NASIG
TRANSFORMING THE INFORMATION COMMUNITY

Transforming Library Data Analytics into Strategic Insights with ChatGPT

Marlene van Ballegooie

Abstract

As artificial intelligence (AI) rapidly evolves, libraries have a unique opportunity to leverage these technologies for enhanced efficiency and impact. This paper examines the University of Toronto Libraries' (UTL) innovative application of ChatGPT in data analysis and decision-making processes. We demonstrate how AI can streamline and improve data analysis, fostering more informed and strategic decisions within the library context. Our exploration showcases ChatGPT's effectiveness across a variety of tasks, including data cleaning and preprocessing, data enrichment, exploratory data analysis, data visualization, and predictive analytics. By presenting concrete examples and outcomes, we aim to demystify AI applications in libraries and highlight their transformative potential.

Keywords: artificial intelligence, data analytics, ChatGPT, data visualization, predictive analytics

Introduction

As artificial intelligence (AI) continues to evolve at a rapid pace, libraries are uniquely positioned to harness these technologies to transform their operations and amplify their impact. Among these innovations,

https://doi.org/10.3998/nasig.7753

ChatGPT—a generative AI developed by OpenAI—stands out for its potential to revolutionize data analytics in library settings. The University of Toronto Libraries (UTL) has embraced this opportunity by integrating AI to enable more robust data analysis and support data-driven initiatives. This paper explores the transformative potential of ChatGPT in library settings, showcasing its practical applications across various data analytics tasks. By leveraging ChatGPT, libraries can streamline data processes, unlock new potential for data management, and enhance their ability to make informed, strategic decisions. Our findings demonstrate that implementing ChatGPT delivers benefits far beyond efficiency gains; it signifies a transformative shift in how libraries engage with their data and extract meaningful insights. This technological leap forward enables libraries to respond more dynamically to user needs, optimize resource allocation, and stay ahead in an increasingly digital landscape.

The Impact of Generative AI on Data Analytics

Generative AI, particularly ChatGPT, is revolutionizing data analytics. This advanced AI model excels in processing, analyzing, and interpreting data by understanding and generating human-like text. It significantly accelerates the data analysis process by automating repetitive tasks, allowing data analysts to focus on deriving strategic insights. ChatGPT's ability to swiftly handle data processing enables real-time identification of trends and patterns, enhancing decision-making. Furthermore, it can uncover hidden patterns within large datasets, extracting meaningful information and identifying correlations that human analysts might miss. One of ChatGPT's most valuable features is its ability to make data analysis more intuitive and accessible, offering a deeper understanding of complex datasets. Its natural language interface facilitates a more conversational interaction with data, democratizing analysis and fostering informed, collaborative decision-making across the library.

Use Cases of ChatGPT in Library Data Analytics

This paper explores seven distinct ways ChatGPT can be utilized within the library environment to enhance data analysis and support data-driven decision-making:

1. Data Preparation and Cleaning: Automating the process of cleaning datasets to ensure accuracy and consistency.
2. Writing Code: Assisting in the development of scripts and algorithms to process and analyze data.
3. Exploratory Data Analysis: Facilitating the initial examination of datasets to discover patterns, spot anomalies, and test hypotheses.
4. Data Visualization: Creating visual representations of data to communicate insights clearly and effectively.
5. Data Categorization: Classifying and organizing data into meaningful categories for easier access and analysis.
6. Predictive Analytics: Using historical data to forecast future trends and behaviors.
7. Textual Analysis: Analyzing text data to gain deeper insights into the connections between textual documents.

Each of these techniques is illustrated with practical use cases to demonstrate their application and effectiveness in real-world scenarios, highlighting the transformative potential of ChatGPT in revolutionizing library data analytics.

Data Preparation and Cleaning

Data preparation is a critical step in data analysis, often consuming a significant portion of a data analyst's time. It is commonly cited that approximately 80 percent of a data analyst's time is spent preparing datasets before analysis or visualization.[1] This underscores the importance of data preparation in ensuring accurate and reliable results.

Common data cleaning tasks include reformatting dates, removing punctuation and diacritics, standardizing names and addresses, filling null values, and identifying and cleaning duplicate values within a dataset. ChatGPT can significantly enhance the efficiency and accuracy of these data cleaning tasks.

