MATH 2015

Renee Tung, Bin Choi, Justin Tan

Fall 2024

1 Introduction

The rapid growth of biomedical research has resulted in an exponential increase in the volume of publications, making it essential for researchers to utilize advanced tools to efficiently explore and analyze this vast body of literature [**gonzalez2024**]. In recent years, literature-mining tools have gained significant attention, helping researchers extract valuable insights and information from large collections of publications. Among these tools, the application of large language models (LLMs) and text embedding techniques has become increasingly pivotal [**simon2024**]. By converting complex scientific texts into unified high-dimensional vectors, LLMs enable the use of machine learning and data mining techniques to facilitate data analysis and visualization [**ma2023**].

However, for visual analytics and tools for research, high-dimensional vectors are not easy to work with. By reducing the dimension to 2D (or 3D), researchers will be able to visualize each publication and interpret its position in the semantic space for a more efficient and thorough literature search. Looking at nearby papers in this semantic space will put papers in the context of the overall research landscape and aid researchers in identifying knowledge gaps or future directions that have yet to be pursued.

Here, we seek to implement a method that will allow for this kind of interpretable visualization of research publications.

Our analysis will leverage dimensionality reduction and visual analytics techniques to help identify patterns and relationships that may be present within a large dataset of published research.

Our main goal is to see if papers with similar topics, methods, or findings group together in a way that we can visualize in 2D or 3D space. We are especially interested in spotting patterns in research themes and finding connections between different subfields in biomedical literature, like areas with similar focuses or methods. We will consider a group of papers (a cluster) meaningful if it has notable differences in its Medical Subject Headings (MeSH terms) compared to other groups. By finding these relationships, we hope to uncover common trends in research, areas that might need more study, and possibilities for collaboration between fields.

2 Implementation

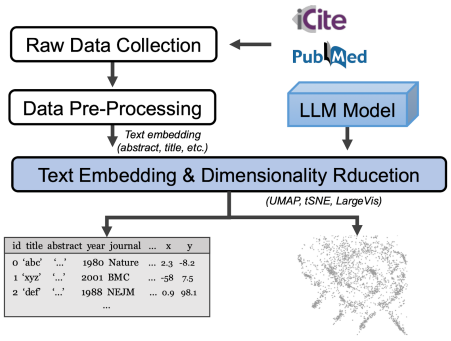

Figure 1: Flowchart of the proposed method

2.1 Data Extraction and Processing

Using the PubMed API, we will query metadata for a subset of recent publications in the biomedical domain. This step involves writing a script to produce raw XML from the PubMed results and extracting the necessary fields.

In order to compile a useful sample, we initially will limit our analysis to a smaller subfield within the domain of biomedical research, such as "working memory". Focusing on these publications yields a smaller number of publications, which allows us to develop a proof-of-concept without consuming an excessive amount of compute resources. This is also a representative starting point for researchers who will be querying largely within their subfields of interest.

Here, we search for the term "working memory"

	pmid	title	journal	year	abstract	mesh_terms	mesh_topics
0	710	Effect of two weeks' treatment with thioridazi...	Psychopharmacologia	1975	Forty paid healthy male students participated ...	Adult;Animals;Anti-Anxiety Agents;Bromazepam;C...	pharmacology;blood;pharmacology;pharmacology;p...
1	2563	EEG sleep studies of insomniacs under flunitra...	International pharmacopsychiatry	1975	This study investigates the effect of flunitra...	Adult;Anti-Anxiety Agents;Dreams;Electroenceph...	therapeutic use;drug effects;adverse effects;p...
2	2812	Alcohol and backward masking of visual informa...	Journal of studies on alcohol	1976	Alcohol increased the time necessary to transf...	Adult;Dose-Response Relationship, Drug;Ethanol...	pharmacology;drug effects;drug effects
3	2813	Recovery of verbal short-term memory in alcoho...	Journal of studies on alcohol	1976	When given a short-term memory distractor test...	Adult;Age Factors;Aged;Alcohol Amnestic Disord...	complications;complications;drug effects;drug ...

