Data Science Research with Newspapers, AI, and LLMs

How researchers are combining TDM Studio with AI to make historical context more accessible

Clarivate™

Research is changing. Identifying trends, patterns and connections among millions of documents has become a necessary capability. The extraction of data from unstructured information is used for methodologies such as natural language processing, machine learning and sentiment analysis. **With the advent of GenAI, research is ready to accelerate further.**

ProQuest TDM studio, from Clarivate™, is a data science solution that enhances research capabilities across all skill levels. It provides access to millions of rights-cleared publications to enable timely interrogation using text analysis and visualization tools. Researchers can write and run Python and R scripts against the full text of this content for their analysis. In addition, **TDM Studio offers access to large language models such as GPT-4**. This new development, combined with TDM Studio's access to news, dissertations, primary sources and Web of Science in a consistent XML format, will enable text and data mining to **unearth ground-breaking insights even faster**.

TDM Studio's analytical tools enable students, faculty and librarians to advance AI and Data Science studies. This platform **transforms libraries into hubs for collaborating on AI research data, maximizing the value of purchased content and demonstrating its return on investment through usage**. Additionally, it provides an opportunity for librarians to equip students and faculty with essential data literacy skills that may serve as a pivotal moment in their academic pursuits or interests.

Read on to explore three innovative projects that show how TDM Studio is empowering students to work with AI and large language models (LLMs) to make troves of historical data more accessible and useful.

Transforming historical newspaper research with multimodal models at the University of Michigan

Efficiently segmenting and parsing historical newspapers

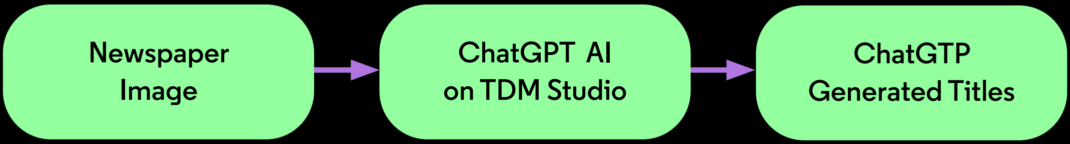
Historical newspapers offer a wealth of information, but accessing specific articles within vast collections can be a daunting task. The University of Michigan wanted to make newspaper pages from 1923 to 1999 more accessible and usable for researchers and the public. Given the variety in newspaper formats over time and the limitations of OCR technology, they required an innovative solution.

Implementing multimodal models with TDM Studio

In collaboration with Clarivate, a team of students from the University of Michigan developed a project to address this challenge using TDM Studio. The goal was to segment and parse historical newspaper images accurately, classify article components and stitch together articles that spanned multiple pages.

LLM Approach: Multimodal for Information Extraction

Use **Multimodal Model (ChatGPT-4o)** to extract information and reconstruct critical components



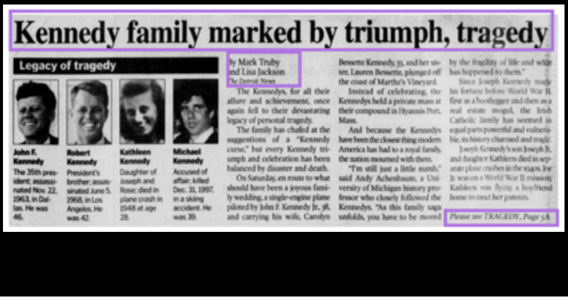
Newspaper Image → ChatGPT AI on TDM Studio → ChatGTP Generated Titles

Prompt: “There are multiple article titles in this newspaper, **identify the titles** for each article and present the title text as it is, do not summarize and don't present it if it's an advertisement. There are at most 30 titles.”

The team:

1. Classified text lines by font size to identify titles, bylines, text and jump lines.
2. Grouped and sequenced components to form complete articles and ensure correct reading order.
3. Stitched together article parts on different pages.
4. Used GPT API within TDM Studio to generate titles and segment articles, overcoming OCR limitations.

