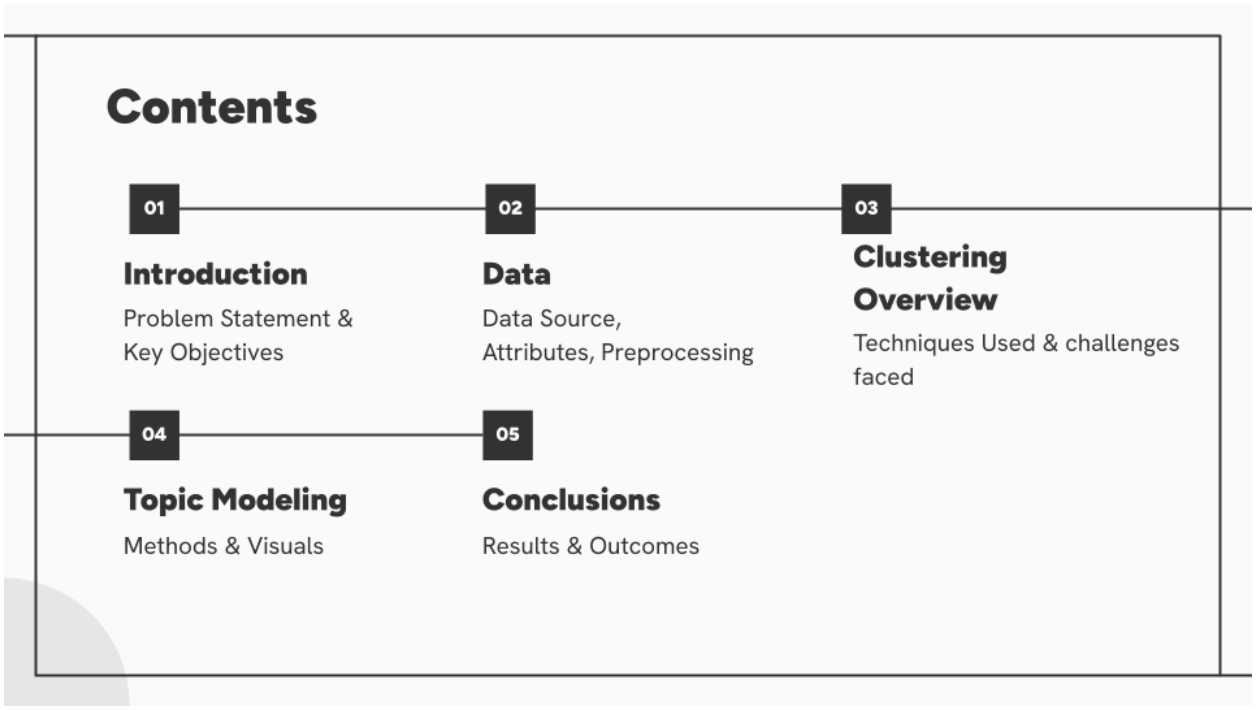
ReSearchMate

An Intelligent Recommender System for Research Papers and Insights

Course: DS5230
Team Members:

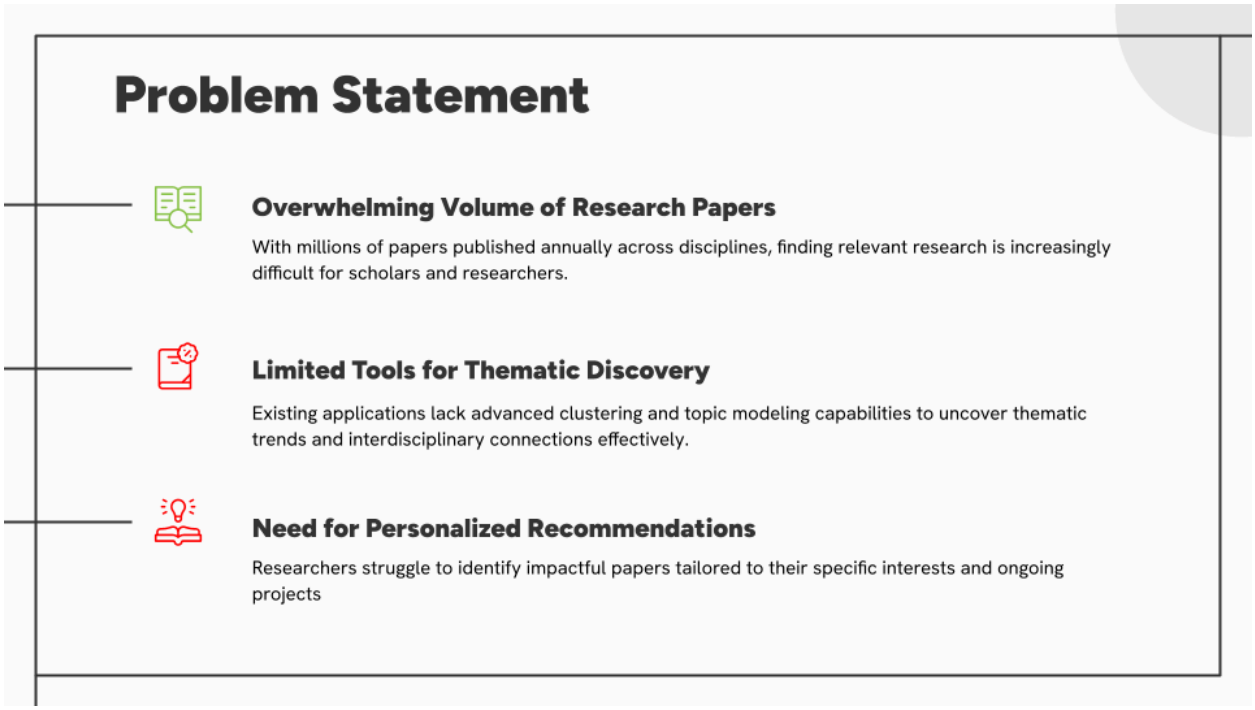
- Isha Singh
- Nisharg Gosai
- Sudarshan Paranjape
- Umang Jain

ReSearchMate aims to revolutionize the way researchers navigate and discover academic literature. By leveraging advanced clustering, topic modeling, and recommendation algorithms, it provides personalized insights and impactful paper recommendations tailored to individual research interests.



