written language; identities are negotiated though text; decisions are justified through written reports; debates are transcribed onto records. The instructions that guide our technologies, the terms that govern our contracts, and the narratives that influence public opinion are all encoded in text. Even our memories are inscribed in language.

From the perspective of contemporary history, these texts form the raw material of the historical record. They document not only what happened, but how events were framed, contested, and remembered. Historiography, too, is shaped by the texts historians choose to read and analyze. As the volume and variety of text continue to grow, so does the importance of methods like text mining, which allow historians to trace patterns and interrogate how language has changed over time.

Being able to analyze and interpret text is a powerful form of literacy. Text mining is no longer a niche research method–it is an essential approach for understanding historic discourses. But simply text mining without a methodological framework is not good practice; it risks producing results without meaning, stripping language of its context, intention, and complexity. Our reason for writing *Text Mining for Historical Analysis* is to embed the techniques behind text mining within the larger conceptual framework laid out by *The Dangerous Art of Text Mining*. As its sister companion, *Text Mining for Historical Analysis* offers a practical guide for bridging historic inquiry and computational methods into a rigorous form of hybrid scholarship. This book demonstrates the relationship between computational methods and historical interpretation where theory, interpretation, and code coalesce to create a critical work.

Throughout this book, we demonstrate how computational methods can be used to support historical analysis. At the heart of this work is a commitment to broadening access to coding and text analysis. We aim to support a wider movement that makes these methods more accessible—grounded in open-source tools and a vision of research that is literate, inclusive, and collaborative.

Computational text analysis allows researchers to identify broad temporal patterns that span decades while also enabling close examination of specific moments in time. This dual capacity helps illuminate events, debates, and cultural shifts with greater precision. At the same time, these methods prompt us to reflect on how we construct meaning as analysts. They encourage us to interrogate our own interpretive frameworks and invite new ways of thinking that reshape our methodological approaches. In doing so, computational research can generate original insights, lead us down unexpected analytical paths, refine our research questions, and challenge conventional habits of historical inquiry.

We suggest that a book demonstrating the methods behind *The Dangerous Art of Text Mining* offers a unique contribution to the field of digital history, and to the digital humanities more broadly. Many excellent books teach programming (Lubanovic 2019, Matthes 2023, Zumel 2019). The same is true for historical analysis

(Wineburg 2021, Fischer 1970). However, few resources exist that demonstrate the hybridization of the two. Yet, analysts have much to gain in this approach.

In addition to offering a useful conceptual framework, *Text Mining for Historical Analysis* is designed to accelerate the learning process in a practical way. It provides a targeted, accessible introduction to digital history that guides readers toward technical proficiency. These dual aims—conceptual clarity and practical utility—also distinguish *Text Mining for Historical Analysis* from other programming books.

Whereas many programming texts are structured around methods developed for numerical datasets, our book enables analysts to begin directly within the domain of history. This is important because textual data, unlike numerical data, must first be processed and transformed before it can be quantified. When that text is historical in nature, additional care must be taken to interpret it in light of its historical context, meaning, and conventions.

The way data is processed—especially historical text—deeply shapes how we engage with it and, ultimately, how we understand the past. The challenges associated with processing both text and historical sources are unique, and our book helps analysts navigate them from the outset.

At the time of writing, few resources offer an extensive introduction that guides analysts from foundational concepts to more advanced and relevant applications within history. Over several years of teaching students at our respective universities, we came to the conclusion that it was vital that analysts wishing to learn these techniques have access to a deliberate, encompassing series of chapters that applies the techniques of computational text analysis to the concerns of History. We shared drafts of the manuscript and the accompanying code with our colleagues and students, and these drafts have received the accolades of several cohorts of users for their transparency.