Use LLM to reconstruct jump-lines and byline / artifacts detection (index, headers)



Kennedy family marked by triumph, tragedy

Legacy of tragedy

John F. Kennedy The 35th president; assassinated Nov. 22, 1963, in Dallas. He was 46.

Robert Kennedy President's brother; assassinated June 5, 1968, in Los Angeles. He was 42.

Kathleen Kennedy Daughter of Joseph and Rose; died in plane crash in 1948 at age 28.

Michael Kennedy Accused of affair; killed Dec. 31, 1997, in a skiing accident. He was 39.

By Mark Truby
and Lisa Jackson
The Detroit News

The Kennedys, for all their allure and achievement, once again fell to their devastating legacy of personal tragedy.

The family has chafed at the suggestions of a "Kennedy curse," but every Kennedy triumph and celebration has been balanced by disaster and death.

On Saturday, en route to what should have been a joyous family wedding, a single-engine plane piloted by John F. Kennedy Jr., 38, and carrying his wife, Carolyn Bessette Kennedy, 33, and her sister, Lauren Bessette, plunged off the coast of Martha's Vineyard.

Instead of celebrating, the Kennedys held a private mass at their compound in Hyannis Port, Mass.

And because the Kennedys have been the closest thing modern America has had to a royal family, the nation mourned with them.

"I'm still just a little numb," said Andy Achenbaum, a University of Michigan history professor who closely followed the Kennedys. "As this family saga unfolds, you have to be moved by the fragility of life and what has happened to them."

Since Joseph Kennedy made his fortune before World War II, first as a bootlegger and then as a real estate mogul, the Irish Catholic family has seemed in equal parts powerful and vulnerable, its history charmed and tragic.

Joseph Kennedy's son Joseph Jr. and daughter Kathleen died in separate plane crashes in the 1940s. Joe Jr. was on a World War II mission; Kathleen was flying a boyfriend home to meet her parents.

Please see TRAGEDY, Page 5A

- **Title:** Kennedy family marked by triumph, tragedy
- **Byline:** By Mark Truby and Lisa Jackson, The Detroit News
- **Jump-line:** Please see TRAGEDY, Page 5A

Prompt: "There are multiple articles in this newspaper, identify the titles, bylines (author) and jump-lines for each article."

Leveraging TDM Studio's Capabilities

TDM Studio provided essential resources, including a powerful virtual machine for handling large datasets, high-performance computing and comprehensive data storage solutions. The platform's integrated environment allowed the team to streamline their workflow, from data import to processing and output.

Improved Accessibility and Research Efficiency

The team's new multimodal model significantly improved the segmentation and parsing of historical newspapers.

- **Accuracy of title generation:** GPT API successfully captured complete titles, even when OCR data was fragmented.

- **Correct match ratio:** Achieved an overall average correct match ratio of 70.6% across the dataset.

- **Enhanced article access:** Users could search for specific articles, authors, and keywords more efficiently, facilitating better content analysis.

Enhancing historical understanding with Retrieval Augmented Generation at University of California San Diego

Addressing the challenge of accessing accurate historical information

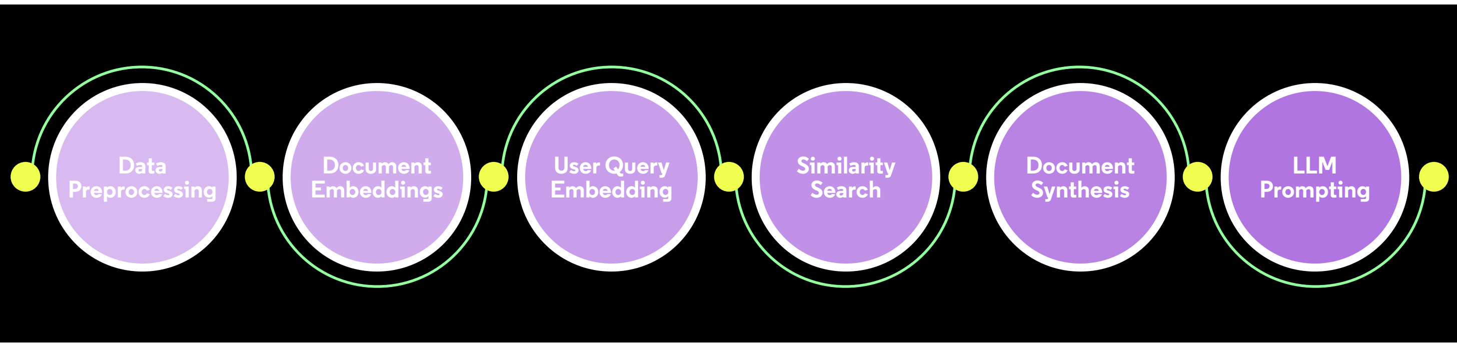
Students and researchers often struggle with accessing accurate and relevant historical information. The vast amount of information available online can be overwhelming, and determining credible sources from those that are less reliable is challenging. Srianusha Nandula and Saachi Shenoy, data science students at University of California San Diego, encountered these issues in their academic work and sought a solution to improve the accuracy and relevance of historical information retrieval. In short, they wanted better answers to questions about history.