Let's Walk Through the Contents

1. **Introduction**
   We'll start by understanding the **Problem Statement** and the **Key Objectives** of the project. This section outlines the challenges faced by researchers in navigating academic literature and introduces our solution, *ReSearchMate*.
2. **Data**
   Next, we'll dive into the **Data Source**, its **Attributes**, and the preprocessing techniques applied to prepare the dataset for analysis. This includes steps like handling missing values, standardizing columns, and cleaning text for meaningful insights.
3. **Clustering Overview**
   This section covers the **Clustering Techniques Used**, including methods like K-Means, along with the **Challenges Faced** in organizing papers into thematic groups.
4. **Topic Modeling**
   We'll explore **Methods & Visuals** from advanced topic modeling techniques such as Latent Dirichlet Allocation (LDA). This section highlights how we captured nuanced research themes and interdisciplinary trends.
5. **Conclusions**
   Finally, we'll present the **Results & Outcomes**, summarizing the impact of the project and discussing potential future directions for *ReSearchMate*

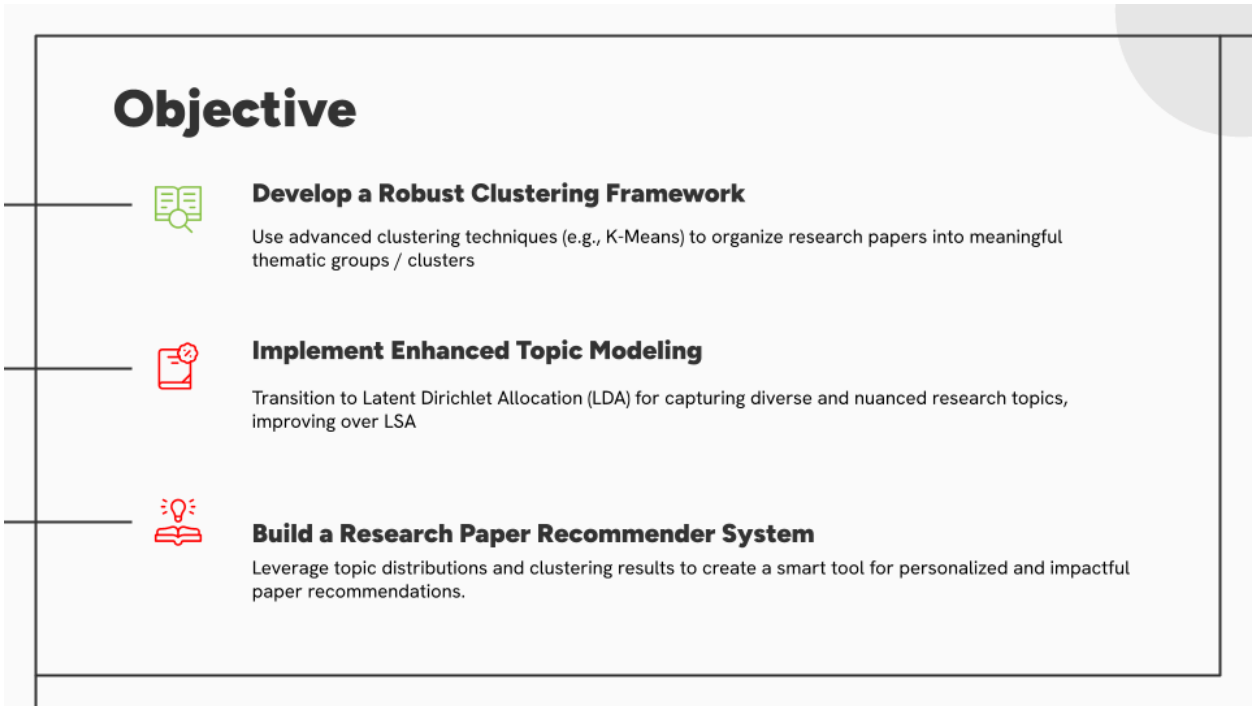


Problem Statement

The sheer volume of research papers published annually across various disciplines poses a significant challenge for scholars and researchers. With millions of papers added to the growing body of academic literature each year, finding relevant and high-quality research has become an overwhelming task. This abundance of information often leads to frustration and inefficiency, as researchers spend valuable time sifting through irrelevant or less impactful studies.

Existing tools for academic search and discovery fall short in addressing this issue effectively. While they provide basic search functions, they often lack advanced clustering and topic modeling capabilities. These capabilities are crucial for uncovering thematic trends and interdisciplinary connections, which can provide fresh perspectives and foster innovative research. Without these features, researchers are left with limited insights into emerging fields and are unable to fully explore the interconnected nature of modern science.

Furthermore, people face difficulties in identifying impactful papers tailored to their specific interests and ongoing projects. Current recommendation systems often fail to deliver personalized suggestions that align with an individual's unique research focus. This gap leaves many scholars struggling to stay updated on critical developments in their areas of expertise, potentially missing key contributions that could influence their work. Addressing this need for personalized, theme-based, and interdisciplinary discovery tools is essential for advancing academic productivity and innovation.

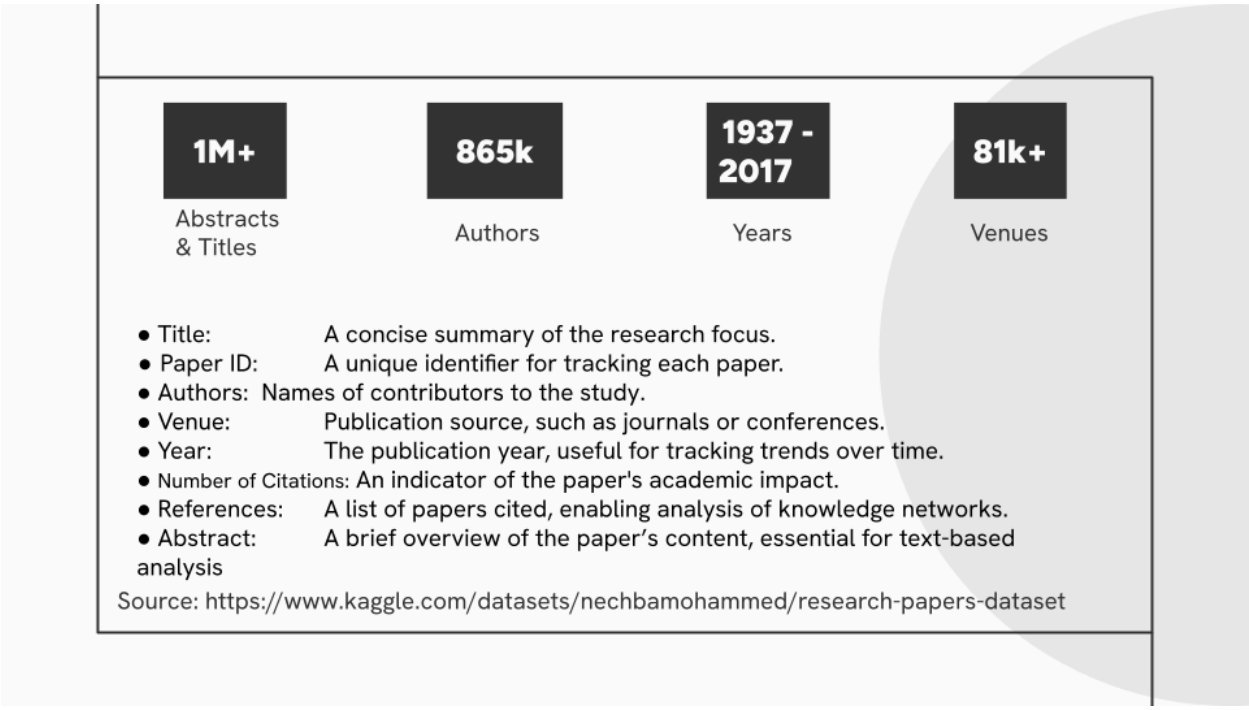


Objective

The objective is to develop a robust clustering framework to streamline the organization of research papers into meaningful thematic groups. Advanced clustering techniques such as K-Means will be employed to categorize papers based on similarities, enabling researchers to quickly identify relevant studies within specific domains. This framework will address the challenge of navigating the overwhelming volume of academic literature by providing a structured and intuitive way to explore research.

