Data Visualization Research Papers in Latent Scope

by Dan Nguyen

Table of Contents

Overview and Motivation

This project originated from my interest in combining machine learning with data visualization to create interactive and insightful ways of understanding complex datasets. Initially, I was inspired by DendroMap, a system developed to support the exploration of large-scale image datasets by organizing them into hierarchical clusters using treemaps [1]. DendroMap demonstrated how clustering and interactive visualization could help machine learning practitioners analyze data distributions, identify subgroups, and examine classification errors efficiently. Motivated by this approach, I intended to apply similar techniques to a dataset suitable for clustering, envisioning an exploration tool that could help users make sense of large collections of data.

However, after further consideration and discussions, I decided to shift the project focus toward LatentScope, a method that applies large language models (LLMs) to textual data visualization. This pivot aligned better with my growing interest in natural language processing and offered a more exciting challenge, moving from structured classification datasets to unstructured text data. LatentScope extends clustering to textual representations, forming meaningful groups based on semantic similarities derived from LLM embeddings.

Building on this, I chose to adapt LatentScope to cluster and visualize research papers. Inspired by the LatentScope demonstration, which organizes textual data into interactive, cluster-based visualizations, I formulated my project around creating a tool to help researchers explore large collections of academic publications and uncover thematic patterns.

While I initially considered web scraping platforms such as ScienceDirect or Google Scholar to gather research articles, I found a more practical alternative in the VisPubData public dataset from VisPubs.com. This dataset offers metadata and abstracts from visualization research papers, making it well-suited for text-based clustering and analysis without the complications of scraping.

In short, this project aims to explore how modern machine learning models — particularly LLMs — can be combined with interactive visualization techniques to facilitate intuitive exploration of large text datasets. This direction builds on insights from prior work like DendroMap [1] while extending the application domain from images to text, which presents unique challenges and opportunities.

Related Work

The main inspiration for this project came from the research paper DendroMap: Visual Exploration of Large-Scale Image Datasets for Machine Learning with Treemaps [1]. This work demonstrated an effective approach for organizing and exploring large image datasets by using hierarchical clustering combined with interactive treemap visualizations. DendroMap's ability to reveal subgroup structures and facilitate user interaction strongly influenced my interest in applying machine learning techniques to large-scale clustering and visualization tasks.

Building on this motivation, I later shifted focus toward LatentScope, a framework that integrates large language models (LLMs) with interactive visualization. Unlike DendroMap, which was designed for images, LatentScope applies clustering methods to textual datasets using LLM-derived embeddings, which enables semantic organization of documents. I found this direction particularly compelling after exploring LatentScope’s demo (https://latent.estate/scope/enjalot/ls-dataisplural/scopes-001), which showcases how textual data can be grouped and navigated in an intuitive, visual manner.

Questions

At the outset of the project, I was driven by the overarching question of how to make research papers more accessible and visually appealing to non-expert users. Specifically, I aimed to address the following:

- How can research papers be summarized effectively so that anyone — not just domain experts — can quickly grasp their content?
- How can papers be categorized into meaningful groups to help users easily find topics of interest?
- How can a visualization be designed to not only show clusters but also display important metadata, such as the paper’s title and authors?
- Is it possible to link each visualized entry directly to the original paper, providing seamless access to the full text?

As the project progressed and I became more familiar with LatentScope and its capabilities, my questions evolved. I discovered that LatentScope is limited in terms of displaying multiple data attributes simultaneously. In particular, only a selected column can be visualized at a time. Given this constraint, I decided to focus primarily on the abstract, as it offers the most concise and informative representation of each paper's content for general users.

Through this process, new and more practical questions emerged, reflecting my shift from high-level ideas to implementation-focused concerns:

- How can I implement a search function that allows users to locate papers by author names?
- Is there a way to display the number of clusters within the search interface to help users understand the scope of the visualization?

These evolving questions guided my project from conceptual design toward a more user-centered and technically feasible implementation.

Data

The dataset used for this project consists of research papers in the field of data visualization, which was obtained from VisPubs.com, a publicly available source recommended by the instructor. This dataset includes metadata and abstracts of published papers and did not require scraping, making it well-suited for direct use in LatentScope.