Many of the exercises in this book are inspired by the *The Dangerous Art of Text Mining* (2023), but have been rewritten with the purpose of offering a point of entry into the practice. This book primarily teaches computational text analysis using the R programming language, with additional examples that explore large language models (LLMs) and chatbot-based approaches like ChatGPT. While many exercises within *Text Mining for Historical Analysis* are applied to British history, and some to the United States Congressional Records, the methods in this book are intended to extend to other questions across historic domains, and even into other fields in which analysts value the concept of change over time. Most of our chapters focus on the 19th-century Hansard corpus, which is a digitized collection of the official transcripts of debates from the UK Parliament (House of Commons and House of Lords) during the 1800s. However the final chapters of this book engage the contemporary United States Congressional Records, and the final chapter of our book extends beyond any singular historic corpus and provides resources for engaging with the broader world of data–such as born-digital archives, open data repositories, and large-scale digital corpora–equipping readers with the perspective needed to navigate, interpret, and make meaningful use of these many data resources that are at the tip of an analyst's fingers. Scholars of different types of historic text–such as legal texts, hip hop lyrics, or records on global affairs–will therefore discover that the methods behind digital history can be generalized to other studies.

Text Mining for Historical Analysis emphasizes the importance of critically reflecting on how different methodological approaches shape what we can see in a corpus. Each approach may illuminate certain features while concealing others. Our book helps analysts recognize and navigate these trade-offs by combining quantitative analysis with close reading of primary sources. We show how linguistic patterns—visible in distant-reading visualizations—are rooted in broader historical structures of meaning, such as discourses and contested concepts. By incorporating contextual historical knowledge into text mining, we demonstrate how analysts can more accurately interpret linguistic trends and generate historically grounded insights.

It is therefore our hope that *Text Mining for Historical Analysis* supplies its readers with a perspective that incites a discerning sense of discovery.

A General Method for Digital History

Historical analysis is a way of understanding how individuals and groups change over time. It involves reflecting on change across a defined time span, which may range from weeks to centuries, depending on the research question at hand, as well as the quality and coherence of the available evidence (such as records, documents, logs, and other historical artifacts). This makes historical analysis a valuable tool for studying the development of political, corporate, or cultural institutions, as well as communities of practice. The theoretical foundation of historical analysis can also inform any work beyond the domain of history that requires critical thinking about the social and cultural dimensions of data.

Historical change is generally understood in terms of events that have a path and impact later episodes of life and thought. These events divide experience into eras or periods that share certain features which can be qualitatively and quantitatively analyzed. But historical analysis is also dense and varied. Individuals may anticipate future events or defy the direction of history in their time. Historical analysis may lead us towards questions of cause and effect, in which we try to unpack the relative impact of individuals on their respective futures. Or historical analysis may instead turn backwards and register the changing impact of memory as contemporary actors reflected on events in their respective past. One of the major challenges for researchers studying history is investigating the categories of temporal experience — including event (what happened), period (when it happened and how long it lasted), agency (who or what caused it), causality (why it happened and what followed from it), and memory (how it has been remembered or forgotten over time). The case studies in this book are intended as demonstrations that computational approaches can drive a thoughtful analysis of change over time driven when in dialogue with historians' categories and concepts.

Applied to the understanding of law, industry, politics, or culture, historical analysis promises a richer portrait of individual and collective behavior. The methods of digital history can be generalized to practically any flavor of inquiry based on text–whether intellectual history, histories of the state, or histories of identity and experience–so long as an archive of text exists with relative cohesion and coverage for a continuous period of time. If we had the specialization, we could have used historical analysis to understand how hip hop lyrics or obituaries changed over time.

One aspect of the approach we promote is awareness of the broader context for the subject matter and its historical context in any choice made about data and its analysis. While we could, in theory, download a dataset of lyrics that document the history of the modern recording industry, we ourselves are no experts in that domain. The case studies we present are mainly from British History because of our extensive experience in the field. In modeling our own expertise around British history, and showing how a familiarity with the parliamentary debates of Great Britain can drive an inquiry into the role of individuals in the abolition of slavery or the Contagious Diseases Acts from 1864 to 1869, we are attempting to model questions about major events, collective action, changing values, and individual interventions that appear at different periods of history and many discourses in different guises. Researchers of hip hop or of Indian history will find questions just as deep and consequential in their own studies. Given the approaches to historical analysis presented here, they may just as easily pursue their own lines of inquiry.