To enhance the understanding of nuanced topics, the framework will implement Latent Dirichlet Allocation (LDA) for topic modeling, moving beyond traditional methods like Latent Semantic Analysis (LSA). LDA is better suited for capturing the diversity and depth of research topics, making it a valuable tool for uncovering interdisciplinary connections and emerging trends. This advanced topic modeling approach will provide researchers with deeper insights into thematic trends across large datasets.

The final component of the objective is to build a research paper recommender system that leverages the clustering and topic modeling results. By utilizing topic distributions and thematic groupings, the system will deliver personalized recommendations tailored to a researcher's interests and ongoing projects. This smart tool aims to enhance academic productivity by helping researchers discover impactful papers efficiently, fostering innovation and collaboration across disciplines.



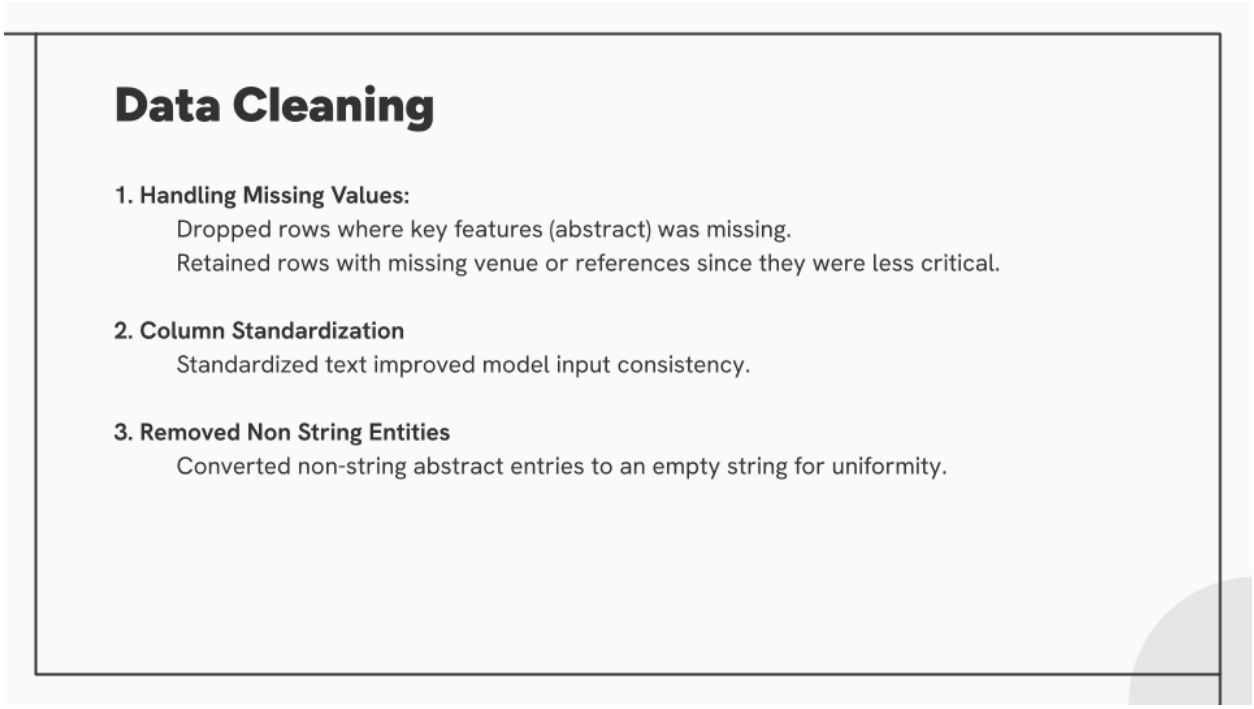
Dataset

The dataset contains over 1 million research papers, with 865,000 unique authors and data spanning from 1937 to 2017. It includes over 81,000 abstracts and titles, providing a substantial foundation for analyzing academic trends and connections. The data offers a rich resource for exploring the evolution of research across decades.

Key metadata includes critical details such as the **title**, offering a concise summary of each paper's focus, and a **paper ID**, a unique identifier for easy tracking. Additional fields include **authors** (contributing researchers), the **venue** (publication source such as journals or conferences), and the **year** of publication, which is crucial for analyzing temporal trends. The dataset also tracks the **number of citations**, serving as a measure of each paper's academic impact, and **references**, which allow for network analysis of the knowledge base.

The **abstracts** and **titles** provide an essential textual foundation for advanced analyses, such as clustering and topic modeling. These elements enable researchers to extract thematic insights and develop tools like personalized recommender systems, making the dataset an invaluable resource for scholarly exploration and innovation.

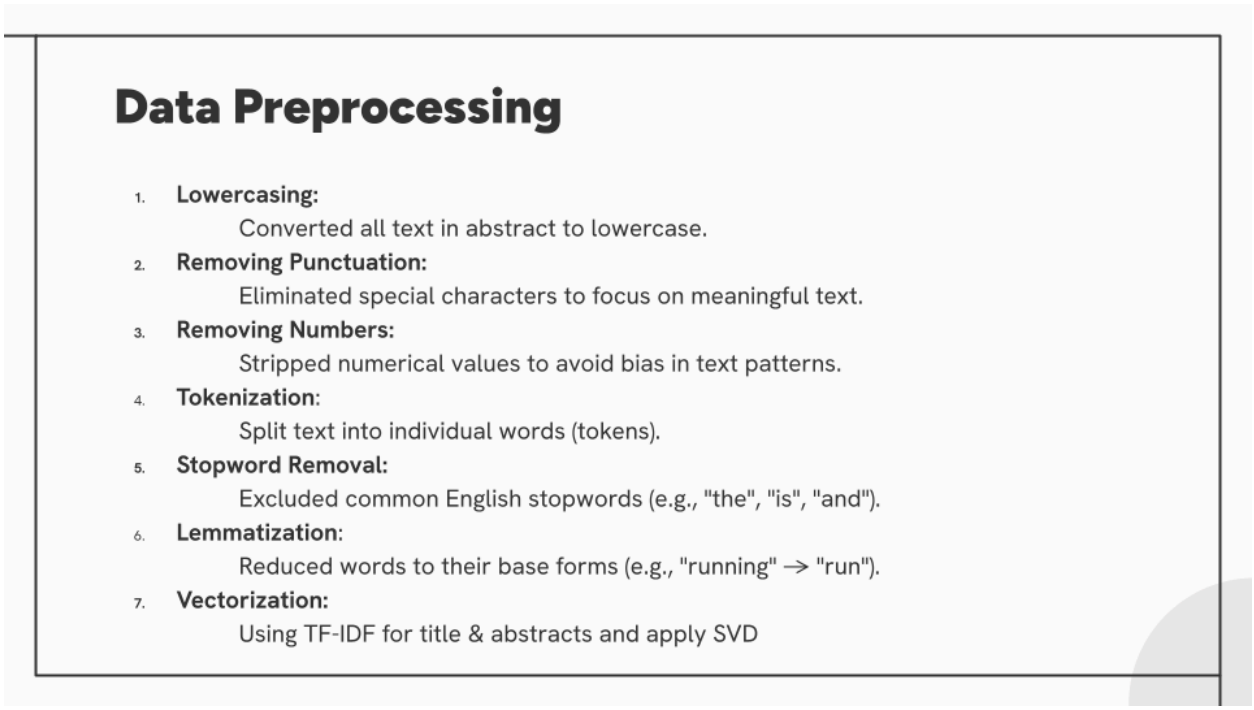
Source :
https://www.kaggle.com/datasets/nechbamohammed/research-papers-dataset



Data Cleaning

1. **Handling Missing Values**
    - Rows with missing **abstracts** were dropped, as abstracts are crucial for text-based analysis.
    - Rows with missing **venues** or **references** were retained since these features are less critical for the core tasks of clustering and topic modeling.
2. **Column Standardization**
    - Text data was standardized to ensure consistency in model inputs. This included normalizing the case, removing extra spaces, and ensuring uniform formats for all text-based columns.
3. **Removed Non-String Entities**
    - Non-string entries in the **abstract** column were converted to empty strings to maintain uniformity. This ensured that downstream processes could handle the data without errors or inconsistencies.