Upon downloading the dataset, I conducted an initial inspection using Python's head() function to examine the structure of the data, including the available columns and their corresponding data types. This step ensured that the dataset was correctly formatted and ready for preprocessing.

For early testing and to become familiar with Latent Scope's workflow, I created a subset of 200 data points from the original dataset. While experimenting with this subset, I encountered issues related to Windows file encoding. Since LatentScope was originally developed on macOS/Linux environments, some discrepancies arose when running the tool on Windows, particularly involving special characters and emojis. These encoding issues were resolved by reviewing the error tracebacks and applying UTF-8 encoding within the code, which successfully addressed the platform-related incompatibilities.

During the pipeline setup, I ran into another challenge — the test subset of 200 papers was too small to generate all the necessary centroids, as LatentScope requires at least 256 data points for complete processing. To address this, I expanded the working subset to 300 data points, which allowed me to complete all steps of the pipeline smoothly.

Once the test workflow was confirmed to be functional, I proceeded to use the entire dataset for the final analysis and visualization. This ensured that the clustering and visualization reflected the full scope of the research papers available in the VisPubs dataset.

Exploratory Data Analysis

Since this project was developed using a pre-existing framework, LatentScope, much of the exploratory analysis focused on iterative testing to understand how the dataset would perform within the system and to identify any data-related issues that could affect the visualization output.

The process involved several key iterations:

1. First iteration (200 data points):

I began with a small subset of 200 research papers to test how LatentScope would process the data. This trial revealed an important requirement: the framework needs a minimum of 256 data points to generate valid scopes and clusters. This insight informed the need to increase the dataset size in future tests.

2. Second iteration (300 data points):

I expanded the dataset to 300 entries, which proved to be sufficient for the system to proceed through all processing stages without issue. This confirmed that, structurally, no additional adjustments to the dataset were necessary at this stage.

3. Third iteration (full dataset):

When testing with the complete dataset, I identified a new problem: approximately 70 entries were missing abstracts, which is the primary text used for clustering and embedding in LatentScope. This missing data caused the clustering process to behave abnormally. Specifically, the framework split the dataset into only two clusters — one consisting of entries with abstracts and one of those without. This unintended split significantly reduced the meaningfulness of the visualization and was not suitable for the project goals. To address this, I used Python to filter out all entries with empty abstracts, thereby eliminating the source of this problem.

4. Fourth iteration (clean full dataset):

With the dataset cleaned to remove incomplete entries, LatentScope processed the data smoothly. While the clustering and embedding results were not perfectly optimized — likely due to model parameters and default settings — the framework was functional. These limitations pointed to opportunities for further fine-tuning, which could be explored to improve cluster cohesion and the semantic accuracy of the visualizations.

Through these iterative tests, I gained important insights about the dataset and its compatibility with the framework. These findings directly informed the design of the visualization pipeline, ensuring that the input data met the structural and content requirements necessary for effective clustering and user-friendly display.

Design Evolution

At the beginning of this project, my goal was to use the most recent large language models (LLMs) to visualize the dataset and generate high-quality clusters. For the first iteration, I selected all-MiniLM-L6-v2, a general-purpose model widely used for text embeddings. While this model performed reasonably well for small datasets, its limitations became apparent when applied to a larger collection of research abstracts. Due to the relatively short length and often similar academic language used in abstracts, all-MiniLM-L6-v2 did not create well-separated clusters, which reduced the clarity and usefulness of the visualization.

After further research, I switched to a more suitable model, sentence-transformers/all-mpnet-base-v2, which is better optimized for sentence-level embeddings and captures semantic nuances more effectively. This change significantly improved clustering performance, as the model was able to better recognize and differentiate between abstracts with subtle topic differences. As a result, the clustering output became more coherent and informative.

In terms of visualization design, my initial plan was to display each paper's title and author information alongside the clusters, providing users with immediate context. However, I encountered a limitation in LatentScope: it only supports displaying a single data column (along with the index and cluster name) for each point. In light of this restriction, I decided to focus on the Abstract column, as it offered the most concise and meaningful summary of each research paper. This decision aligned with the principle of maximizing the information density of the visualization while adhering to technical constraints.

