Leveraging ChatGPT as a Coding Tool for Systematic Analysis of Extensive Mendeley Document Collection

Mohamed Abuella

December 19, 2023

Abstract

The efficient management and organization of an extensive document library in the academic research reflects the ongoing development of knowledge and expertise. Since 2011, I have diligently accumulated a substantial collection of academic documents in Mendeley with diverse research interests. Several attempts have been made to extract meaningful insights from this extensive archive, with the most recent endeavor taking place in the summer of 2022. This study explores the use of ChatGPT as an assistant-style Chatbot to convert a vast collection of academic documents from Mendeley into insightful graphs, concise summaries, and other tasks tailored to user needs. Moreover, this study can also be applied to alternative reference management tools such as Zotero. Additionally, it demonstrates practical scalability to handle the analysis of substantial number of publications in PDF format. The GitHub repository for the source code and the output of this study is available at: https://github.com/MohamedAbuella/Analysis_Mendeley.

1 Introduction

In the realm of academic research, the diligent assembly of an extensive number of documents serves as a testament to the evolution of research knowledge. Over the years, since 2011, I have diligently built up a large collection of academic documents using Mendeley.

Multiple efforts have been made to derive more valuable information from this extensive archive.

The first attempt for analysis the Mendeley documents was during my doctoral study, when I encountered a review paper that employs text mining techniques to analyze 1000 publications in the field of solar energy forecasting [1]. Inspired by this approach, I considered applying similar techniques to analyze my Mendeley document collection. However, I faced challenges in implementing these techniques effectively.

Then, in the summer of 2022, I came across a paper by Wang et al. introduces the COVID-19 Open Research Dataset (CORD-19) [2]. The COVID-19 dataset comprises over 192,000 scholarly articles focusing on COVID-19, SARS-CoV-2, and related coronaviruses. The paper highlighted a collaborative effort among leading research groups, which, through Kaggle, launched a challenge and called upon artificial intelligence experts worldwide to contribute tools and approaches [3]. The aim was to assist the medical community in addressing high-priority scientific questions related to the ongoing health crisis.

After that, at the end of 2022 with the emergence of ChatGPT [4], a large language model capable of generating human-quality text and codes, I believe that was the opportune time to revisit my analysis of the collected Mendeley documents. ChatGPT's ability to comprehend and process information from a wide range of code sources makes it an ideal tool with a query-based tool for obtaining coding assistance these documents into informative graphs and summaries.

In the course of this study, a related works have been reviewed, such as [5, 6, 7].

This study chronicles a personal project undertaken to harness the potential of ChatGPT as a utility for transforming Mendeley documents nto valuable and informative insights. The aim is to leverage ChatGPT's capabilities to enhance the extraction and interpretation of meaningful information from the Mendeley document collection, thereby unlocking the full potential of this resource for research and knowledge development.

2 Methodology

Flowchart depicting the systematic analysis process of Mendeley documents is shown in Figure 1.

