International Journal of English Linguistics; Vol. 15, No. 1; 2025
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

An Introduction to Quantitative Text Analysis for Linguistics: Reproducible Research Using R

Hanaa Alqahtani[1]

[1] Department of Foreign Languages, Taif University, Taif, Kindgom of Saudi Arabia

Correspondence: Hanaa Alqahtani, Department of Foreign Languages, Taif University, Taif, Kindgom of Saudi Arabia.

Received: September 15, 2024 Accepted: November 19, 2024 Online Published: December 13, 2024

doi:10.5539/ijel.v15n1p71 URL: https://doi.org/10.5539/ijel.v15n1p71

Abstract

Jerid Francom's book An Introduction to Quantitative Text Analysis for Linguistics: Reproducible Research Using R is an essential textbook for researchers and students alike, who are exploring quantitative text analysis. This book is designed with beginners in mind, it emphasizes reproducible research, offering a structured approach to text analysis through the programming language R. Spanning five interconnected parts, beginning with foundational concepts like the Data-Information-Knowledge-Insight (DIKI) hierarchy, corpus creation, and data curation, advancing to topics like tokenization, dimensionality reduction, vector space modeling, and hypothesis testing with the {infer} package. This book contains practical exercises alongside detailed explanations that guide readers through the entire process of text analysis, starting from data acquisition to predictive modeling and statistical designs. Computational methods including readability measures, sentiment analysis, semantic modeling, and topic modeling are highlighted within this book, ensuring that readers are equipped to extract meaningful insights from linguistic data. Through the incorporation of Tidyverse tools and additional resources like GitHub repositories, Francom successfully bridges theoretical understanding with hands-on application. Transparency and reproducibility have been prioritized within the text, and meticulous data documentation and open-source methodologies have been meticulously advocated by the author. Although the book is an accessible resource for English-language data, readers might be challenged due to its focus on breadth over depth when their focus might be on seeking advanced exploration or on the other hand for those without basic programming experience. Regardless of this, Francom's pedagogical approach combines clarity with practical guidance, making this book a valuable resource for students, researchers, and professionals who aim to integrate quantitative methods into their linguistic research.

Keywords: Quantitative Text Analysis, Reproducible Research, Linguistics, R Programming, Data Curation, Computational Methods

An Introduction to Quantitative Text Analysis for Linguistics: Reproducible Research Using R by Jerid Francom is an introductory textbook providing fundamental concepts and practical programming skills that can be applied to quantitative text analysis. The target audience for this book is undergraduate students, although it may prove useful for graduate students and researchers looking to broaden the range of their methodological approaches. It takes on such an approach which makes this book accessible for beginners with no previous experience in either statistics or programming, making it an excellent resource for those exploring quantitative text analysis.

The book is structured into 5 interdependent parts which are further divided into chapters expanding into each topic, beginning from the very fundamentals of text analysis (such as characteristics of good research, transparency in data acquisition, ensuring reproducibility through documentation, structuring data into datasets, the transformation of datasets for analysis, predictive modeling workflow, tidy models framework, etc.) leading towards more complex concepts and topics (such as Text normalization, data frame integrations, tokenization, recoding, generation, Clustering, dimensionality reduction, vector space modeling, cross-validation, Hypothesis testing using {infer}, statistical designs, exploratory, predictive and inferential analysis, etc.), in an attempt to walk a beginner through each topic providing enough information.

To present a clearer picture of what is involved in the process of text analysis and the range of its application, the first part of this book, titled "Orientation" provides a thorough preliminary understanding of how quantitative text

analysis fits into the broader context of data analysis and linguistic research. Part one of the book consists of Chapter 1, which expands into the basics of text analysis. The author briefly discusses the complexity of the world around us and the ability to decipher these complexities is a limitation of human nature. To be able to work around these limitations successfully; is a task fit for science. The repertoire for scientific research and exploration has significantly expanded within the last couple of decades, now in the 21st century there are highly advanced tools for data analysis, enough computational power for data storage and management, etc. Finally, the author describes how text analysis is based on real-world language and the observational data extracted from that which can be used in a broad spectrum of fields.

Moving on to part two of the book, the author highlights that foundational knowledge must be established regarding the Data, Information, Knowledge, and Insight Hierarchy (DIKI), before getting into the inner workings of a data project. Therefore, the reader must be well prepared to understand the levels of the DIKI hierarchy and fully understand the role each level plays in terms of extracting insights from data. Part two of this book has been subdivided into 4 chapters, further expanding into the fundamentals of text analysis.

Chapter two is centered around data and information. Within this chapter, the author delves deeper into what distinguishes populations and samples. Corpus is the deliberate selection to denote a population of interest. It's pointed out how an existing corpus can either be selected or the researcher can develop one on their own, what's important is that once a viable corpus is chosen, it is then able to establish itself as a curated dataset which becomes the source of information from which research will originate from. Even curated datasets require certain levels of transformations to improve the utility of the information to be analyzed. And lastly, to produce viable and reproducible research, thorough documentation of data acquisition, curation, and transformation at any given step of analysis, must be implemented.

