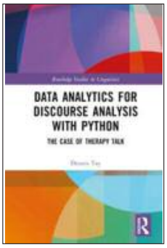

Data Analytics for Discourse Analysis with Python: The Case of Therapy Talk

Dennis Tay (Ed.)

New York: Routledge, 2024. 190 pages. ISBN: 978-1-00-336029-5

Data analytics is the process of examining, cleaning, transforming, and modeling data to extract useful information, identify patterns, and support decision-making. In this field, the data available for analysis includes not only statistical numbers but also textual corpora. In recent years, the development of data analytics, especially text mining, has provided novel approaches for quantitative research in linguistics, including discourse analysis. The book under review, featuring popular text-mining techniques and an abundance of psychotherapeutic texts, constitutes a timely contribution to quantitative discursive research.

The volume consists of six chapters, including an introduction and a conclusion. The introductory chapter starts with the definition of data analytics and its interdisciplinary significance, with a spotlight on its relevance to discourse analysis, especially the analysis of psychotherapy talks. It then outlines methods used in this book for quantifying the features of language and texts, such as word embeddings and LIWC (Linguistic Inquiry and Word Count, p. 18), and introduces Python as a basic tool for data-driven linguistic analysis. Here, the author provides many examples to explain complex linguistic features in an easy-to-understand manner. For example, the author uses visualization forms to illustrate the basics of word vector calculation and how this can be used to measure the similarity between words. When explaining LIWC, the author provides specific examples to demonstrate how to quantify the emotional and cognitive features of language. The introductory chapter is particularly informative and methodologically useful for readers who are interested in learning research methods that are complementary to traditional qualitative approaches for discourse analysis and linguistics study at large.

The main body of the volume comprises Chapters Two to Five, each dedicated to a distinct analytical technique. These chapters are thoughtfully

ISSN: 1139-7241 / e-ISSN: 2340-2784
https://doi.org/10.17398/2340-2784.49.342

structured with an initial conceptual overview, followed by a practical case study, and complemented by annotated Python scripts. Such an arrangement ensures a clear and coherent presentation of each technique, maintaining the clarity and conciseness of the book.

Specifically, Chapter Two explores a computational technique called Monte Carlo simulation (MCS), and its application in handling uncertainties. The core of MCS lies in the simulation of multiple potential outcomes to approximate complex phenomena (p. 39). To put it simply, the aim of MCS is to predict the unknown with the known. Through interesting daily-life examples such as "birthday problem" (p. 29) and "casino roulette" (p. 33), the author demonstrates how to employ MCS to infer missing data by leveraging the properties of existing data. Such a technique deserves the particular attention of researchers interested in analyzing psychotherapy data. Psychotherapy is often a long process that spans several years and includes hundreds of sessions of conversations. As such, discourse analysis of psychotherapy is often faced with the issue of missing texts. Therefore, the MCS technique introduced in this chapter is highly useful in the field of psychotherapy discourse research, since it can be used to simulate missing texts based on an incomplete dataset, so as to achieve approximation and present a complete picture of the data.

Chapter Three focuses on the application of clustering to psychotherapy talk analysis. The author first illustrates the technique of clustering with a country grouping task, showing three distinct clusters of countries based on GDP, life expectancy, and COVID-19 statistics. It also explains the mechanism of k-means, one of the most frequently used clustering methods. The chapter then applies clustering to gauge the linguistic synchrony of therapist-client interaction, i.e., the degree of alignment of the language used by the therapist and the client (p. 73). Specifically, the author demonstrates how to categorize therapy talks into synchronized and asynchronized ones based on emotional and cognitive features of the language of therapist-client interaction. In this way, this chapter highlights the practical utility of cluster analysis in assessing the quality and outcome of psychotherapy treatment.