No book teaches all subjects or ways of thinking at once. This book offers an introduction to text mining for historical analysis via a singular programming language, R. It also teaches how to perform some computational inquiry using chatbots backed by large language models (LLMs). Throughout the book, we will lightly summarize the background thinking that a reader is required to have in order to arrive at an

adequate interpretation of why the language changed the way it did and what the analyst can learn from changing language. As we will emphasize again and again, there must be a dance between computational methods and expertise about historical context. Our desire in such a hybrid approach is to offer a book that will be useful to our colleagues and friends around the world–from San Francisco, Dallas, Atlanta, Bogota, Melbourne, Hong Kong, to Chennai–and to formulate an approach to historical analysis designed to elevate the methods of scientists and historians alike.

Insight Comes from Understanding Background Context: The Value of History

Producing insight from historical documents that is worthy of the attention of multiple fields is no simple game. It cannot be produced at the touch of a button or the application of a new algorithm to new data; it requires adjusting the algorithm, rethinking the questions, examining the data, and iterating through the work until something truly interesting has come to light. In general, this caliber of work can only be accomplished by analysts that are equally serious about history as they are about computational methods. At moments in the book, we contribute and engage theory and debates around history as well as computation. At other times, we introduce computational methods and tell a history that is already well established. We suggest that demonstrating computational methods that confirm existing knowledge will allow the reader to focus on understanding the computational method, and will allow us to make the point that examples showing what scholars already know are important for understanding how a given method can be used to process data. When both the historical context and the methodological approach are novel, it becomes challenging for an analyst to critically assess the behavior of the metric, algorithm, or data processing procedures. As we propose later, an essential step in producing digital history involves iteratively comparing computational results with established scholarship and secondary sources to develop a well-grounded rationale for one's own interpretation.

The point of this example is that a reflective process of text mining requires more than data processing skills. One must know a little about the history of a subject before one moves on to making new discoveries. It is also essential to reflect on how language can function like a game—where statements may be recorded in text that may not reflect reality, yet they can spread, accumulate, and form discourses that appear factual. These discourses, in turn, can produce real world consequences for our understanding of the historic record. Text mining can give us hints about how discourses changed, but visualizations in themselves are not sufficient to produce a significant understanding of real historical events. This process of engaging other sources and gaining a more grounded understanding of historical events is key to the iterative nature of historical inquiry.

When analysts skip over the need to engage with what other analysts and historians have found, one frequent result is a data-driven analysis that does not relate back to our broader knowledge of history. We are concerned that this mode of analysis has become more frequent over time–an issue diagnosed in *The Dangerous Art of Text Mining* as one of the perils of a practice of data science as practiced without engaging the existing body of knowledge generated within the humanities (Guldi 2024).

An analyst who aims at genuine insight from a corpus should begin by understanding the wealth of these debates on different areas of expertise, from market behavior to popular culture, and the limits of what any dataset can provide no matter how skilfully the data has been organized. Even when datasets are well organized or clean, they are not necessarily conclusive or complete. Insight into data begins with understanding the conditions under which the data was originated. These conditions manifest as biases, and include omissions and slants. We suggest that a first step to engaging data is to articulate the limitations of how a given dataset can be interpreted. For instance, the dataset most frequently used in this book – Hansard – excludes the voices of women, most of the working class, as well as colonized subjects of British empire. Even in the most pristine and organized version of the Hansard corpus, this reality is not necessarily

evident without historical background knowledge. An analyst wholly new to historical analysis will find a model in these exercises for developing a meta-awareness. Techniques for moving between multiple sources – secondary, primary, and data-driven analysis – is one of the most important contributions that this book has to offer.

Engagement with secondary sources is one of the markers of work that takes historical analysis seriously. Throughout the book, the reader will find summaries of arguments about research from the field of history and the domain of computational text analysis. Because we are writing for a broad readership, we will not assume a general knowledge of history or the significance of the examples here; we will instead introduce popular knowledge such as British Prime Ministers and major events. Throughout these case studies, we will undertake basic explanations that presume no background knowledge in a way that may feel meaningful to some readers, or simplistic to others with a greater acquaintance with a given specialty. These summaries, if simplistic, are meant to introduce the topic to readers who may come to the book from a background in another field, without any idea of why British historians talk about the year 1832 with heightened investment.