Figure 2: Example dataset resulting from PubMed API (ncbi-entrez-direct) with query "working memory"

2.2 Text Embedding

Leveraging a text embedding model to embed the abstract, title, and other relevant metadata of each publication will allow us to capture and represent the complex semantics of each paper as a high-dimensional vector. The collection of these vectors can form a matrix, which we call $\mathbf{X}$, of size $(n_{\text{features}} \times n_{\text{papers}})$. The vector length of the embeddings produced by our transformer-based language model BAAI/bge-small-en-v1.5 is 384 dimensions. Hence, we fix n_{features} to be 384.

In our prototype implementation, we generate a unique text embedding for each publication by using its title and abstract (source code). This embedding is designed to encapsulate the semantic content of the publication. Essentially, each publication is represented as a point in a 384-dimensional semantic space, where the spatial relationships between points reflect their semantic similarities. In this space, publications with similar meanings are positioned closer together, enabling us to quantify and analyze their semantic relationships effectively.

2.3 Dimensionality Reduction

To facilitate the development of visualization tools for exploring the semantic space of publications, we seek to reduce the dimensionality of the embedding matrix $\mathbf{X}$. The objective is to project the high-dimensional data into a lower-dimensional space that enhances interpretability and usability while preserving the essential spatial relationships that encode semantic similarity. To this end, we evaluate several state-of-the-art dimensionality reduction techniques, including:

- **Principal Component Analysis (PCA)** - PCA is a linear matrix decomposition method that calculates the axes that capture the highest variance within the data. It works by identifying the eigenvalues and eigenvectors from the covariance matrix $\mathbf{S}$, solving the equation:

  $$\mathbf{S}\mathbf{x} = \lambda\mathbf{x}$$

  where λ represents the variance along the direction defined by the eigenvector $\mathbf{x}$. The eigenvectors of $\mathbf{S}$, which are the columns of the matrix $\mathbf{P}$, are called the principal components. These components form new axes for the transformed data:

  $$\mathbf{Y} = \mathbf{P}^T\mathbf{X}$$

  The data $\mathbf{Y}$ is uncorrelated, and its variances are given by the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. The total variance of the data is equal to the sum of the eigenvalues of the covariance matrix:

  $$\text{Total Variance} = \sum \lambda_i$$

  PCA allows us to reduce the dimensionality of the data while retaining the most important patterns. For example, the first principal component explains the largest portion of the total variance.

In our PubMed dataset, PCA helps us reduce the complexity of the 384-dimensional embedding matrix by capturing the most important patterns in fewer dimensions. This allows us to visualize how publications relate to each other, revealing clusters of similar research topics or methods within the "working memory" subfield.

- **t-distributed stochastic neighbor embedding (t-SNE)** - t-SNE is a nonlinear dimensionality reduction technique designed for visualizing high-dimensional data in two or three dimensions. It works by preserving the local structure of the data, ensuring that similar points in the high-dimensional space remain close in the lower-dimensional representation. The algorithm achieves this by constructing a pairwise similarity probability distribution in both the high- and low-dimensional spaces. It then minimizes the Kullback–Leibler (KL) divergence between these distributions to optimize the placement of points in the low-dimensional map.

- **Uniform Manifold Approximation and Projection (UMAP)** - UMAP is another nonlinear dimensionality reduction technique. It is grounded in concepts from Riemannian geometry and algebraic topology, allowing it to approximate the manifold structure of high-dimensional data and project it into a lower-dimensional space. Unlike t-SNE, UMAP explicitly models both local and global relationships, making it more versatile for understanding overall patterns in data.

2.4 Dimensionality Reduction using t-SNE

We used the dimensionality reduction techniques to produce 2D projections of the high-dimensional embedding matrix. These methods allowed us to visualize how research publications are distributed in semantic space.