The ultimate goal of the author is to communicate the exact steps that may aid the reader in producing viable and reproducible research. In the third chapter, we have moved a level up on the DIKI hierarchy where we focus on description and analysis. At this stage, descriptive individual and intervariable assessments are conducted to extract knowledge from the data. Only after this can the researchers move onto the analysis stage, according to the author. Three main data analysis types were highlighted in Chapter 3, with each type being a separate approach for knowledge acquisition using the provided data. Although, the choice of data analysis type rests on the shoulders of the researcher and their goals. Analyzing data is not the only important part of conducting research, how that data is presented plays a huge part as well. As far as traditional reporting is concerned, the author believes it leaves much to be desired in terms of the concealment of key analysis points. That being said, the programming approach is a viable non-traditional approach where exact applied methods can be shared with the audience and provides a definite step-by-step guide to run an analysis in a reproducible manner.

In chapter 4 the author's main focus is on research. He makes it a point to provide the optimum guidance for prospective researchers to help them develop a viable research project. He points out how good, purposeful research is never a linear process. It always requires a ton of trial and error to get right, therefore the researcher must always adopt a methodological and informed approach. To refine their research skills, the author recommends the researcher refine their research by thoroughly exploring the area of interest and connecting that with the existing research. Although there are certain practical implications (technical skills, time constraints, existence of viable data, etc.) that tend to present themselves as hurdles in the path of the researchers which essentially forces them to rethink and reevaluate their projects in ways in which they may have to implement small or significant changes. Therefore, to produce reproducible research that is beneficial to the researchers and their community, the formulation of a viable research question itself holds value for successful and insightful research.

In Chapter 5 the author covers methods for acquiring corpus data along with analyzing various components of R programming language. Custom functions and writing control statements are some core ideas discussed along with general topics such as interaction with data online. What's important is that these methods must be transparent to researchers and potential collaborators as well as the general research community. Similar to previous discussions, data documentation and its acquisition are essential and must be stored within code and human-facing documents. From this point onwards, the researcher has a comprehensive overview of data that's available on the web and accessibility to the majority of that data.

In chapter 6 the author expands on the subject of structuring raw, unprocessed data into tidy datasets. Beginning with identifying the main types of data, and moving on to the level of structure, the author points out how each original data(set) is unique in terms of its structure and file format for supporting its metadata. To keep track of the changes that we have made along the process of data curation, the newer versions of data are saved separately from the original. Having this curated version serves as a starting point for analysis since different types of analysis

might be applied to the original data. A data dictionary is formed when a researcher utilizes code to create a curated dataset, this dictionary can explain all variables and measurements within the dataset. This aligns to make research reproducible in the sense that it becomes far easier for others to comprehend the data structure.

In chapter 7, the concept of tokenization is introduced, a crucial step in preparing linguistic data for quantitative analysis, where text is broken down into smaller units, or "tokens," like words or phrases. This process is fundamental for further analysis, allowing models such as clustering and vector space modeling. Here, the author discusses the practicalities and computational requirements of tokenizing large texts, noting that it can be challenging in terms of processing time and memory, especially with wide-ranging datasets. This section marks a turning point in the book by introducing practical, resource-aware tools for handling substantial linguistic datasets, paving the way for more complex analyses like frequency and syntactic analyzing.

Curated datasets can be manipulated in ways that make them better suited for analysis. The author covers various transformation procedures in this chapter, ranging from text normalization to data frame indications. The steps provided in this chapter are bound to be applied in any given research project, although they may not be implemented in the exact order in which they are presented here. It's important to note that mixing procedures to meet specific requirements is not rare. It is crucial to conduct thorough data checks while applying transformation techniques since they are responsible for tweaking the shape and contents of the provided dataset(s). Doing so makes it easier to catch any errors along the way and ensures that transformations are functioning as required. Again, the author points out the importance of correctly documenting the transformed dataset into a data dictionary, especially when dealing with multiple data sets. Another valuable insight provided by the author is that proper and descriptive naming of any form of file is often overlooked but it's the best decision one can take.

In part five, analysis, the author dives deep into the essence of the book. In chapter 8 of the book, we are introduced to the two main divisions of exploratory data analysis, which are descriptive and unsupervised analysis. Here we go through a wide variety of methods from descriptive and predictive analysis, especially in the absence of a pre-established hypothesis. Each method of analysis provides its own sets of pros and cons, and this was highlighted by providing practical and real-world examples to the reader. The main takeaway from this chapter is whether we apply descriptive or unsupervised learning methodologies, the questions we ask must be data-driven, and we should apply methods to summarize, reduce, and sort through the intricately detailed and complex datasets.