There is much that a singular book cannot hope to teach. Even while we undertake preliminary explanations of multiple disciplines–computer science and history–this book-length introduction to digital history cannot substitute for an introduction to the field of history nor to the history of democracy, the history of the United States or Great Britain – all subjects that intersect with the themes treated here. A data-driven practice of text mining that takes historical research seriously, however, can provide a higher standard of innovation in scholarship – by demonstrating processes and demanding findings that raise the bar for computational investigations of the human past.

Hybrid Scholarship and the Role of the Interdisciplinary Researcher

The digital humanities emphasizes the importance of "hybrid" teams—collaborative groups where computer scientists and humanists work together across multiple iterations to generate new knowledge. A "hybrid" practice is one where a singular researcher or teams of researchers draw from different backgrounds to continue to synthesize knowledge in new ways over months or years, developing new strategies to approach cross-disciplinary problems. The principle of hybrid work – of learning about where the data comes from and what its limits or biases are, of studying who has approached the dataset and its questions before, of returning to the text – is important whether researchers expect to find themselves analyzing hip hop lyrics, political speeches, social media, or corporate reports.

We suggest cross-disciplinary teams provide fertile ground for iterative experimentation, where the interpretive richness of history meets the power of computation to produce new forms of knowledge, or new reflections on the methods through which we produce knowledge. At the same time, we believe it is important to equip disciplinary researchers–historians and computer scientists alike–with perspective that blurs the rigid boundaries between fields and gives each researcher the ability to think across domains. This way, researchers of all backgrounds might not see themselves as distinct from the study of change over time, overlooking the ways in which historical inquiry can inform responsible system design and the social impact of their work. History is a framework for understanding how knowledge and our technologies evolve—making it relevant to every field.

A General Introduction to using Large Langauge Models for Computational Text Analysis

Since 2023, broadening literacy in computational methods has become even more important. Artificial intelligence tools—such as chatbots powered by large language models—and user-friendly platforms like Voyant Tools and Democracy Viewer have significantly increased access to computational text analysis (Buongiorno et al. 2024; Sinclair 2016). These tools lower the technical barrier to entry, enabling researchers to experiment with different methodologies and examine how they affect textual corpora, all without requiring extensive programming expertise.

However, this accessibility does not diminish the importance of learning to program. On the contrary, a basic introduction to R, such as the one provided in this book, is more relevant than ever. Even for readers who do not aim to become expert coders, the ability to understand a data processing pipeline—and to read or modify code—encourages more thoughtful and transparent analysis. It allows researchers to move beyond treating scripts or applications as opaque black boxes and instead ask critical questions about how such tools shape the interpretation of historical data.

Without this foundational knowledge, the inner workings of online text analysis tools remain obscure, limiting our ability to evaluate how data processing decisions influence our conclusions. Learning the basics of programming thus fosters deeper critical engagement with computational methods and equips scholars to challenge their assumptions about how historical knowledge is produced.For example, AI-based tools can quickly process large amounts of historical text, summarize content, and even assist in identifying linguistic patterns that might be difficult to detect through traditional close reading. However, the use of AI chatbots also raises critical questions about accuracy and interpretability, issues that we aim to address in this book. Chatbots generate probable sounding text in confident sounding ways, but they generate text probabilistically rather than through an understanding of historical nuance. This means chatbots can introduce errors or fabricate sources. Since LLMs are trained on contemporary and often skewed datasets their outputs may reflect present-day assumptions and biases rather than meaningfully representing historical contexts.

In this book, we will explore the use of LLM-based chatbots for historical analysis by using them to brainstorm and write code. We suggest chatbots can support brainstorming by proposing alternative ways to visualize results or by identifying different angles that may not have been considered by the analyst. One practical use of a chatbot is using it to generate a list of "stop words"—or, common words to be removed from a text in order to support a more meaningful analysis. Although a human analyst can create a list of stop words by hand, doing so is often tedious and time-consuming. AI can speed up this process, allowing researchers to devote more time to interpreting historical patterns and insights. However, generating something like a stop words list with a chatbot does not eliminate the need for critical judgment (Schofield 2017). Analysts must still carefully review the suggested stop words, since chatbots may mistakenly flag words that hold historical significance. Words that appear unimportant in a general linguistic context may carry unique meaning in specific historical periods or texts. Without oversight, an automatically generated stop word list could erase crucial evidence, skew results, or reinforce modern biases. Thus, chatbots must be used with discretion to ensure they serve historical inquiry.