A key part of the design evolution was the fine-tuning of clustering parameters to balance the number of clusters. Specifically, adjusting the minimum cluster size and the minimum number of samples per cluster proved to be crucial. I found that using lower thresholds produced more clusters, which helped capture subtle differences between abstracts. However, creating too many clusters introduced a new issue: the automatic labeling process became less effective, as the labels for fragmented clusters often failed to convey clear and general topics. To resolve this, I iteratively tested different configurations and ultimately determined that setting the minimum cluster size to 10 and the minimum samples to 3 produced the most balanced and meaningful results. This setup allowed the visualization to maintain both topic clarity and fine-grained separation, supporting a better user experience.

Overall, while the project remained aligned with my original proposal, the design naturally evolved to accommodate the practical realities of working with LatentScope and large-scale textual data. Rather than deviating significantly from the initial plan, I refined my approach based on model performance, framework constraints, and perceptual principles, ensuring that the final visualization was informative, balanced, and user-friendly.

Implementation

The primary goal of this project was to create an interactive visualization that enables users — particularly those without deep expertise in research or data visualization — to easily explore and understand the key topics covered in research papers. The design focuses on helping users quickly grasp the general themes and subjects of the papers through cluster-based visual representation.

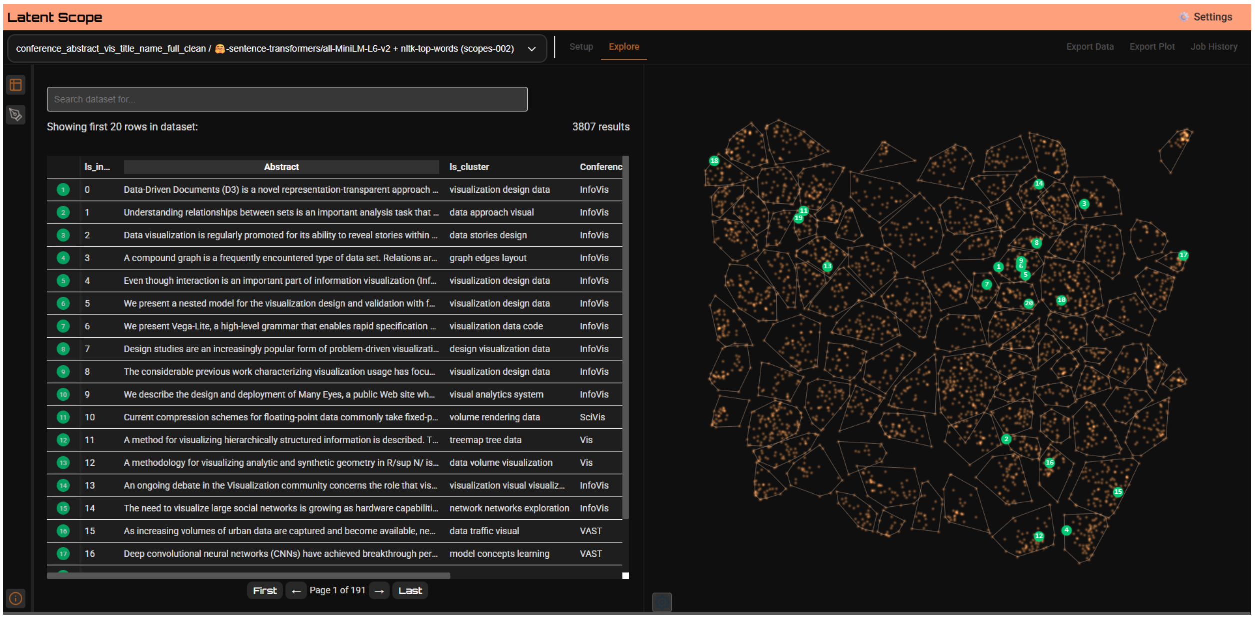
In this visualization, each cluster represents a major topic area, grouping together papers with similar themes based on their abstract content. This allows users to visually navigate the research landscape and observe how papers are organized by topic. There are 96 clusters in total, reflecting the diversity of subjects present in the dataset.