Chapter Four introduces the algorithms of text classification, with an emphasis on the k-nearest neighbors (k-NN) algorithm. The major difference between text classification and text clustering lies in the reliance on labeled texts for the training of the classification model. Through a case study, the chapter demonstrates the practical use of k-NN classification in predicting therapy types (psychoanalysis, CBT, and humanistic therapy, p. 109) from linguistic features in psychotherapy transcripts. The results show that, in spite of an acceptable accuracy (71%), there is still room for improvement in the classification performance.

In Chapter Five, the author shifts to time series analysis, a methodological approach that promises to unlock the temporal dynamics of discourse data. The basics of time series analysis are introduced with financial data, such as stock prices. The focus of this chapter, however, is on the prediction of future data based on the statistics from the past. Then, a case study is presented using LIWC scores to model therapist and client language over a total of 40 sessions, and apply time series models to capture and predict language patterns. Such an approach would be useful in psychotherapy practice, since the therapist could better plan or adjust their counselling strategies, including language styles, if the predicted effect of future sessions is not desirable.

In the concluding chapter, the author summarizes the application of data analysis in discourse research and suggests directions for future studies. The author encourages readers to apply the techniques presented in the book to their own research and emphasizes the importance of interdisciplinary collaboration.

In general, this book is worth reading with the following strengths. Firstly, the book excels in integrating data analytics with discourse analysis, particularly within the context of psychotherapy. This integration not only provides novel tools for discourse analysts but also points to a new application domain of data analytics. Such an interdisciplinary approach also echoes the ongoing trend of digital humanities and computational social science. The second strength lies in its practicality and real-world significance. The text type analyzed in this book (psychotherapy talks) is an under-researched genre in discourse analysis. However, the importance and necessity of studying psychotherapy talks cannot be overestimated, especially in the context of the frequent occurrence of psychological issues in the post-COVID era. The analysis presented in this book can help therapists quantitatively and accurately measure the effectiveness of treatment from a linguistic perspective and thus improve treatment plans. Thirdly, perhaps aware that some readers may lack a technical background, the author utilizes many examples from everyday life to illustrate data analytic techniques. For instance, MCS is explained with casino games and the birthday problem, both of which are straightforward and comprehensible examples for lay readers. This approach reduces the difficulty of understanding and helps to broaden the potential readership to include researchers from various disciplines in the humanities and social sciences. Lastly, for each case study, the author not only presents the experiment results but also provides Python codes that can be run directly, thus facilitating hands-on learning and practical application for readers who are beginners in programming.

That being said, however, several potential limitations of this book should be noted. First, from a methodological perspective, some algorithms in the book may need updates. For instance, the k-NN algorithm introduced in Chapter 4 was proposed in 1967 (Cover & Hart, 1967). Given that k-NN is considered a somewhat antiquated model, the prediction of therapy types may achieve higher accuracy with more sophisticated models, such as those based on neural networks or transformers. The second is concerning the neglect of non-verbal information. The volume only analyzes text data, whereas psychotherapy is a face-to-face interaction that involves visual, auditory cues, as well as body language. Such non-verbal signals also convey an important message in psychotherapeutic communication. Therefore, future research may consider conducting multimodal discourse analysis in psychotherapy to obtain more comprehensive results.

Overall, this book offers a comprehensive introduction to the application of data analytics in psychotherapy discourse research. It has provided an in-depth look at the use of NLP methods for sophisticated data analysis methodologies within the field of discourse analysis, particularly focusing on therapeutic dialogues. However, the methods applied herein are not confined to psychotherapeutic discourse analysis. Therefore, we recommend it as a useful resource for scholars in the field of critical discourse analysis, corpus linguistics, psycholinguistics, and other readers interested in applying quantitative methods to discourse analysis.

Received 29 July 2024
Accepted 29 July 2024

Reviewed by **Xiaoqian Li**
Chongqing University (China)
xiaoqianlee_cqu@163.com

References

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27. http://dx.doi.org/10.1109/TIT.1967.1053964