Maintaining control over the analysis is so important because LLMs do not "understand" coding best practices, research goals, or the nuances of historical context. They generate output based on statistical patterns in their training data. As a result, the code they produce may contain errors, and their suggested methods or interpretations may be inappropriate for historical research. We demonstrate this limitation in Text Mining for Historical Analysis by prompting an AI chatbot to interpret 19th-century Parliamentary speech—an exercise that reveals the risks of relying on LLMs for interpretive analysis without human judgment.

Another important risk is that chatbots may generate code that runs without error but subtly distorts the

data, leading to potentially misleading conclusions. This risk underscores the need for researchers to develop the skills to critically review, interpret, and revise chatbot-generated code to ensure it aligns with both their research goals and the nature of their data. In *Text Mining for Historical Analysis*, we demonstrate how to assess code with a critical eye, equipping readers with the mindset and strategies necessary to ensure that code supports—rather than compromises—rigorous historical inquiry.

A General Introduction to Computational Methods

For many analysts new to the domain of digital history, it is hard to know where to get started if their aim is to analyze text data. They may find themselves searching instructions online, all the while asking themselves: "should I begin with cleaning my data or learning the basics of command-line instructions? Should I start by counting words? Should I commit to Python or R?" So many choices may appear at once.

Analysts need a structured and focused program of study—one that equips them with practical skills for analyzing text and understanding change over time, rather than a collection of disconnected lessons that may not align with their research concerns. To meet this need, we have organized the core challenges faced by text analysts into a coherent series of chapters, each designed to build on the previous one. Our approach ensures that scholars can develop a functional understanding of text mining for historical analysis without unnecessary distractions.

Many of the concepts demonstrated in this book include detailed explanations and step-by-step guidance tailored for readers who have never programmed before, as well as those with some programming experience who are new to R or its application for historical analysis. To support this range of learners, the early chapters provide structured, guided introductions to key programming concepts, which are then built upon progressively throughout the book.

We have selected R as the primary programming language for this book because it offers a balance between ease of use and analytical capability. Compared to other languages such as Python–which often require a baseline familiarity with complex syntax, multiple data structures, and strict data management–R enables researchers to engage in text analysis more quickly and with fewer technical barriers. This makes R particularly well-suited for historians and other researchers in the humanities. By minimizing technical overhead we can concentrate on mastering computational methods that directly advance our research goals.

Throughout the book we try to communicate basic concepts originating from computer science and history and define key terms. A glossary of terms from both history and computer science are also provided in the appendix.

The Longevity of the R Programming Language

Readers may question the long-term viability of the R ecosystem. These questions may stem from concerns about the long-term stability of the R packages used in this book, the instructions we provide for interacting with the RStudio interface, or the rapid advancements of generative AI as a primary means of generating code (as opposed to reading or writing code ourselves). These questions are especially important in the context of a book that has technical content, where technological updates can render examples or instructions as less applicable.

Software packages can change, and it is possible that RStudio could change its look or feel. We agree understand these concerns, which is why we have taken several steps to ensure the durability and reliability of the instructions in this book. First, we have chosen R and RStudio deliberately for their longevity and

wide adoption. RStudio's user interface has remained largely consistent for over a decade, and we expect our instructions to remain accurate for many years to come, regardless of changes to operating systems or software updates.

In respect to the R packages used throughout the book, some are widely adopted and have become "stable." This means they are well-tested and unlikely to change significantly. We avoid relying on "bleeding-edge" or experimental versions, which are more likely to introduce changes that can "break" our code. Other packages, like `hansardr`, were created by the authors. We are committed to maintaining such packages into the future. We also welcome reader feedback on future updates.

Finally, we acknowledge a broader concern: in an era of fast-moving technological change–especially around generative AI–how can readers trust that a book like this will remain useful? Ultimately, *Text Mining for Historical Analysis* is not a code cookbook. It introduces foundational concepts and methodologies for digital history beyond the use of a singular tool. In doing this, our book also teaches readers how to interpret and engage R. Our book extensively uses Hadley Wickham's `tidyverse`–a set of R packages that share an underlying philosophy and grammar of data processing and visualization (2019). However, the particular functions we use from the `tidyverse`—along with our selective method of employing functions, the process for ordering them in a larger code body, and the way we pair them with other R libraries like quanteda—demonstrates our methodological approach to structuring a reproducible and interpretable text analysis pipeline for the field of digital history.