Equipped with various Python libraries, such as NumPy and Pandas, ChatGPT can automate and streamline the data cleaning process, ensuring data quality and consistency. It can also use regular expressions to manipulate text strings into the desired format. One notable experiment we conducted involved using ChatGPT to write a script for normalizing call numbers. The process began with a description of the call number pattern and the specific changes needed, such as removing the Cutter number and publication year. ChatGPT generated a regular expression to match the pattern and provided a complete Python script to implement the changes. We tested this script on a sample dataset, and it successfully transformed the call numbers as required, demonstrating ChatGPT's practical utility in data preparation and data cleaning.

Writing Code for Data Analysis

ChatGPT excels in writing code, enabling even novice users to generate scripts and perform data analysis tasks efficiently. This capability removes barriers and allows non-technical staff to undertake sophisticated data analysis without needing assistance from a data analyst or programmer. ChatGPT can create code in Python and other programming languages and can also translate code between different languages. To promote scalability in data analysis tasks, ChatGPT allows users to download the generated code for reusability in other Python environments, making it possible to apply the scripts to larger datasets. Additionally, ChatGPT's coding abilities extend to explaining, debugging, and commenting on code, making it a versatile tool for various coding tasks. This versatility empowers library staff to tackle a wide range of data-related challenges, ultimately enhancing the library's data analysis capabilities.

For users who are not proficient in Python coding, ChatGPT can be an invaluable resource for conducting large-scale data analysis. To leverage ChatGPT's coding abilities, users simply need to describe the desired outcome in natural language and optionally provide a dataset for analysis. For instance, to identify trends in print format usage over time, we created a prompt that outlined the desired results and supplied a dataset of usage statistics categorized by Library of Congress classification. In response to the prompt, ChatGPT developed a Python script that not only performed the analysis but also visualized the results using graphs and charts. This practical application demonstrates how ChatGPT can simplify and automate the process of data analysis, making it accessible to a broader range of users.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is crucial for gaining a comprehensive understanding of a dataset, particularly at the initial stages of data analysis. It involves examining key features of the dataset to understand its underlying structure, identify outliers, and test assumptions. Through the use of summary statistics and visualizations, data analysts can discern key trends and relationships between variables. ChatGPT can streamline this process by providing quick summaries and visual representations of the data, making it easier to identify important patterns and insights. This capability enhances the efficiency and effectiveness of EDA, allowing data analysts to make more informed decisions.

ChatGPT can significantly enhance the EDA phase of a data analysis project. In our experimentation, we used ChatGPT to analyze COUNTER reports enriched with Library of Congress (LC) call number data. We crafted a prompt that outlined the dataset's key fields and requested specific insights related to LC Classification usage patterns, monthly usage trends, and platform preferences. In response, ChatGPT generated comprehensive summaries and suggested relevant visualizations. It identified the most frequently used LC classifications, highlighted seasonal usage fluctuations, and revealed user preferences across different platforms. Such AI-driven exploratory analyses serve as a foundation for more in-depth investigations, guiding librarians and researchers towards data-informed decision-making and collection development strategies.

Data Visualization

Data visualization is essential for making data more understandable and actionable. ChatGPT can generate various visual elements, such as charts, graphs, and maps, to represent information graphically. By utilizing Python libraries like Pandas, Matplotlib, Seaborn, and Plotly, ChatGPT can create diverse visualizations that highlight patterns, trends, and correlations that may not be apparent in the data. The primary goal of data visualization is to transform raw data into visual formats, enabling quick comprehension of complex information and uncovering insights hidden in the numbers. This capability allows users to more effectively interpret and utilize data for informed decision-making.

Several examples illustrate ChatGPT's ability to create data visualizations based on various datasets and specific prompts. For instance, when provided with a spreadsheet containing historical circulation data, ChatGPT generated a line graph showing overall trends in print circulation. Additionally, by uploading a COUNTER report enhanced with LC call number information, ChatGPT created a radar chart illustrating e-book usage by LC Classification. By providing a spreadsheet containing historical circulation data by stack range, ChatGPT produced a heat map visualizing resource usage within a physical space. Lastly, when given a list of top collaborating institutions, geographic coordinates, and the number of publications produced, ChatGPT created an interactive map. This experimentation demonstrates

ChatGPT's capability in creating effective visualizations that provide valuable insights and support data-driven decision-making.