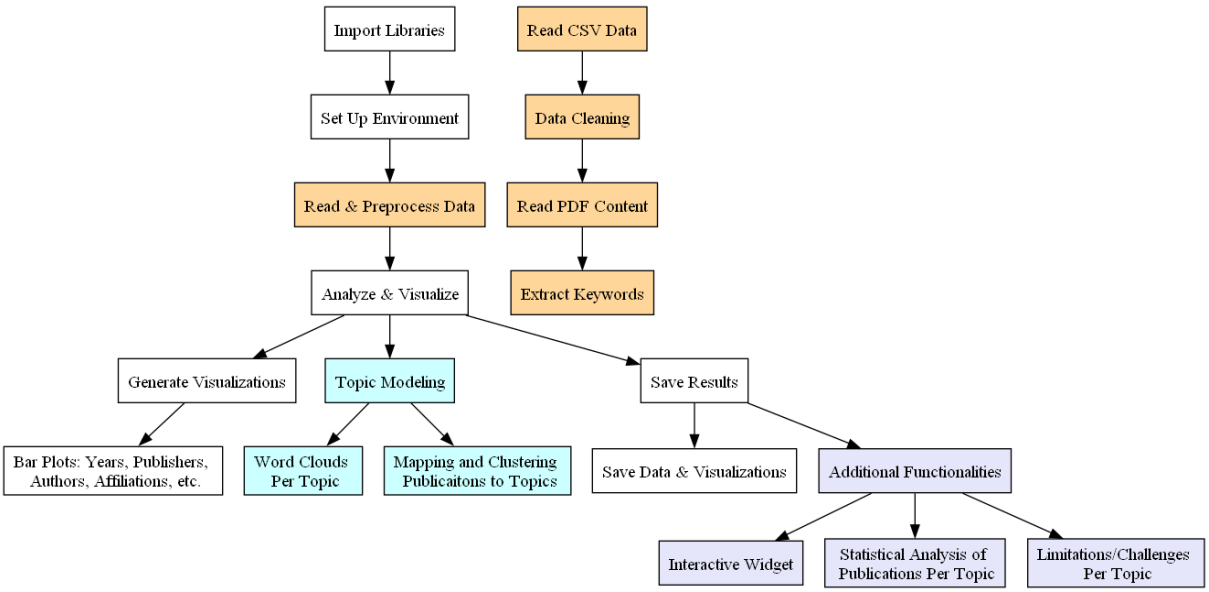

Figure 1: Flowchart for the proposed systematic analysis of Mendeley documents.

The initial step involves the extraction of data from Mendeley, comprising metadata such as titles, authors, and publication years. Subsequently, the contents of associated PDF files are processed using methods like textract and PyPDF2 to capture valuable information embedded within the documents.
Once the metadata are extracted from the documents, the pivotal role of the chatbot becomes apparent as it assists in coding and designing analysis tasks tailored to the user's needs.

Keywords are extracted to facilitate a more granular analysis of the document content. The script not only extracts essential information but also performs topic modeling using NMF, enabling the identification of key themes within the document corpus.

3 Results and Discussion

The results of the analysis are presented through a series of visualizations, including bar plots depicting publication trends over the years, distribution of document types, and insights into prominent publishers and journals. Word clouds and interactive widgets offer an engaging visualization of user tags and enable detailed exploration of individual publications.

Furthermore, the incorporation of ChatGPT for topic modeling provides a comprehensive understanding of the overarching themes within the document collection. The generated graphs and summaries offer a distilled representation of the knowledge landscape encapsulated in the Mendeley documents.

4 Conclusion

This study showcases the successful integration of ChatGPT as a utility for analyzing a substantial collection of Mendeley documents. The use of advanced natural language processing techniques, coupled with dynamic visualizations, transforms the traditional approach to document analysis. As a result, researchers and academics can derive valuable insights from their document repositories, enhancing the efficiency and depth of knowledge extraction.

In summary, the synergy between Mendeley and ChatGPT presents a compelling narrative of innovation and adaptability in the ever-evolving landscape of academic research. This project not only underscores the potential of AI-driven tools in document analysis but also serves as an inspiration for future endeavors aimed at unlocking the wealth of knowledge embedded in extensive document repositories.

Future work may involve further optimization of the search algorithm, exploration of additional natural language processing techniques, and extending the application to other domains. Moreover, integration of ChatGPT 4.0 plugins into the analysis pipeline could provide a dynamic and context-aware approach to understanding the nuances of the documents.

Acknowledgment

I wish to thank the diverse group at the Center for Applied Intelligent Systems Research (CAISR), Halmstad University, for helpful discussions.

Supplementary Materials

The source codes that are implemented on Python 3.9.7 to produce the results are available at: https://github.com/MohamedAbuella/Analysis_Mendeley

References

[1] D. Yang, J. Kleissl, C. A. Gueymard, H. T. Pedro, and C. F. Coimbra, "History and trends in solar irradiance and pv power forecasting: A preliminary assessment and review using text mining," *Solar Energy*, vol. 168, pp. 60–101, 2018.

[2] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. Kinney *et al.*, "Cord-19: The covid-19 open research dataset," *ArXiv*, 2020.

[3] Covid-19 open research dataset challenge (cord-19). [Online]. Available: https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge

[4] Chatgpt. [Online]. Available: https://chat.openai.com/

[5] I. E. Pratama. (2023) Covid eda: Initial exploration tool. [Online]. Available: https://www.kaggle.com/ivanegapratama/covid-eda-initial-exploration-tool

[6] Llama 2 is here - get it on hugging face. [Online]. Available: https://huggingface.co/blog/llama2

[7] M. Ekin. (2023) Covid-19 literature clustering. [Online]. Available: https://www.kaggle.com/maksimeren/covid-19-literature-clustering#Loading-the-Data