Due to the lack of official documentation on how to deploy a fully functional Latent Scope demo, users can only currently interact with an exported interactive plot version. This version still provides important interactive features. Users can zoom in on larger clusters to reveal smaller, more specific sub-clusters, enabling them to examine topics at different levels of granularity. This hierarchical zooming helps users move from broad subject areas to more focused subtopics.

In the demo, users can also hover over each cluster to view the automatically generated general topic label, providing immediate context about the cluster's contents. Furthermore, clicking on any individual data point (representing a research paper) displays detailed information, including the paper's abstract, cluster number, and index. This ensures that users can quickly learn what each paper is about without needing to leave the visualization.

Additionally, the demo includes a search feature, allowing users to filter papers based on specific criteria such as the conference at which the research was presented or by selecting from four primary cluster groups to focus on major themes.

Overall, even with current deployment limitations, the interactive plot offers meaningful functionality. It supports intuitive exploration through zooming, searching, and cluster inspection — helping users make sense of complex research datasets in an accessible and engaging way.

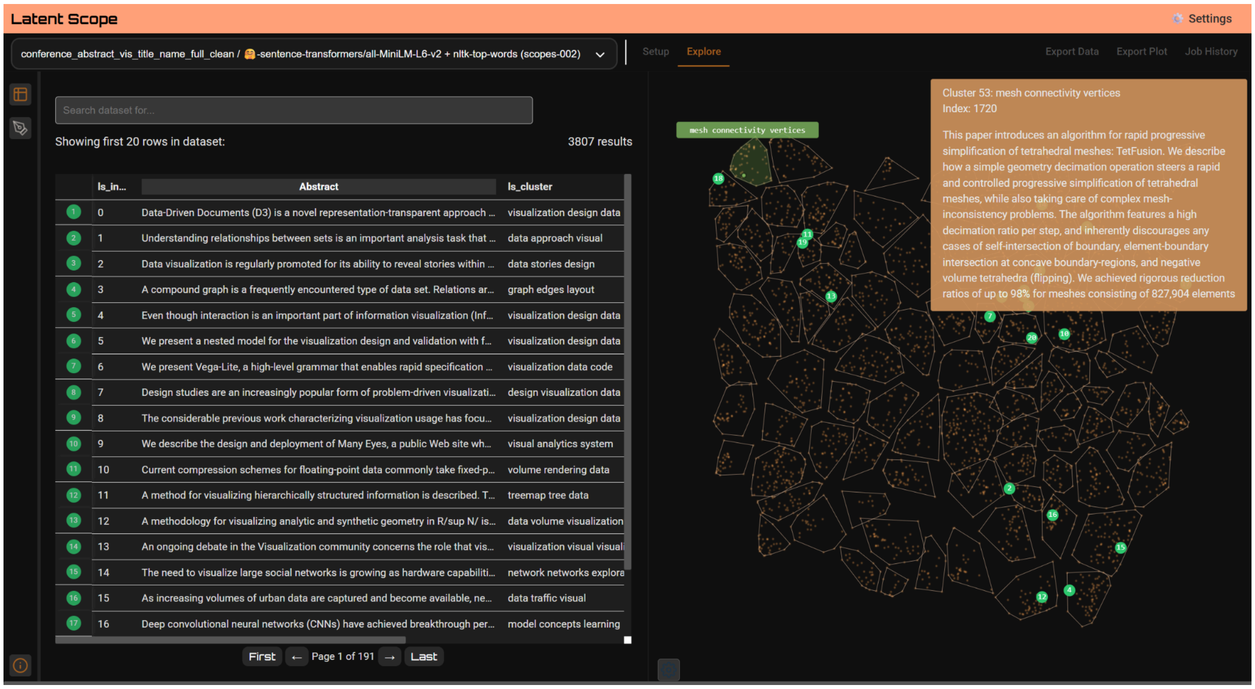

Latent Scope
Settings
conference_abstract_vis_title_name_full_clean / 🤗-sentence-transformers/all-MiniLM-L6-v2 + nltk-top-words (scopes-002)
Setup
Explore
Export Data
Export Plot
Job History
Search dataset for...
First
←
→
Last