Data Categorization

Another significant area where ChatGPT can enhance library data analysis is data categorization. This process involves assigning established subject categories to text strings or creating new categories based on the data's content. ChatGPT can efficiently manage this task, offering a streamlined approach to organizing large volumes of data. Unlike human professionals, who often require specialized domain knowledge to perform such tasks, ChatGPT leverages its access to a vast repository of information to deliver accurate and consistent categorizations. This capability enables libraries to organize large volumes of data more effectively and with greater speed. By leveraging ChatGPT's advanced language processing skills, libraries can enhance their data management practices, leading to better resource organization and accessibility.

To evaluate ChatGPT's potential in categorization tasks, we assessed its ability to analyze and categorize user search queries from the library's discovery system. The goal was to test ChatGPT's data categorization capabilities by having it assign three subjects to each search string and classify them into one of ten broad disciplines. The results were impressive, demonstrating ChatGPT's efficiency and accuracy. For example, when given the search term "mesothelioma," ChatGPT correctly assigned the subjects "Medical Oncology," "Asbestos Exposure," and "Cancer Treatment," categorizing it under the broad discipline of "Clinical and Life Sciences." In this analysis, a notable observation was the differing results obtained when using ChatGPT's programmatic data analysis functions, such as the Data Analyst GPT, versus the more creative native ChatGPT interface. The programmatic approaches required a data dictionary and were unable to categorize a list of over 500 search strings. In contrast, the native ChatGPT interface applied subject terminology more creatively and handled the task with ease, highlighting the importance of selecting the appropriate tool based on the task requirements. This experiment underscores ChatGPT's potential in streamlining categorization processes and handling unstructured data sets that require nuanced interpretation.

Predictive Analytics

ChatGPT offers significant potential for developing predictive analytics models within the library context. Predictive analytics involves forecasting future events or values based on historical data. For those new to predictive analytics, the initial steps can be daunting. ChatGPT can assist in predictive modeling by guiding users through the selection of appropriate predictive models based on specific data and objectives, interpreting statistical results, and assisting with code writing and debugging. In practical implementation, ChatGPT can assist with the technical aspects of predictive analytics projects, ensuring that even individuals with limited expertise can effectively utilize predictive models to derive valuable insights from their data.

Using ChatGPT, we conducted an experiment to predict the future performance of specific call number ranges within the LC classification system. The hypothesis was that accurate predictions of specific subject areas would enable libraries to make informed purchasing decisions. Utilizing a dataset with twenty years of circulation data (2003–2022), the objective was to predict trends for 2023 to 2025, providing actionable insights for collection management. ChatGPT assisted in the selection and testing of various predictive models. After comparing the results, an ensemble model combining Holt-Winters exponential smoothing and Autoregressive Integrated Moving Average (ARIMA) was chosen for its superior predictive capabilities. ChatGPT then created a training dataset using data from 2003 to 2021 and validated the model by generating predictions for 2022. By predicting known data, this step ensured the model's accuracy before making predictions for

2023–2025. To standardize the predictive analytics process, ChatGPT was used to develop a Python script that generated a detailed spreadsheet and line graphs of historical and predicted circulation numbers. This script ensured the consistent application of the predictive model across all Library of Congress classifications, and it enabled the analysis to be conducted in various Python environments, such as Jupyter Notebooks. This project demonstrates the potential of AI to advance predictive analytics and enable a forward-looking approach to library management.

Textual Analysis

ChatGPT's application in textual data analysis offers significant advantages for library professionals, enabling the extraction of meaningful insights from text-based information. This AI tool excels at processing, summarizing, and identifying common themes within documents, making it particularly useful for extracting core ideas from texts quickly and efficiently, without requiring exhaustive reading. These capabilities enable librarians to better understand user feedback, literature reviews, and research outputs, allowing for improved knowledge synthesis, collection development and resource management. The use of AI-driven textual analysis facilitates more informed decision-making, promotes a deeper comprehension of textual data, and efficiently highlights areas of commonality or alignment within diverse literary collections.