In short, while no software is immune to change, we have made every effort to ensure that *Text Mining for Historical Analysis* offers a conceptual understanding that will remain relevant, and that exercises in this book are built on a foundation of stable, well-supported tools. We believe this will keep the book relevant and useful to readers for many years to come.

The Data in this Book

The analyses in this book primarily concern Hansard, which contains the official record of the debates of Britain's House of Lords and House of Commons, 1803-1899. It is an interesting data set because the debates cover the transformations of the Industrial Revolution, the abolition of the trans-Atlantic slave trade by the British navy, the struggle for women's rights, and many other debates of general interest to readers of modern history.

One advantage of exploring the British parliamentary debates, as we do for most of the volume, is that we have jointly worked for at least a decade on the datasets presented in this book. Our familiarity allows us to model an engagement with issues of data quality and historical inquiry – something that would be more difficult to organize in a book that moved from dataset to dataset at random. Deep familiarity with a dataset, its historical context, and the kinds of historical questions that might arise when looking at a dataset is one of the prerequisites for serious analysis of change over time.

As we will explore in the chapters that follow, Hansard is in many ways an imperfect data set. The Hansard data dates from a moment in history before the advent of modern technologies of transcription, including shorthand and electronic recording. Instead, it was compiled from press reports and notes, highly edited for what journalists assumed would be of interest to contemporary readers, and sometimes paraphrased to keep pace with the speaker. Nonetheless, even in the early versions of the reports, a great deal of data about the social, economic, and political realities facing parliament remain preserved. Though far from a verbatim transcript of the records of parliament, it is nevertheless an important historical source that has served generations of historians, many of them working with concerns social and cultural as well as political and intellectual.

Part of what computational text analysis can accomplish is an investigation into the record of bias left by an earlier age. The Hansard dataset is an artifact from an era where the vote was limited to a tiny portion of the aristocracy (before 1832) and the middle class (to 1867). The record is also almost exclusively male, as women could not vote in Britain before 1918. Britain of course controlled a global empire, and many of the speakers expressed disdainful attitudes towards their Roman Catholics, Indians, Africans, and Irish subjects; indeed, historians have reasoned that it was the British willingness to fabricate arguments about racial superiority and the necessity of violence that allowed the empire to exist. The techniques in this volume allow the student of history to examine British attitudes about gender, race, and class for themselves.

The Challenging Skill of Detecting the Unexpected

There are many software packages that allow analysis to import datasets and quickly perform text mining, such as in the form of sentiment analysis, word counts, and topic models. To discover the genuinely unexpected from a historical standpoint, analysts must already have a deep familiarity with a field. Few books offer such a deep, conclusive history of Britain. Instead, it is more typical that several books on history and culture must be read in unison, for which *Text Mining for Historical Analysis* might be one. Our goal in teaching digital history methods are not to challenge assumptions about history itself. While some of the interpretations presented in this book may do so, that is not our primary aim.

We hope that teaching digital methods initiates a form of engagement that may produce genuine surprise and meaning for those familiar with the historical events. For those less familiar with these events, we hope that learning text mining provides not only a deeper understanding of the past but also a critical framework for interpreting historical narratives and patterns. One of the crucial skills we demonstrate in *Text Mining for Historical Analysis* is the ability to familiarize oneself with the many different sources and conversations that form the broader context for a dataset. Historical documents do not exist in isolation; they are embedded within a larger intellectual community where researchers contribute findings, interpretations, and connections that enrich and contextualize the data. Engaging with these networks of scholarship allows for a more informed and nuanced approach to historical analysis. These additional sources–often termed as "secondary sources"–often provide give analysts an ability to engage documents by themes. Just some examples of these themes may include: The operation of democracy, the problems of gender and race, or the challenges of understanding technology. The insights derived from these themes are often termed "theory."