These cleaning steps aimed to optimize the dataset for clustering, topic modeling, and recommendation system development while preserving the integrity of essential information.



Data Preprocessing

1. **Lowercasing**
   - All text in the **abstracts** was converted to lowercase, ensuring uniformity and reducing redundancy during analysis (e.g., "Research" and "research" treated the same).
2. **Removing Punctuation and numbers**
   - Special characters and punctuation marks were eliminated to focus on meaningful textual content and improve tokenization accuracy.
   - Numerical values were stripped from the text to avoid biasing patterns based on irrelevant numeric sequences.
3. **Tokenization & Lemmatization**
   - Text was split into individual words (tokens), forming the basis for further text processing and analysis. And then were reduced to their base forms
4. **Stopword Removal**
   - Common English stopwords (e.g., "the," "is," "and") were excluded to retain only significant terms, improving topic modeling and clustering effectiveness.
5. **Vectorization**
   Text data (titles and abstracts) was transformed into numerical representations using **TF-IDF** (Term Frequency-Inverse Document Frequency).
   - **Singular Value Decomposition (SVD)** was applied to reduce the dimensionality of TF-IDF vectors, retaining the most informative features for clustering and topic modeling.

These steps prepared the text data for advanced analysis by focusing on relevant features, minimizing noise, and optimizing computational efficiency

Clustering techniques were employed to analyze and categorize the research paper dataset based on **Title and Abstract features**

K-Means Clustering:

Description: K-means is a centroid-based clustering algorithm that partitions data into k clusters by minimizing the intra-cluster distance (variance).

Why K-Means?:

- Effective for large datasets.
- Provides interpretable results through defined cluster centroids.
- Computationally efficient compared to hierarchical methods.

Key Insights:

- K-means was effective in capturing distinct themes in the dataset when applied to both Title and Abstract features.
- The use of evaluation metrics such as WCSS, Silhouette Score, and Davies-Bouldin Index ensured the selection of an optimal k, maximizing clustering quality.

Agglomerative Clustering:

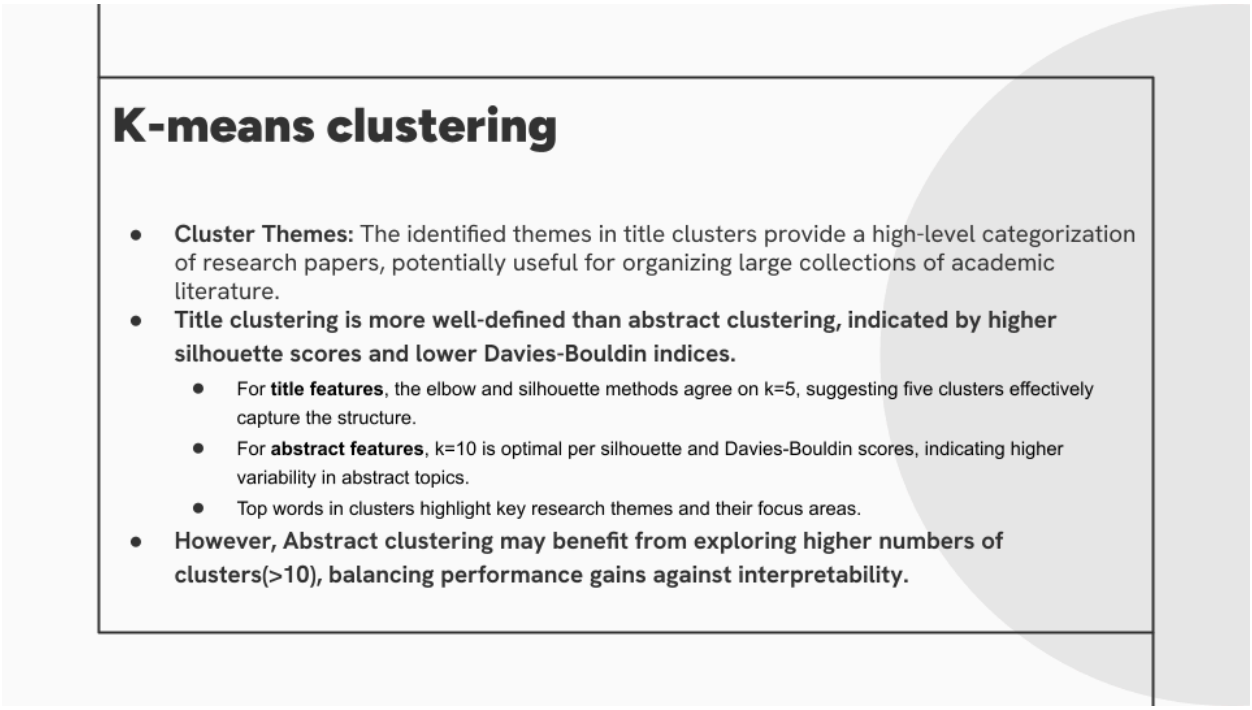
Description: Agglomerative clustering is a hierarchical technique that builds clusters through a bottom-up approach, iteratively merging the closest pairs of clusters based on a linkage criterion.

Why Agglomerative Clustering?:

- Provides a hierarchical view of the data, which is useful for understanding relationships between clusters.
- Capable of handling different cluster shapes and densities.

Key Insights:

- Agglomerative clustering complemented the K-means results by offering a hierarchical structure, allowing further exploration of sub-themes within the data.
- Different Linkage methods were experimented with to optimize results.



K-means clustering was applied to the Title and Abstract separately, and experiments were run to find the optimal value of k for both the title and abstract. Experiments were run for **k values from 2 to 10** iteratively and below scores were recorded for each k value

1) K value
2) WCSS
3) Silhouette Score
4) Davies-Bouldin Score

A **custom function** was created that provides a structured approach to evaluate the optimal number of clusters using three widely recognized metrics: **Within-Cluster Sum of Squares (WCSS), Silhouette Score, and Davies-Bouldin Index.**

Key Steps in the Function:

1. **Elbow Method for WCSS:**
    - Calculate the first differences (deltas) and second differences (double_deltas) of the WCSS values.
    - Identify the index where the second difference is maximized, corresponding to the elbow point.
    - Set elbow_k as the number of clusters at this elbow point.
2. **Silhouette Score:**
    - Find the index of the maximum Silhouette Score and retrieve the corresponding number of clusters.
3. **Davies-Bouldin Index:**
    - Find the index of the minimum Davies-Bouldin Index and retrieve the corresponding number of clusters.