Showing first 20 rows in dataset: 3807 results

	ls_in...	Abstract	ls_cluster	Conferenc
1	0	Data-Driven Documents (D3) is a novel representation-transparent approach ...	visualization design data	InfoVis
2	1	Understanding relationships between sets is an important analysis task that ...	data approach visual	InfoVis
3	2	Data visualization is regularly promoted for its ability to reveal stories within ...	data stories design	InfoVis
4	3	A compound graph is a frequently encountered type of data set. Relations ar...	graph edges layout	InfoVis
5	4	Even though interaction is an important part of information visualization (Inf...	visualization design data	InfoVis
6	5	We present a nested model for the visualization design and validation with f...	visualization design data	InfoVis
7	6	We present Vega-Lite, a high-level grammar that enables rapid specification ...	visualization data code	InfoVis
8	7	Design studies are an increasingly popular form of problem-driven visualizati...	design visualization data	InfoVis
9	8	The considerable previous work characterizing visualization usage has focu...	visualization design data	InfoVis
10	9	We describe the design and deployment of Many Eyes, a public Web site wh...	visual analytics system	InfoVis
11	10	Current compression schemes for floating-point data commonly take fixed-p...	volume rendering data	SciVis
12	11	A method for visualizing hierarchically structured information is described. T...	treemap tree data	Vis
13	12	A methodology for visualizing analytic and synthetic geometry in R/sup N/ is...	data volume visualization	Vis
14	13	An ongoing debate in the Visualization community concerns the role that vis...	visualization visual visualiz...	InfoVis
15	14	The need to visualize large social networks is growing as hardware capabiliti...	network networks exploration	InfoVis
16	15	As increasing volumes of urban data are captured and become available, ne...	data traffic visual	VAST
17	16	Deep convolutional neural networks (CNNs) have achieved breakthrough per...	model concepts learning	VAST

Page 1 of 191

Graph 1. Overview of Data Vis Research Paper on Latent Scope Demo


Latent Scope
Settings
conference_abstract_vis_title_name_full_clean / 🤗-sentence-transformers/all-MiniLM-L6-v2 + nltk-top-words (scopes-002)
Setup
Explore
Export Data
Export Plot
Job History
Search dataset for...
mesh connectivity vertices
First
←
→
Last