Implementing Retrieval Augmented Generation information

The team envisioned a large language model (LLM) that could pull and consolidate content from a universe of credible sources – beyond the data it was trained on. This new research tool would gather and combine content from assorted historical news resources to precisely respond to questions about the past, helping researchers and students. Using TDM Studio, Nandula and Shenoy were able to realize their vision through the use of Retrieval Augmented Generation (RAG). RAG marries the generative prowess of large language models (LLMs) with conventional data retrieval systems. It refines LLM outputs by allowing them to pull from a trusted external knowledge base beyond their original training datasets.

To create the RAG tool, the team:

1. Collected and filtered newspaper articles from 1925 to 1929, using major news sources.
2. Used a sentence transformer to create “embeddings” to represent the meaning and sentiment of each article.
3. Created query embeddings to match user queries with the relevant articles stored in their database.
4. Used an LLM model to generate responses to user queries using the retrieved articles as context.



Leveraging TDM Studio

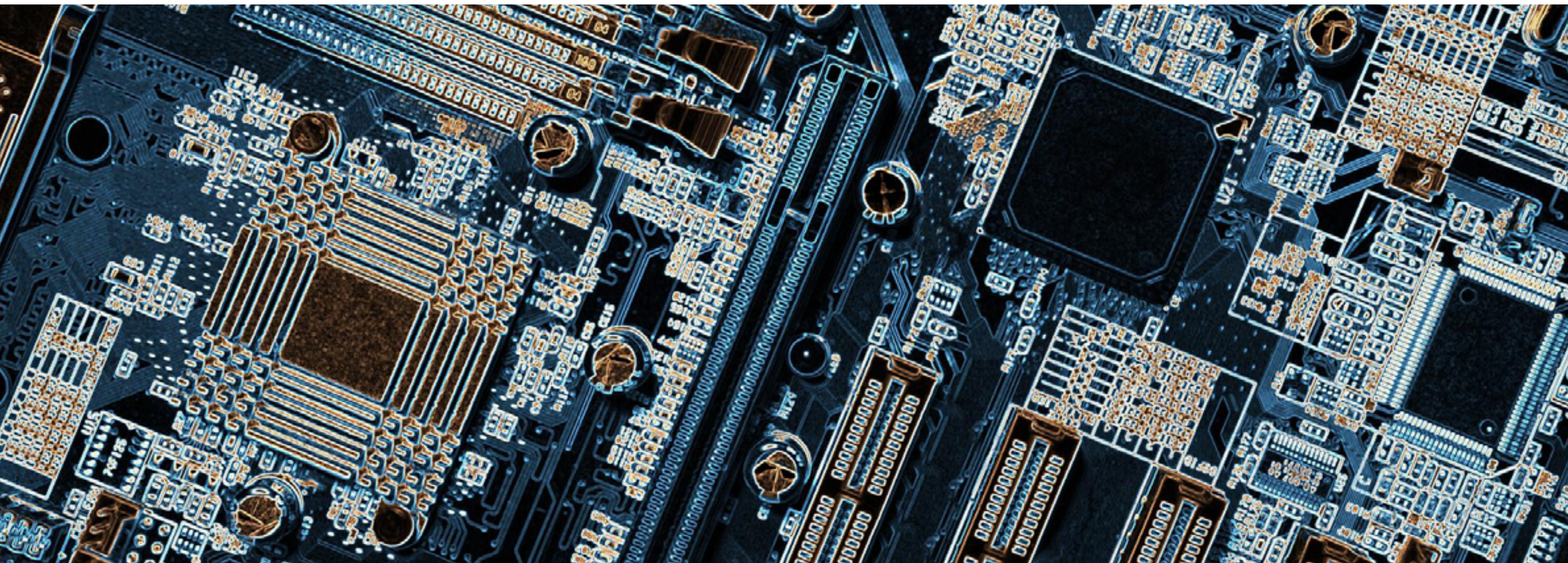
TDM Studio provided a powerful virtual environment that enabled efficient storage and access to millions of articles. The platform's high-performance computing capabilities and support for large datasets were crucial for the project's success. Additionally, TDM Studio's features for data import, storage and output facilitated the reproducibility of the code and the overall project workflow.