By combining these three metrics, the function provides a comprehensive way to evaluate the optimal number of clusters.

The **output** of the custom function is below
Optimal clusters for Title:
{'elbow_k': 5, 'silhouette_k': 5, 'davies_bouldin_k': 2}.
Optimal clusters for Abstract:
{'elbow_k': 3, 'silhouette_k': 10, 'davies_bouldin_k': 10}

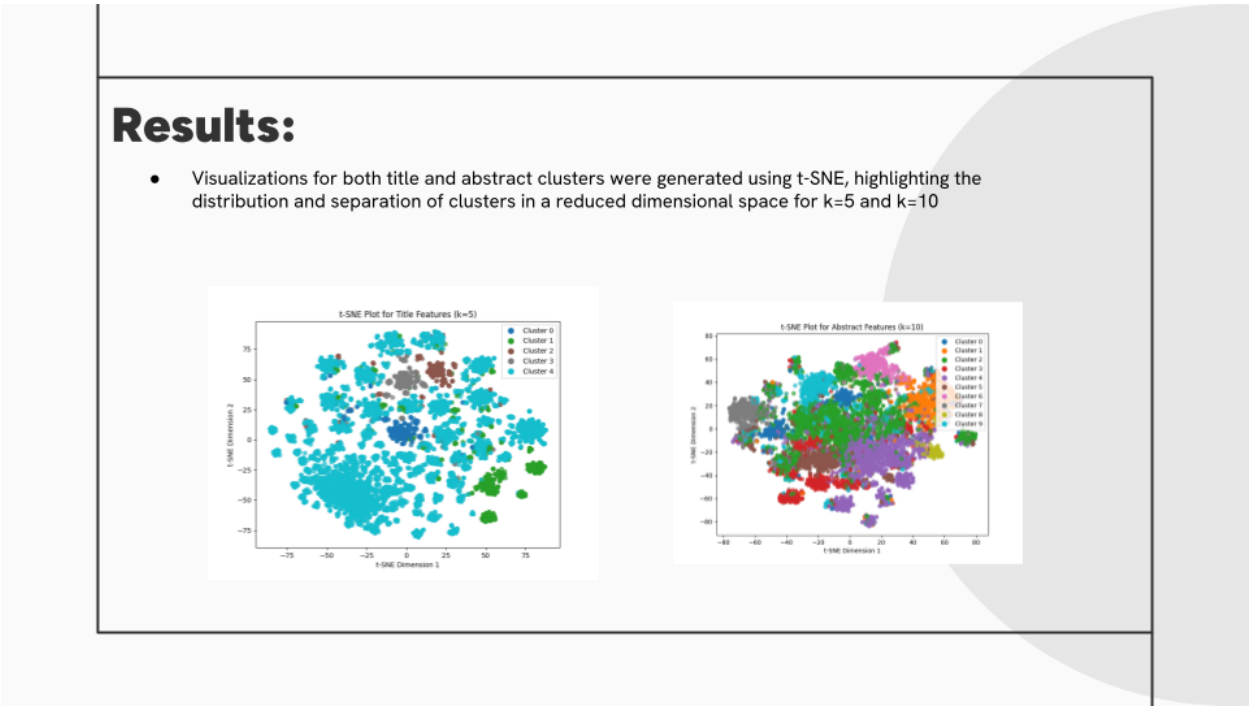
Based on the above output we decided to **select k=5 as the optimal k value for Title** as the elbow point was found to be at k=5 and the silhouette score was also the highest at k=5

F**or Abstract, k=10 as the optimal k value** was selected as the silhouette score was the highest at k=10 and the Davies Bouldin score was the lowest at k=10.

However, it is important to note that since the experiments were run for k values from 2 to 10, the k=10 value is the upper limit and therefore should be considered as an optimal value within the scope.

Future work could include experimenting with higher k values for abstract features.

The above slide shows how k-means captured themes in the data. **The identified themes in the title and abstract clusters provide a high-level categorization of research papers, useful for organizing large collections of academic literature.**



The slide above shows two **t-SNE** (t-distributed Stochastic Neighbor Embedding) visualization plots representing title and abstract feature analyses:

Abstract Features Plot

The plot displays clustering results with k=10, showing a complex distribution of data points across 10 distinct clusters. The data points are spread across a 2D space ranging from approximately -80 to +80 on both dimensions. Notable characteristics include:

- Dense cluster formations in multiple regions
- Significant overlap between some clusters, particularly in the central region
- Clear separation of some clusters, especially those at the periphery

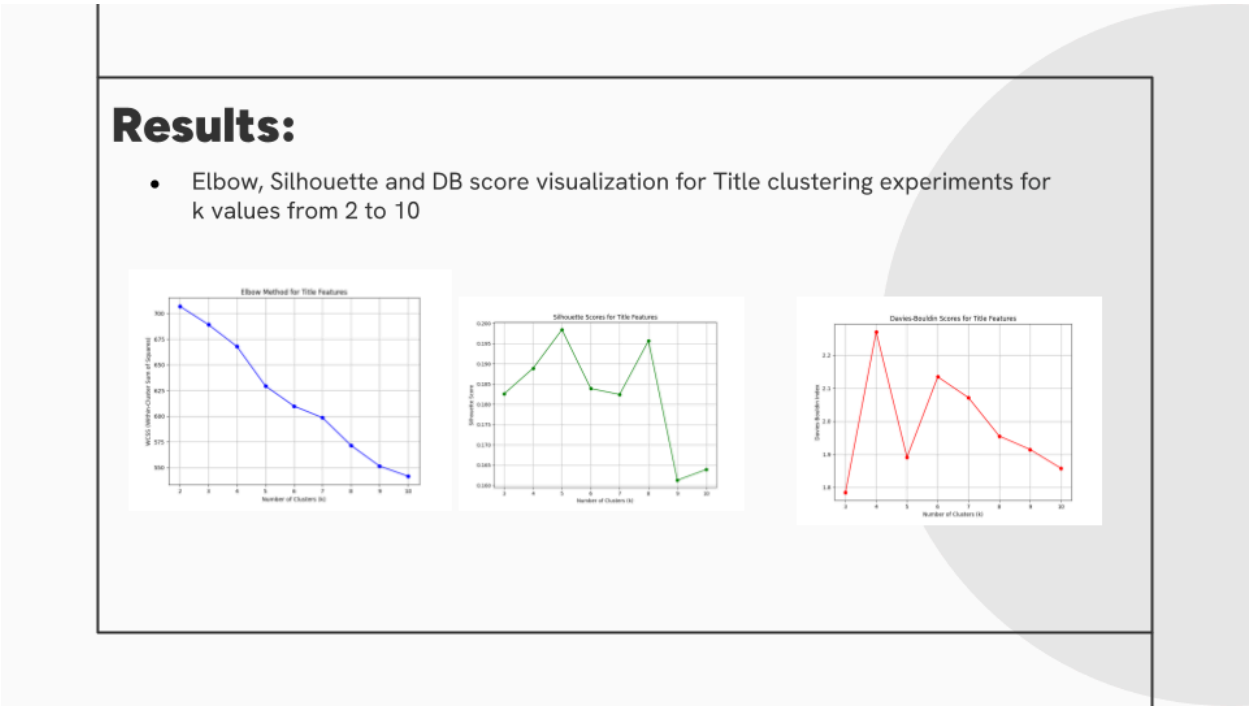
Title Features Plot

The plot shows an optimal clustering with k=5, revealing:

- More concentrated grouping patterns
- Dominant presence of turquoise-colored clusters (Cluster 4)
- Fewer overlapping regions compared to the abstract features
- Data points are distributed in a similar range but with different cluster density patterns
- Clear separation of the green clusters (Cluster 1) in the lower right region

Both visualizations demonstrate dimensionality reduction from high-dimensional space to 2D while preserving local relationships between data points. The different number of clusters (k=10 vs k=5) suggests that the title features might have more natural grouping tendencies compared to the more complex abstract features.