Showing first 20 rows in dataset: 3807 results

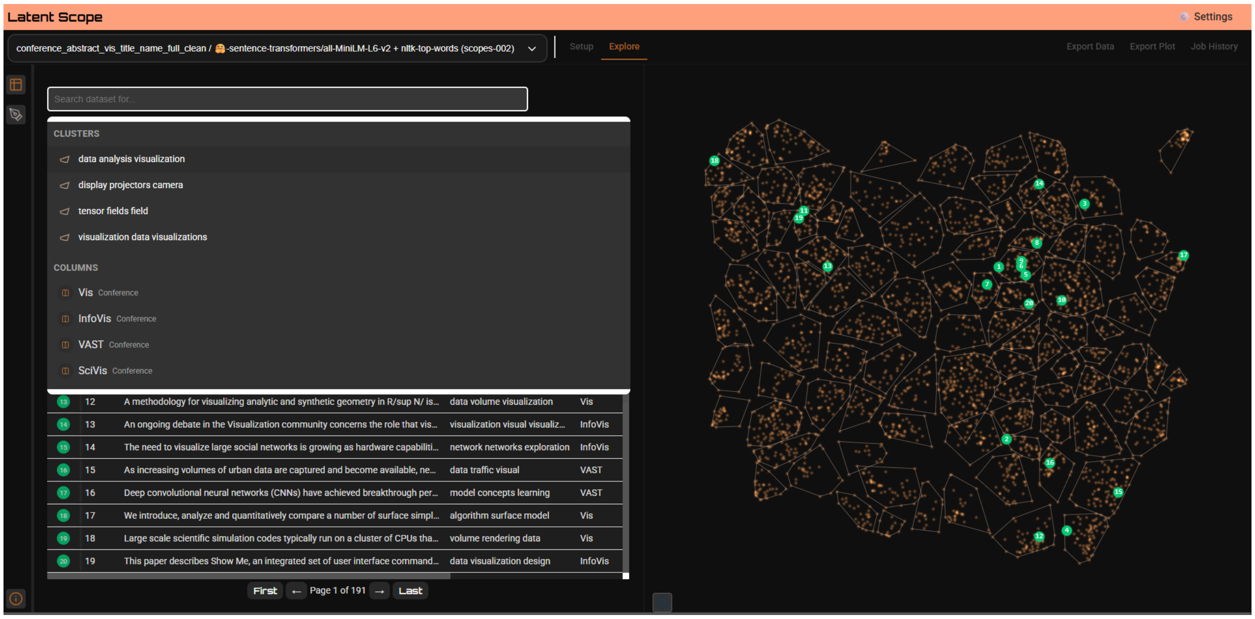
	ls_in...	Abstract	ls_cluster
1	0	Data-Driven Documents (D3) is a novel representation-transparent approach ...	visualization design data
2	1	Understanding relationships between sets is an important analysis task that ...	data approach visual
3	2	Data visualization is regularly promoted for its ability to reveal stories within ...	data stories design
4	3	A compound graph is a frequently encountered type of data set. Relations ar...	graph edges layout
5	4	Even though interaction is an important part of information visualization (Inf...	visualization design data
6	5	We present a nested model for the visualization design and validation with f...	visualization design data
7	6	We present Vega-Lite, a high-level grammar that enables rapid specification ...	visualization data code
8	7	Design studies are an increasingly popular form of problem-driven visualizati...	design visualization data
9	8	The considerable previous work characterizing visualization usage has focu...	visualization design data
10	9	We describe the design and deployment of Many Eyes, a public Web site wh...	visual analytics system
11	10	Current compression schemes for floating-point data commonly take fixed-p...	volume rendering data
12	11	A method for visualizing hierarchically structured information is described. T...	treemap tree data
13	12	A methodology for visualizing analytic and synthetic geometry in R/sup N/ is...	data volume visualization
14	13	An ongoing debate in the Visualization community concerns the role that vis...	visualization visual visuali
15	14	The need to visualize large social networks is growing as hardware capabiliti...	network networks explora
16	15	As increasing volumes of urban data are captured and become available, ne...	data traffic visual
17	16	Deep convolutional neural networks (CNNs) have achieved breakthrough per...	model concepts learning

Page 1 of 191

Cluster 53: mesh connectivity vertices
Index: 1720

This paper introduces an algorithm for rapid progressive simplification of tetrahedral meshes: TetFusion. We describe how a simple geometry decimation operation steers a rapid and controlled progressive simplification of tetrahedral meshes, while also taking care of complex mesh-inconsistency problems. The algorithm features a high decimation ratio per step, and inherently discourages any cases of self-intersection of boundary, element-boundary intersection at concave boundary-regions, and negative volume tetrahedra (flipping). We achieved rigorous reduction ratios of up to 98% for meshes consisting of 827,904 elements

Graph 2. Latent Scope Demo with viewing Abstract



conference_abstract_vis_title_name_full_clean / 🤗-sentence-transformers/all-MiniLM-L6-v2 + nltk-top-words (scopes-002)

Search dataset for...