A practical application of textual analysis using ChatGPT is the identification of common themes in departmental work plans within the UTL Central Libraries. Each year, these libraries create detailed plans that outline projects and goals, which are essential for setting objectives and allocating resources. Manually reviewing these plans to identify alignments has traditionally required significant effort. By automating this process with ChatGPT, library staff saved considerable time and gained a clearer understanding of departmental alignments and potential collaboration opportunities. Initially, ChatGPT identified themes independently within the documents, producing interesting but incomplete results. To improve accuracy, we refined the approach by querying ChatGPT on specific themes, including the usage of AI tools, Equity, Diversity, and Inclusion (EDI) initiatives, cross-institutional collaborations, student wellness and experience, and metadata enhancements.

The thematic analysis of departmental workplans provided valuable insights into the priorities and strategies of various library departments. Moreover, by gaining a broad understanding of upcoming projects across different areas, ChatGPT facilitated the identification of inter-departmental collaboration opportunities. For example, it revealed that both the Metadata Services Department and the Metadata Technologies Unit were exploring AI to enhance workflows and efficiencies. This alignment highlights potential synergies in combining their expertise and strategies to accelerate AI adoption in metadata management. Leveraging AI for textual analysis not only improves the understanding of textual data but also effectively identifies areas of alignment, fostering collaboration that benefits the entire organization.

Lessons Learned

The application of ChatGPT in library data analysis has yielded valuable insights and several key lessons. One recommendation emerging from this experience is the utilization of ChatGPT's paid version when feasible. The ChatGPT Plus subscription, available at twenty United States dollars per month, grants access to the advanced ChatGPT 4o model, offering significant advantages for data analysis tasks. These benefits include enhanced processing speed, higher usage limits, and the capability to handle diverse data formats such as spreadsheets and code. Moreover, subscribers gain access to specialized GPTs designed for specific tasks and can create custom GPTs, substantially expanding the platform's analytical capabilities. This upgraded access is particularly beneficial in the context of complex library data analysis, where efficiency, versatility, and advanced features can significantly impact the quality and depth of insights generated.

A critical aspect of effectively utilizing ChatGPT for data analysis is recognizing that it is generative, not deterministic. ChatGPT generates responses based on patterns learned from extensive training data. It does not produce the same output every time for the same input; rather, it can create a variety of plausible responses. For repetitive tasks requiring consistent results, it is often more reliable to use ChatGPT to generate Python scripts, which can execute the work systematically and consistently without variability. Additionally, treating interactions with ChatGPT as a dynamic conversation can greatly enhance outcomes. Breaking tasks into smaller, precise prompts and providing examples of both desired and undesired responses can more effectively guide the model towards producing the intended results. This strategic approach to prompting not only improves the accuracy and relevance of ChatGPT's outputs but also allows for iterative refinement of the analysis process.

Effective management of ChatGPT's context window is crucial for optimal performance in data analysis tasks. The model's current context window spans 32,000 tokens, equivalent to approximately 26,000 words, beyond which it begins to lose older information. To maintain coherence and ensure relevant responses, it is essential to employ strategies such as token counting and periodic summarization. These techniques help preserve context throughout extended interactions. Furthermore, leveraging the @mentions feature enhances the interactivity and precision of the analysis process. This functionality allows users to direct specific queries or commands to particular tools or capabilities within ChatGPT, facilitating more dynamic and efficient interactions. Such targeted communication is particularly valuable when executing coding tasks, requesting specific analytical actions, or switching between different aspects of the data analysis workflow. By utilizing these context management and interaction techniques, analysts can significantly improve the depth, accuracy, and efficiency of their AI-driven data analysis projects.

Conclusion

ChatGPT is more than just a tool—it is a catalyst for transforming library data analytics. This technology has the potential to revolutionize library operations and decision-making processes. With the capability to quickly and accurately derive insights, it enables libraries to make informed decisions, optimize core functions, and enhance user experiences. Through continuous experimentation and refinement, ChatGPT enhances how libraries handle data, making processes more efficient, insightful, and accessible. By optimizing data processes and enhancing decision-making, libraries can not only improve current operations but also anticipate and prepare for future challenges and opportunities. Integrating ChatGPT into library practices ensures that libraries remain relevant and impactful in the digital age, continually evolving to meet the needs of their communities.

Contributor Notes

Marlene van Ballegooie is the Metadata Technologies Manager, University of Toronto Libraries, Toronto, Ontario, Canada.

Note

[1] Gil Press, "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says," *Forbes*, March 23, 2016, last updated April 14, 2022, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/.