The plots represent abstracts showing more complex relationships requiring finer clustering granularity, while titles demonstrate more distinct and separable features.



In the above slide, the three plots show different evaluation metrics for determining the optimal number of clusters (k) for **Title features** in a clustering analysis:

Silhouette Score Analysis

The Silhouette score plot shows two notable peaks:

- First peek at k=5 with a score of ~0.198
- Second peak at k=8 with a score of ~0.195
- Sharp decline after k=8, dropping to ~0.161 at k=9

This suggests that either 5 or 8 clusters could be optimal configurations, with k=5 showing slightly better cohesion and separation.

Elbow Method Analysis

The Within-Cluster Sum of Squares (WCSS) plot shows:

- Steady decrease from k=2 (~705) to k=10 (~540)
- Notable "elbow" point around k=4-5
- Diminishing returns in reduction of WCSS after k=5

This indicates that k=5 might be a good balance between cluster complexity and explained variance.

Davies-Bouldin Score Analysis

The Davies-Bouldin index reveals:

- General downward trend after k=6
- Lowest score at k=10 (~1.85)

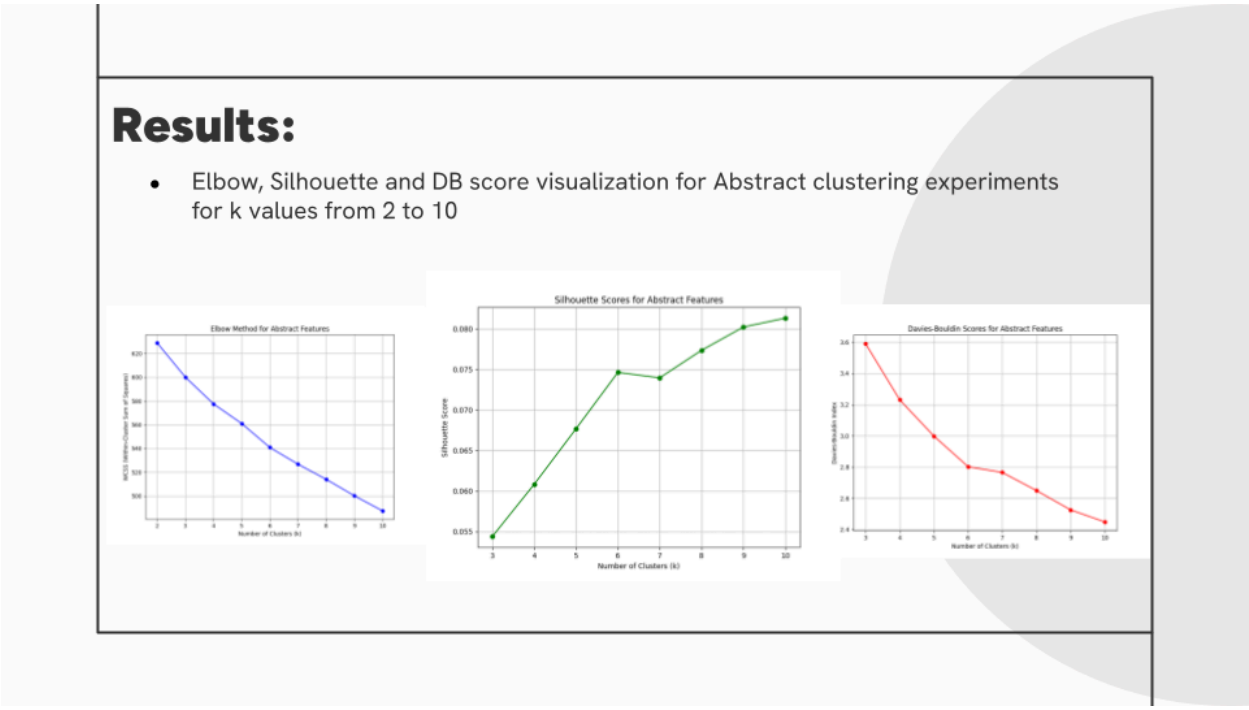
Lower Davies-Bouldin scores indicate better clustering, suggesting that higher k values provide better cluster separation, though this should be balanced against model complexity.

Considering all three metrics together, **k=5 appears to be the optimal choice** because of it:

Achieves the highest Silhouette score

Corresponds to the elbow point in the WCSS plot

Offers a reasonable trade-off before the secondary peak in Davies-Bouldin scores



The above slide has three plots that show evaluation metrics for determining the optimal number of clusters (k) for **Abstract features** in a clustering analysis:

Elbow Method Analysis

The WCSS plot demonstrates:

- Steady decrease from k=2 (~630) to k=10 (~490)
- Gradual curve without a sharp elbow point
- The most significant drop occurs between k=2 and k=4
- Diminishing returns in reduction after k=6

Silhouette Score Analysis

The Silhouette score shows:

- Continuous upward trend from k=3 (0.054) to k=10 (0.081)
- Notable increase between k=5 and k=6
- Small plateau around k=6-7
- Highest score at k=10, suggesting better cluster separation with more clusters

Davies-Bouldin Score Analysis

The Davies-Bouldin index displays:

- Consistent downward trend from k=3 (3.6) to k=10 (2.4)
- The steepest decline between k=3 and k=5
- More gradual decrease after k=6
- Lower scores indicate better clustering separation