CLUSTERS

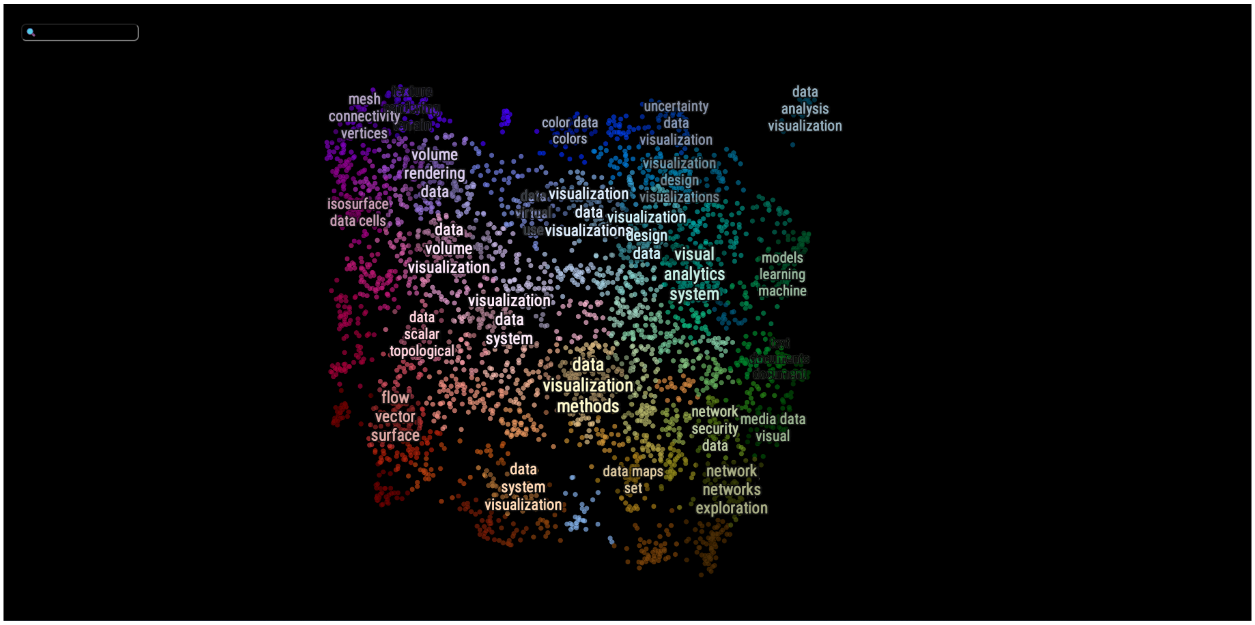
- data analysis visualization
- display projectors camera
- tensor fields field
- visualization data visualizations

COLUMNS

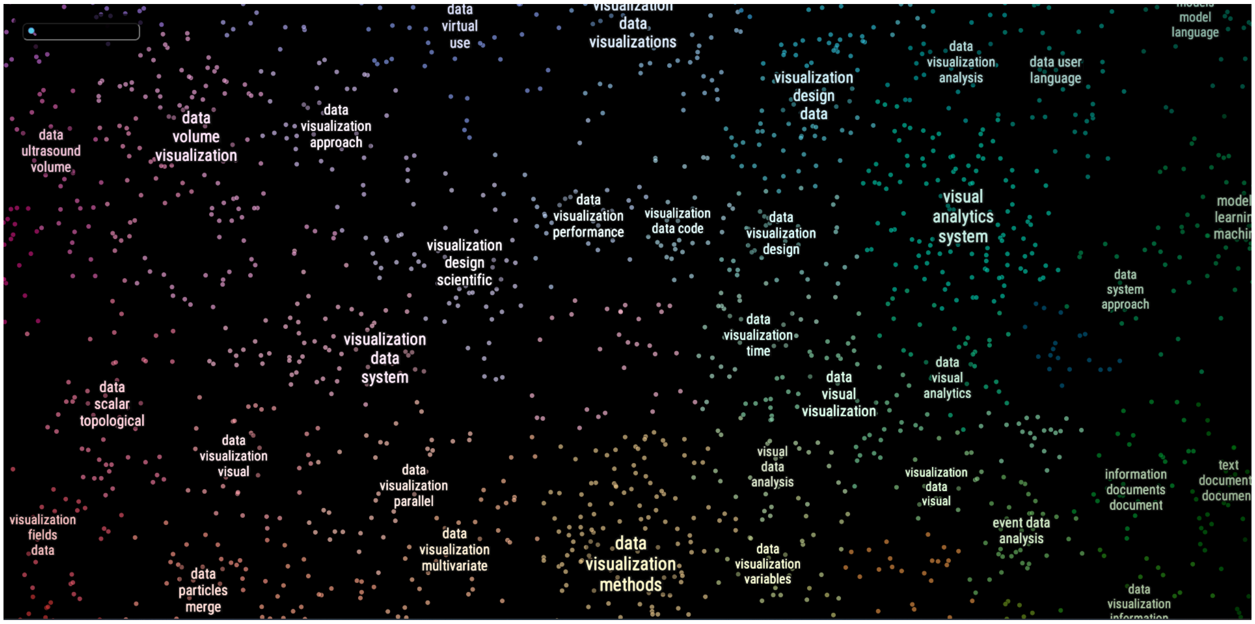
- Vis Conference
- InfoVis Conference
- VAST Conference
- SciVis Conference