To aid readers in effectively approaching the framework for predictive modeling and Tidy models, the author provides a detailed outline in chapter 9. This outlined workflow can then be applied to text classification and regression tasks. This approach has multiple benefits, to begin with, the reader gains the skill of identifying, selecting, and engineering features, additionally, they gain experience building and tuning models. In an attempt to evaluate these models, readers are guided to cross-validate the performance and to finalize their interpretation with the right methods of extracting feature importance.

In chapter 10 the author walks us through testing the null hypothesis utilizing a tool called {infer}. This tool conducts statistical analysis utilizing simulations. Different types of study designs were explored within this chapter, for example, univariate, bivariate, and multivariate analyses. The authors also discussed how to explore hypothesis testing for different types of data such as numerical and categorical. The details mentioned in this chapter lead to a solid conclusion, it showcases how {infer} is a versatile and powerful tool for statistical analysis. Since the methods utilizing {infer} enable the researcher to test a wide variety of hypotheses straightforwardly and consistently.

The final section of this book is all about communication within the scientific realm. Communication is the key to being understood and to successfully convey your message across to an audience. For the majority of the part, regardless of the audience, a good presenter is always able to hook people in, thereby having a greater impact. Chapter 11 covers the importance of effective communication. It then digs deeper into the strategies that ensure a reproducible research project.

The final section focuses on presenting and sharing research findings. It covers the structure of scientific reports, the importance of peer review, and strategies for making research more accessible to the broader community. The emphasis on communication highlights the book's practical approach, ensuring that students and researchers can not only perform analyses but also share their findings effectively. The role of public-facing research and the importance of thorough documentation for reproducible results has been highlighted throughout the book. These factors are imperative within the world of modern scientific inquiry, something that is constantly evolving and growing at a significant speed with the passage of time. The fundamentals of scientific inquiry such as transparency, reliability, and accessibility, according to the author, regardless of the rapid progress remain the same.

For the purpose of analyzing written content, this book covers numerous computational methods. It starts from

readability measures, this is helpful for assess how easy the text can be read and understood by its readers, to help tailor the content for a diverse audience. Next, we are introduced to sentiment analysis which aids the readers in determining the emotional tone being conveyed through the text- whether it is positive, negative or a neutral sentiment- further providing insights into the emotional context of a body of text. Semantic models analyze the meanings of words based on the context in which they appear, this aids researchers in gaining a better understanding of how words relate to each other, semantically. Identification of hidden themes or topics within large collections of text is a topic modeling technique, something which is also covered within the pages of this book. It helps to uncover the underlying subjects within a corpus. Lastly, the book touches the topic of text classification methods, in which text is categorized into predefined groups using supervised learning techniques. These classifications are based on genre and sentiment, enabling the automated analysis of large text datasets for predictive tasks. Valuable insights can be extracted from textual data and content can be organized efficiently through the utilization of these methods.

To sum up, An Introduction to Quantitative Text Analysis for Linguistics: Reproducible Research Using R is a comprehensive, authoritative, and well-written textbook, which prepares the reader with the confidence and skills to conduct successful quantitative text analysis all the while making sure that it is reproducible. One of the key strengths of this book is its educational design. The author, Jerid Francom, does not automatically assume that each reader will be well-equipped with beginner-level knowledge regarding the subject, and is patient and thorough with his pedagogical approach, which can gradually guide the reader toward a higher level of understanding and expertise. He also places a strong emphasis on reproducibility throughout the book, pointing towards transparency, the rigorous advocation of data documentation, and utilizing tools such as R which are open source. All this ensures the data being presented can be verified by other scholars. The Tidyverse suite of packages used in this book makes it highly practical for the reader to follow along. Hands-on experience with real-world data is provided through additional resources such as the accompanying GitHub repository and the {gkit} package. These external resources enrich the experience beyond the bounds of the book. Keeping the positive aspects in mind, this book did present some limitations. One limitation is the emphasis of breadth over depth, in the sense that this book touches upon a long list of subjects but not exactly diving deeper into each one, rather skimming over them. This is not a fault per se, but its effectiveness is limited for the more advanced readers. While Francom makes it evident that the techniques presented in this book apply to a wide range of languages, the content of this book focuses solely on English-language data, which largely limits its immediate adaptability to other linguistic contexts. Lastly, regardless of the relatively easy-follow guide paired with thorough instructions and exercises, some beginners might find the learning curve rather steep and challenging, especially those without a programming background, and may struggle to find their way around the programming-heavy analysis within this book. Irrespective of its limitations, the book does a great job covering the promised topics of discussion and is highly recommended for students, researchers, and even experts looking to refresh their skills.

Acknowledgments

Not Applicable.

Authors' contributions

Not Applicable.

Funding

Not Applicable.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Canadian Center of Science and Education.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.