Among the most important skills for digital history is "iteration," a practice long been theorized in scientific as well as humanistic inquiry as a key method of working with data. The origins of such an approach might be traced back to the 1970s, where scientific scholars, like John Tukey, formalized an approach called "Exploratory Data Analysis" (EDA), which involves a series of steps that move from initial data visualizations to those focused on answering a specific question. Such an approach is applicable to historical analysis (Tukey 1977). We will show that users can transition from a high-level overview of a time period and into a close reading of the text itself. The importance of taking an iterative approach to research has been emphasized by other fields, too. Political scientists, like Justin Grimmer and his collaborators, stress the importance of iterative engagement by moving between theory-based lines of questioning (also termed "induction") and data-based lines of questioning (called "deduction") in a cycle referred to as "abductive logic" (Grimmer 2022). In her article, "Critical Search," Guldi outlines an iterative process where the researcher moves from reading historical overviews to working with data, to reading critical theory to engage with these questions, to finding data insights, to then adjusting the algorithm to align with the scholar's questions, to testing those insights against the additional literature to see whether and how scholars have engaged these episodes (Guldi 2018). In a more concrete example of iterative research: scholars like Lauren Tilton have examined how EDA can serve as an impetus to hone a research problem from a timeline of all the photographs in a section

of the Library of Congress to the timing of content in photographs taken by two prominent photographers (Arnold 2023, Wexler 2014).

Our book models an iterative approach to analysis. For example, when examining how speech patterns varied across decades, we show that slicing and re-slicing the data—based on slightly different research questions—and applying different metrics to the corpus can yield entirely different results than applying a single statistical measure once. A one-time application of a method may overlook the multiplicity of narratives and insights embedded in the data. While no single approach can capture this complexity, using multiple methods helps illuminate the richness and layered meanings within the corpus.

Another skill for producing insight from historical analysis is an appreciation of the dynamics of "validation" and "discovery." As stated, the methods of digital history blend scientific and humanistic modes of inquiry. In result, there are times when we write and run code that replicates what is already known about the historical record. These are not failed studies. Instead, they show that the analytic approach (in this case, topic modeling) is capable of replicating a specific mode of humanistic inquiry. This ability to replicate is key to a scientific appreciation for the mode of analysis. The interplay between validation and discovery ensures that computational techniques not only reinforce existing understandings but also open new avenues for historical interpretation.

As a rule of thumb, we do not prioritize discovery over validation or vice-versa; rather, we look to discovery as an ultimate step after a series of visualizations designed to validate different aspects of theory and secondary research. Researchers might colloquially refer to an "80/20" rule of validation to discovery, where 80% of work with data is expected to recapitulate already known features of the data, and only 20% of the work leads to new discoveries. In keeping with this set of expectations, we will lead readers through a process where some analyses given here will not lead to earth-shattering discoveries to expert historians of Britain's parliament. Indeed, we do not anticipate the sole readership of this book are experts in British history but rather analysts of diverse backgrounds who are engaging historical records and wish to supplement their engagement with theory and methodology from the domain of digital history. The promise of this kind of work is a process of interrogation, whereby writing code leads to more precise questions about language, which are in turn driven by broader reading. Those questions ultimately lead to data-driven analysis which may constitute new findings, whether in British history or in the analyst's own domain of research. As we mentioned above, we anticipate and hope the content taught here initiates additional inquiry into vast historical domains, and that *Text Mining for Historical Analysis* might serve as the basis–or foundation–for much larger questions.

Another key to producing insight is growing curious about data processing, algorithms, and the many ways in which they can be applied. *Text Mining for Historical Analysis* encourages a curiosity in the following exercises by walking readers through the production of an analysis, step-by-step, and then recreating a new analysis by making minor modifications to the existing code. For instance, we may first explain a computational method and then guide readers through iterative applications, demonstrating its versatility and encouraging exploration through small modifications. Both of us have proposed the value of using different algorithms to investigate the same substantive question because each algorithm and each adjustment to algorithm parameters will produce different, subjective results, and some of these results may be more or less meaningful for historical analysis. In short, there is no one "best" algorithm for any investigation. A more strategic approach leverages several related algorithms to investigate the hidden dimensions of difference latent in the data. This approach argues that a singular algorithm cannot produce a "definitive" interpretation of history. As an example, multiple times throughout the book we will compare the effect of using different algorithms designed to measure similar phenomenon, eschewing the claim that there is one "best" algorithm for any subject in favor of a genuine curiosity about patterns hidden in the data that may produce surprise and insight.