13	12	A methodology for visualizing analytic and synthetic geometry in R/sup N/ is...	data volume visualization	Vis
14	13	An ongoing debate in the Visualization community concerns the role that vis...	visualization visual visualiz...	InfoVis
15	14	The need to visualize large social networks is growing as hardware capabiliti...	network networks exploration	InfoVis
16	15	As increasing volumes of urban data are captured and become available, ne...	data traffic visual	VAST
17	16	Deep convolutional neural networks (CNNs) have achieved breakthrough per...	model concepts learning	VAST
18	17	We introduce, analyze and quantitatively compare a number of surface simpl...	algorithm surface model	Vis
19	18	Large scale scientific simulation codes typically run on a cluster of CPUs tha...	volume rendering data	Vis
20	19	This paper describes Show Me, an integrated set of user interface command...	data visualization design	InfoVis

Graph 3. Search feature based on Conference and main clusters in demo



Graph 4. Interactive Plot


data virtual use
visualization data visualizations
model language
data ultrasound volume
data volume visualization
data visualization approach
visualization design data
data visualization analysis
data user language
data visualization performance
visualization data code
data visualization design
visual analytics system
visualization design scientific
data system approach
data visualization time
visualization data system
data scalar topological
data visual visualization
data visual analytics
data visualization visual
visual data analysis
data visualization parallel
visualization data visual
information documents document
text document
visualization fields data
data visualization multivariate
event data analysis
data visualization methods
data visualization variables
data particles merge
data visualization information

Graph 5. Interactive plot Zooming in feature

Evaluation

Through the use of the visualization framework, I gained valuable insights into both the dataset and the clustering process. Specifically, working with the interactive plot helped me understand the importance of tuning clustering parameters when applying LLM-generated embeddings. I learned that selecting the right clustering size and minimum sample threshold is critical: too few clusters oversimplify the data, while too many can fragment meaningful topics. Through iterative testing, I found a balance that allowed the clusters to represent distinct and interpretable categories of research papers. As a result, I now have a better understanding of how research papers in this dataset are grouped into topical categories based on their abstracts.

Although the visualization successfully allows users to explore the structure of the research dataset, certain limitations remain. Due to the lack of official documentation on deploying a full Latent Scope web demo, my output is limited to an interactive plot version. While this plot supports zooming, it does not offer live display of abstracts directly in the interface. However, I addressed this limitation by preparing a demonstration video that showcases the intended interactive features more fully, such as how users would navigate clusters and read abstracts.

Another area for improvement is the automatic cluster labeling process. Currently, the labels are generated using the nltk-top-words method, which extracts the most frequent terms in each cluster to form a label. While this approach provides a basic representation of each cluster's theme, it is limited in its ability to generate precise or descriptive labels—especially when the clusters cover nuanced research topics. This limitation is largely due to the free and simple nature of nltk-top-words. In the future, using advanced LLMs via paid API services, such as OpenAI or Hugging Face, could greatly improve label quality and make the visualization more informative and user-friendly.

Overall, the visualization meets its primary objective: it helps users gain an overview of the dataset, explore clusters interactively, and examine abstracts to understand what each paper is about. Nonetheless, there is still room for improvement in terms of deployment, real-time interactions, and the accuracy of cluster labels, which would further enhance its accessibility and overall user experience.

Reference Research Paper

[1] Donald Bertucci, Md Montaser Hamid, Yashwanthi Anand, Anita Ruangrotsakun, Delyar Tabatabai, Melissa Perez. "DendroMap: Visual Exploration of Large-Scale Image Datasets for Machine Learning with Treemaps." IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 1, pp. 320–330, 2023.