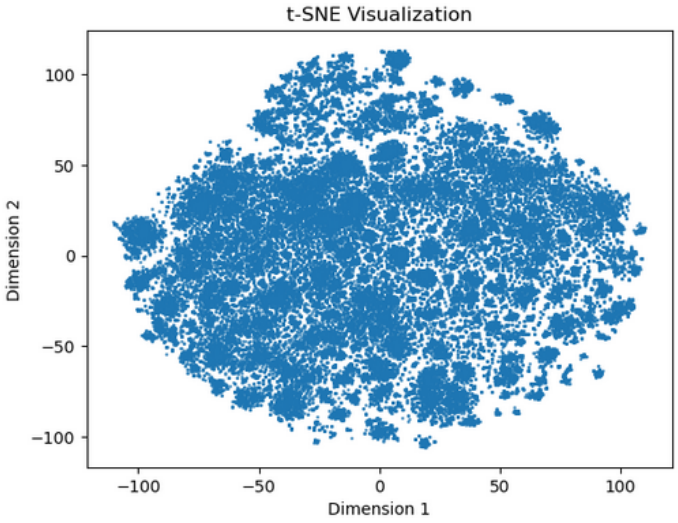

Figure 3: t-SNE visualization

Once in 2D, we applied clustering algorithms to identify meaningful groups within the data. Here we demonstrate using HDBSCAN, a density-based algorithm that groups points in dense regions while marking sparse points as noise. Applying it to the t-SNE embedding, with parameters: min_cluster_size = 250, and min_samples = 50 yielded 63 clusters.

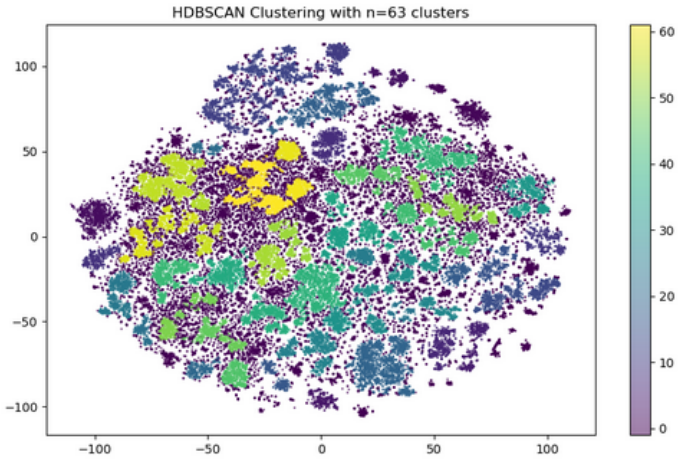

Figure 4: HDBSCAN Clustering

To explore the content of individual clusters, we analyzed the Medical Subject Headings (MeSH) terms associated with the papers. For example, Figure 5

shows the distribution of MeSH terms for one cluster.

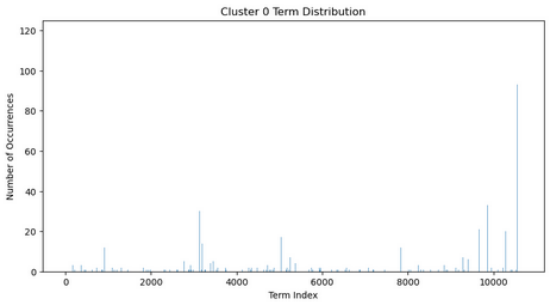

Figure 5: Histogram terms for a single cluster

2.5 Interpretation

To understand if the clusters are meaningful, we analyzed the associated Medical Subject Headings (MeSH terms) for each paper. We vectorized the MeSH terms and used statistical tests to compare distributions across clusters.