Text Mining for Historical Analysis provides readers with approaches that move from high-level overviews of words over time, to inquiry about the speaker of those words, to the use of those individual words in context. This is just one example for how this book models the process of moving through the data from quantitative overviews back to the original words in context. In the chapters that follow, we have set out to demonstrate how researchers produce insight by putting actual texts in dialogue with quantitative visualizations. Data modeling is a crucial part of the process, but it is not the sole process we demonstrate. This is because great insight almost never comes by simply reading a word cloud of top words; it comes after moving between data-driven visualizations and actual words in context over several iterations.

In *Text Mining for Historical Analysis*, we demonstrate how to integrate background readings from the humanities and social sciences to provide a more nuanced and comprehensive introduction to digital history. For example, someone with a background in data science might instinctively focus on counting feminine and masculine pronouns, while someone trained in the humanities might consider broader questions of gender and agency, exploring how these concepts are embedded in grammatical structures. By bringing these perspectives together, we can reveal how computational and humanistic approaches complement each other, leading to a richer and more holistic understanding of historical documents. This example, based on exercises in this book, is merely one case among thousands that a creative analyst might draw from in their analysis of language. But, what other strategies might an analyst leveraging the methods of digital history pursue? Perhaps they will analyze gender and place in the landscape, or ethnicity and agency with respect to the state. They may examine the gender of political speakers and which subjects women in parliament were willing to introduce. Or they may create another textual investigation entirely on the basis of their readings of critical theory. These engagements are each examples now known as the tip of the iceberg of what will become possible as scholars even more deeply engaged with theory begin to engage with computational text analysis.

Above all, an analyst's attention to "insight" means using critical thinking to interpret visualized linguistic patterns and understand what they truly communicate about the words on a page from which the visualization was derived. Without such care, analysts are liable to produce interpretations of text that are misleading. For example, an analyst might conclude that an upward tick of the phrase "ignorant women" across the nineteenth century means that there were actually more ignorant women over time. Instead, the phrase indicates the formation of a "discourse," that is, an increasingly accepted representation of reality that is communicated and shared through language. A representation, however, is not the same as reality. In the nineteenth century, speakers in parliament found that talking about the ignorance of women was a means to justify the legislation they were trying to pass, especially in the era of the Contagious Disease Acts (1864-69). Parliamentary speakers – again almost exclusively men – blamed women for the spread of venereal disease through the British navy. The habit of blaming women had profoundly negative consequences for contemporary women, who lost important liberties during this period. Because of political habits of blaming women, the laws were written in such a way that it was women who were arrested in the attempt to stem the spread of disease. The historically-informed interpretation of the changing pattern of words produces an answer that is much more interesting, meaningful, and surprising than the naive interpretation.

Using the Book With Your Own Data

What if you're not a historian of nineteenth-century British Parliament? Digital history methods can still be applied to many other datasets. New sources are continually becoming available, and researchers studying virtually any historical period may find datasets to which the following methods apply—with only minor adjustments, such as aligning column names for speaker, date, and speech content. Toward the end of the book, we will shift our focus to the Congressional Records of twentieth-century America. What about

when an analyst is ready to model their own data? Where might one acquire the data and how does one structure it for analysis? In the final chapter, we will expand our scope even further, exploring the concept of "data" more broadly by demonstrating how to collect, clean, and analyze data from diverse sources and formats, including large-scale collections hosted by modern and historic government websites. One of the great delights of text mining is that a set of code can be applied to a new dataset with an adjustment of few lines of code.

For contemporary records or certain kinds of archival material that has already been digitized by libraries and archives, the process of assembling a dataset may be relatively straightforward. Exercises in the Programming Historian give a straightforward set of steps to scraping the records from an archival digitization project into a database of a kind that can be examined with the processes in this book.

The landscape of data is swiftly changing. We already hear about historians who have had great success in digitizing photographs from the archives of typescript or even handwritten manuscripts by using large language models (LLMs). Such practices stand to dramatically lower the cost of digitization. This horizon will doubtlessly change over the near horizon.

Book Organization

Put chapter overview here.