Significant Improvements in User Satisfaction

The RAG tool significantly improved the accuracy and relevance of responses to historical questions. The responses were more detailed and relevant when the model was provided with context from the retrieved articles. As a result, user satisfaction scores increased from an average of 2.8 to 4.8 out of 5.

> "The main benefit of TDM Studio for our project was the powerful virtual machine. This resource allowed us to search over documents efficiently, and we realized how helpful it was as we scaled up the amount of articles in our data set."

Srianusha Nandula,
UC San Diego



Exploring the evolution of ESG investing at the University of Florida

Finding patterns in public interest through historical newspaper data

Environmental, Social, and Governance (ESG) investing has gained significant attention in recent years. However, understanding its historical evolution and the public's changing attitudes toward corporate responsibility over time presents a challenge. A team at the University of Florida's Warrington School of Business sought to explore how public attention to ESG issues has evolved over a long historical horizon and what implications this has for asset pricing.

Analyzing historical newspaper data with TDM Studio

Using Clarivate's TDM Studio, the team analyzed over 130 years of historical newspaper data from major sources like *The Wall Street Journal* and *The New York Times*. By employing natural language processing (NLP) techniques, they constructed an index capturing the evolving public attention to ESG issues.

The team:

1. Gathered over 4 million business-related news articles spanning 130 years and conducted standard preprocessing steps such as removing punctuation and insignificant words.
2. Created a common set of seed words such as pollution, inequality and discrimination to generate an ESG dictionary for each decade.
3. Calculated the frequency of ESG-related keywords in each article to construct a monthly ESG attention index.

> "We are essentially leveraging natural language processing to over 130 years of news articles, which come from ProQuest's TDM Studio."