We applied the Mann-Whitney Test to compare MeSH term distributions non-parametrically between every pair of clusters. For significance, we focused on p-values less than 0.05. About 24.8% of cluster pairs showed significant differences, suggesting many clusters reflect distinct research themes.

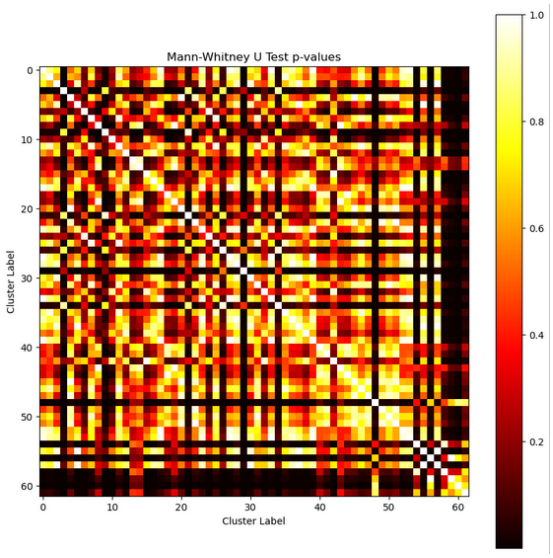

Figure 6: Heatmap of p-values computed using Mann-Whitney test, pairwise between all clusters

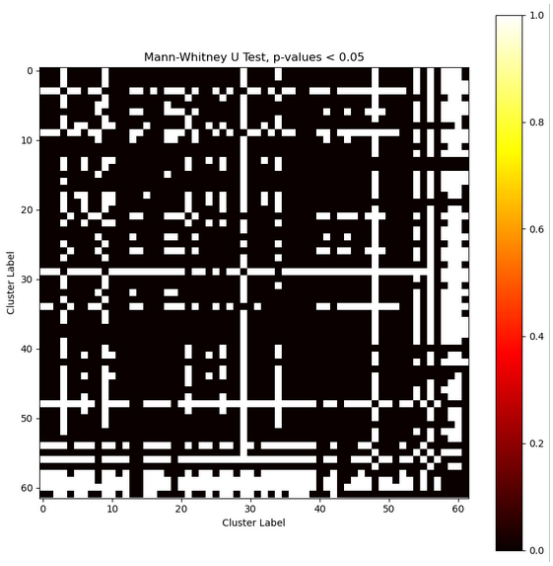

Figure 7: Pairs where p-values were significant (p¡0.05) are colored in white

2.6 Dimensionality Reduction using PCA

To be completed.

3 Results

Dimensionality reduction and clustering grouped papers by their topics. Using t-SNE and HDBSCAN, we identified 63 clusters, with 24.8% showing significant differences in MeSH terms. These differences suggest the clusters represent distinct research themes. By looking at the most common MeSH terms in each cluster, researchers can link these groups to specific research topics. ¡will do comparison for PCA and UMAP¿

4 Discussion

- shortcomings and future analyses including: using different clustering algorithms, clustering on higher-dimensional embedding, etc

5 Conclusion

This method helps researchers organize and explore literature more efficiently. By grouping similar papers, it's easier to spot trends, find related works, and

identify research gaps. It's a useful tool for navigating large datasets and discovering new areas to explore.

6 Temporary notes for this submission

There are some edits we are considering for the final submission, including reorganization to the following structure:

I. Method
II. Test Dataset
III. PCA ....
IV. t-SNE
V. UMAP
VI. Further analysis
VII. Discussions + Evaluation / or further works

References

[VH08] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[MHM18] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[BAA23] BAAI. *bge-small-en-v1.5*. `https://huggingface.co/BAAI/bge-small-en-v1.5`. Hugging Face. 2023.

[SSZ24] E. Simon, K. Swanson, and J. Zou. "Language models for biological research: a primer". In: *Nat Methods* 21 (2024), pp. 1422–1429. DOI: `10.1038/s41592-024-02354-y`.