Boyuan Li,
University of Florida

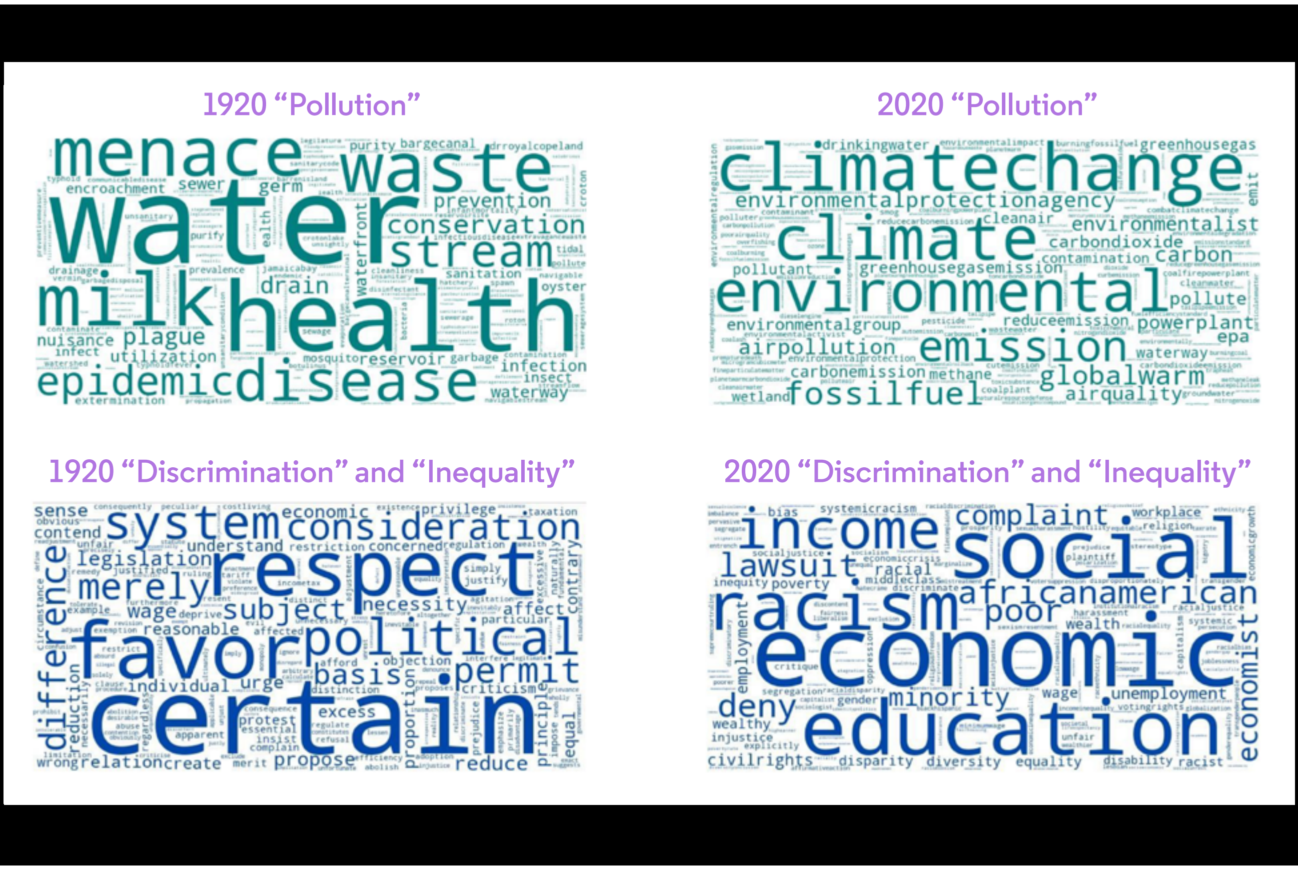
Applying TDM Studio's capabilities

TDM Studio provided the necessary tools and computational power to handle the vast dataset and complex analyses required for this project. The platform's support for large data storage and high-performance computing facilitated efficient processing and analysis.

Insights into ESG evolution and asset pricing

The analysis revealed significant patterns in public attention to ESG issues:

- **Variation over time:** The ESG attention index showed notable variation over 130 years, reflecting changes in societal concerns and priorities.
- **Historical context:** Key events, such as the Great Depression, were marked by heightened attention to social issues, while environmental concerns became more prominent in recent decades.
- **Asset pricing implications:** The study challenged previous notions by showing that ESG investing might not create long-term value, contrary to recent trends.



1920 "Pollution"

2020 "Pollution"

1920 "Discrimination" and "Inequality"

2020 "Discrimination" and "Inequality"

The development of responsible AI at Clarivate

At Clarivate we're focused on providing Academic AI tools that users and institutions can trust. We're transparent about the benefits and risks associated with AI and adhere to academic standards. Our current initiatives include integrating Academic AI in existing solutions, such as AI powered research assistance, as well as transforming teaching and learning with services like Alethea and increasing efficiencies through meta-data improvements

> "I want to thank TDM Studio for providing us the platform and all the tools we relied on for this project. Without it, it's impossible to pull off a project like this."

Boyuan Li,
University of Florida

Take the next step

Empower your research community with tools that make it possible to efficiently access and analyze millions of documents across thousands of providers.

[Request your free trial now!]