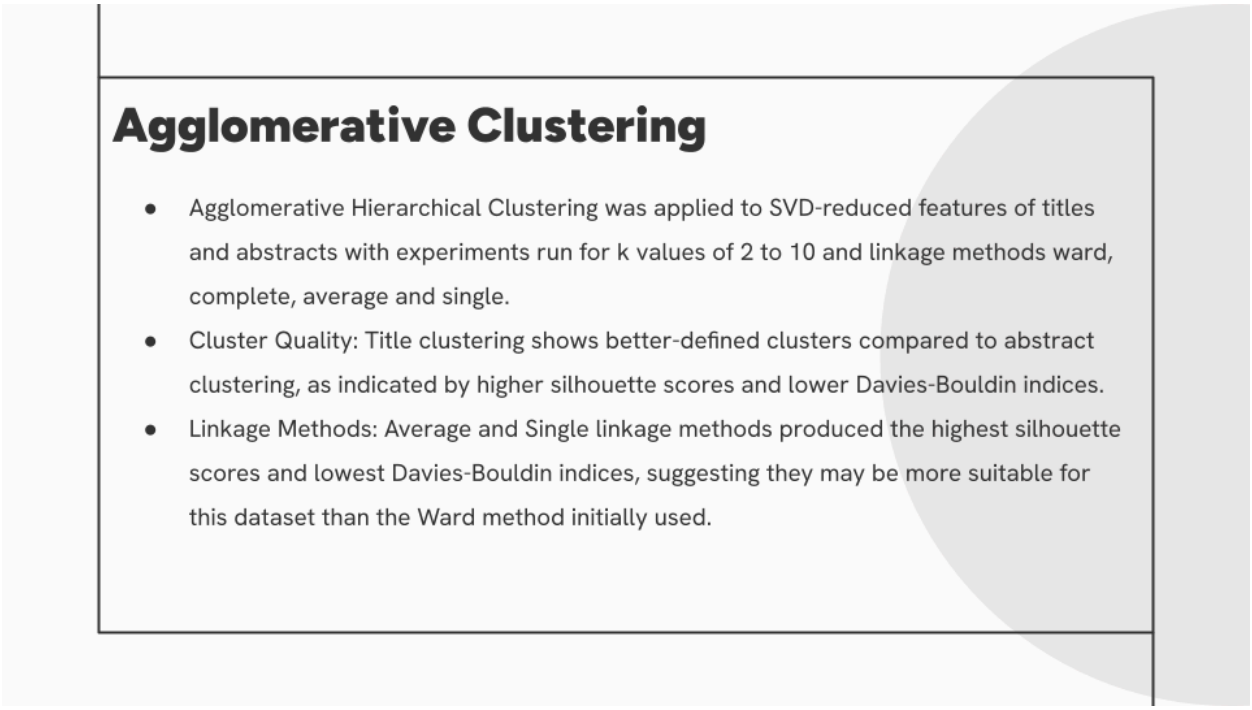
Unlike the title features analysis, the abstract features show a less clear-cut optimal k value. The metrics suggest:

Higher k values (8-10) might be more appropriate for abstract features

More complex clustering structure compared to titles

No single obvious optimal k value, but k=10 shows promising results across metrics

The trade-off between model complexity and cluster quality needs to be considered



Agglomerative clustering was applied to TF-IDF vectorized and SVD-reduced research paper titles and abstracts to identify thematic groupings.

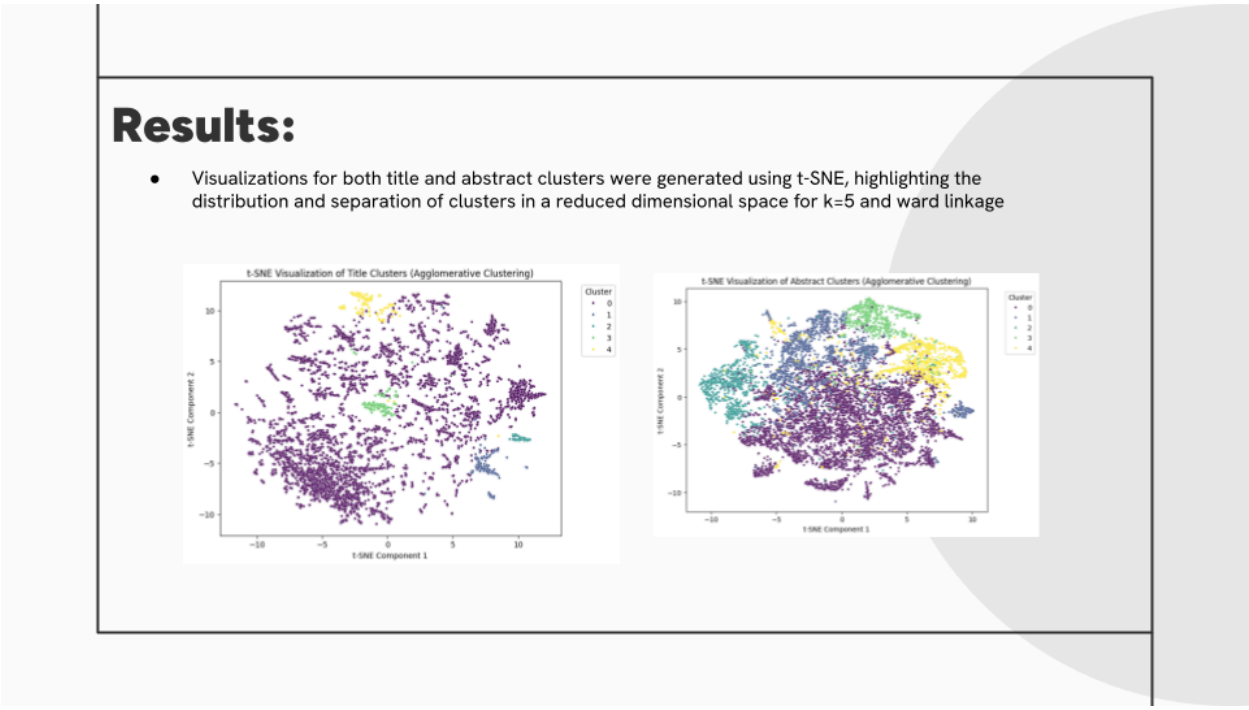
Clustering performance was evaluated for cluster sizes (k=2 to k=10) using four linkage methods: **ward, complete, average, and single.**

The models were assessed using the **Silhouette Score** (measuring cluster cohesion and separation) and the **Davies-Bouldin Index** (evaluating compactness and separation).

For Title features, the optimal configuration was achieved with **k=2** clusters using the **average linkage method**, yielding a Silhouette Score of 0.634 and a Davies-Bouldin Index of 0.254, indicating well-separated, compact clusters.

For Abstract features, the best Silhouette Score (0.400) was observed with k=2 and the **single linkage method**, while the lowest Davies-Bouldin Index (0.449) occurred at k=3, suggesting a trade-off between cohesion and compactness.

These results demonstrate agglomerative clustering's ability to capture meaningful patterns in the data and identify optimal configurations for grouping research titles and abstracts based on thematic content.



The above slide shows two t-SNE visualization plots which show the results of agglomerative clustering applied to abstract and title features:

Abstract Features Clustering (k=5)

The plot reveals:

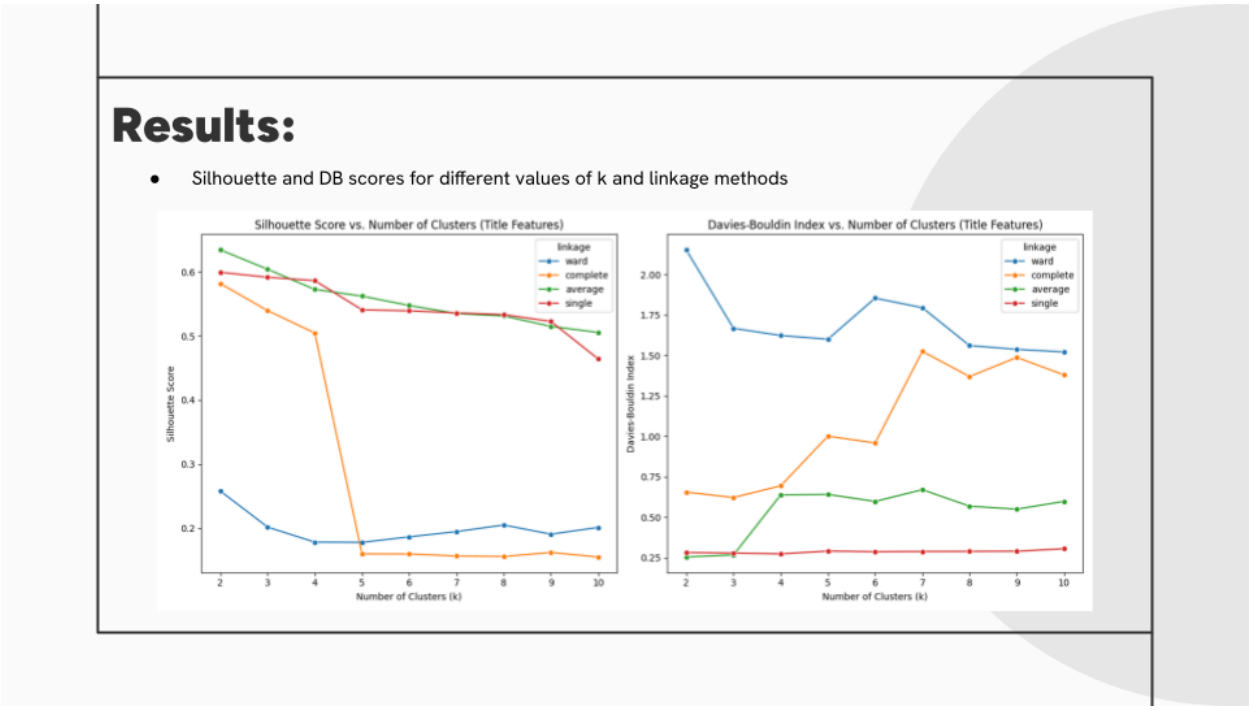
- Five distinct clusters with varying sizes and densities
- Large central purple cluster (Cluster 0) showing high density
- Well-separated yellow cluster (Cluster 4) on the right side
- Distinct turquoise cluster (Cluster 2) on the left
- Green cluster (Cluster 3) showing clear separation in the upper right
- Some overlap between clusters, particularly near the central region
- Data points spread across a range of approximately -10 to +10 in both dimensions

Title Features Clustering (k=5)

The plot demonstrates:

- More dispersed distribution with less defined boundaries
- Dominant purple cluster (Cluster 0) spread across multiple regions
- Small, isolated clusters for other groups (1-4)
- A higher degree of scatter compared to abstract clustering
- Less cohesive cluster formation
- Similar dimensional range (-10 to +10) as abstract features

The abstract features show more coherent clustering patterns with clearer boundaries between groups, while title features exhibit more fragmented clustering with scattered subgroups. This suggests that abstract content may contain more structured and distinguishable patterns compared to titles, which show higher variability and less distinct grouping tendencies.



The above slide shows a comprehensive analysis of agglomerative clustering performance on title features for different values of k using multiple evaluation metrics:

Silhouette Score Analysis

The left plot shows Silhouette scores for different linkage methods:

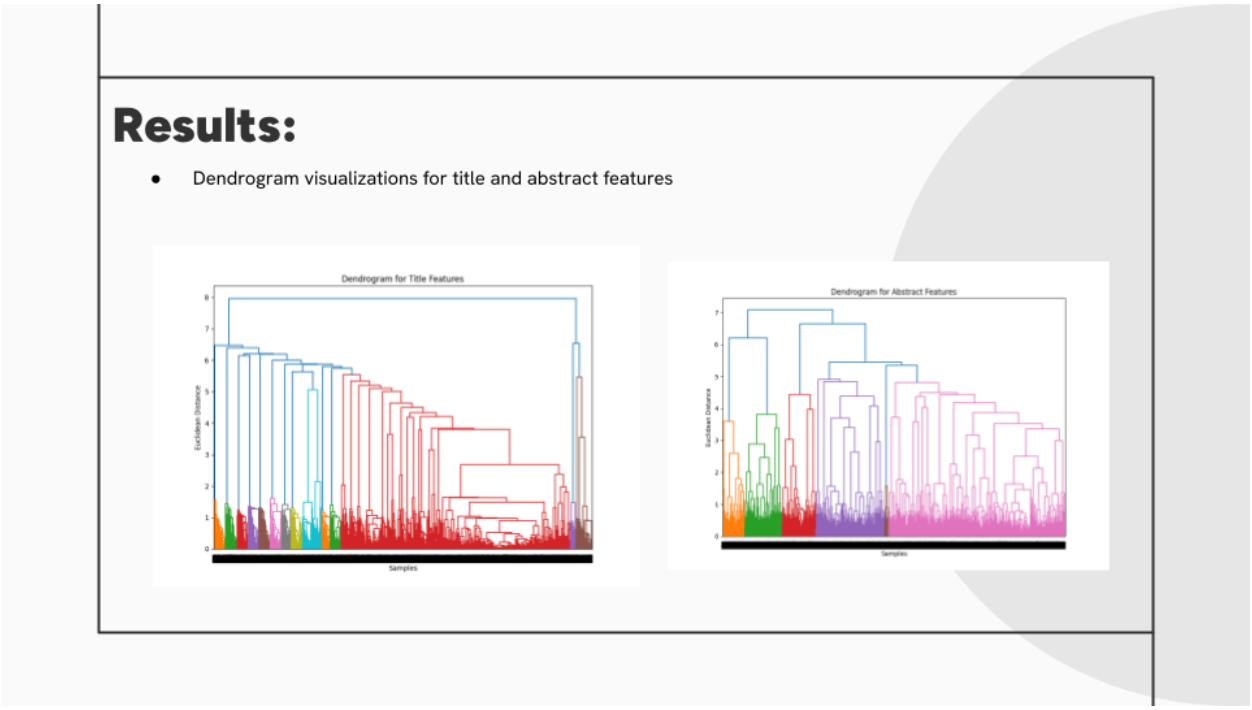
- Average and single linkage methods perform best, starting at ~0.65 for k=2
- Complete linkage shows a dramatic drop after k=4
- Ward linkage consistently shows the lowest scores (~0.2)
- All methods show a general declining trend as k increases

Davies-Bouldin Index Analysis

The right plot reveals:

- Ward linkage starts highest (~2.1) and gradually decreases
- Complete linkage shows a significant increase after k=4
- Single linkage remains most stable (~0.3)
- Average linkage maintains moderate scores throughout

The metrics suggest different optimal cluster numbers depending on the method:
Silhouette scores favor lower k values (2-4)
Davies-Bouldin index suggests better separation at higher k values for ward linkage
The elbow plot indicates diminishing returns after k=6
Single and average linkage methods generally outperform ward and complete linkage



In the above slide, the two dendrograms reveal distinct clustering patterns between title and abstract features.

Title features exhibit a more complex branching pattern with multiple small clusters merging at various heights, indicating a more nuanced relationship between features. In contrast, the abstract features show more distinct, cohesive groupings with clearer separation between major clusters.

Exploration of Other Techniques

Other clustering methods were also evaluated but were found to have limitations in this specific context:

DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

Strengths: DBSCAN is robust to outliers and works well for clusters of varying densities.

Limitations:

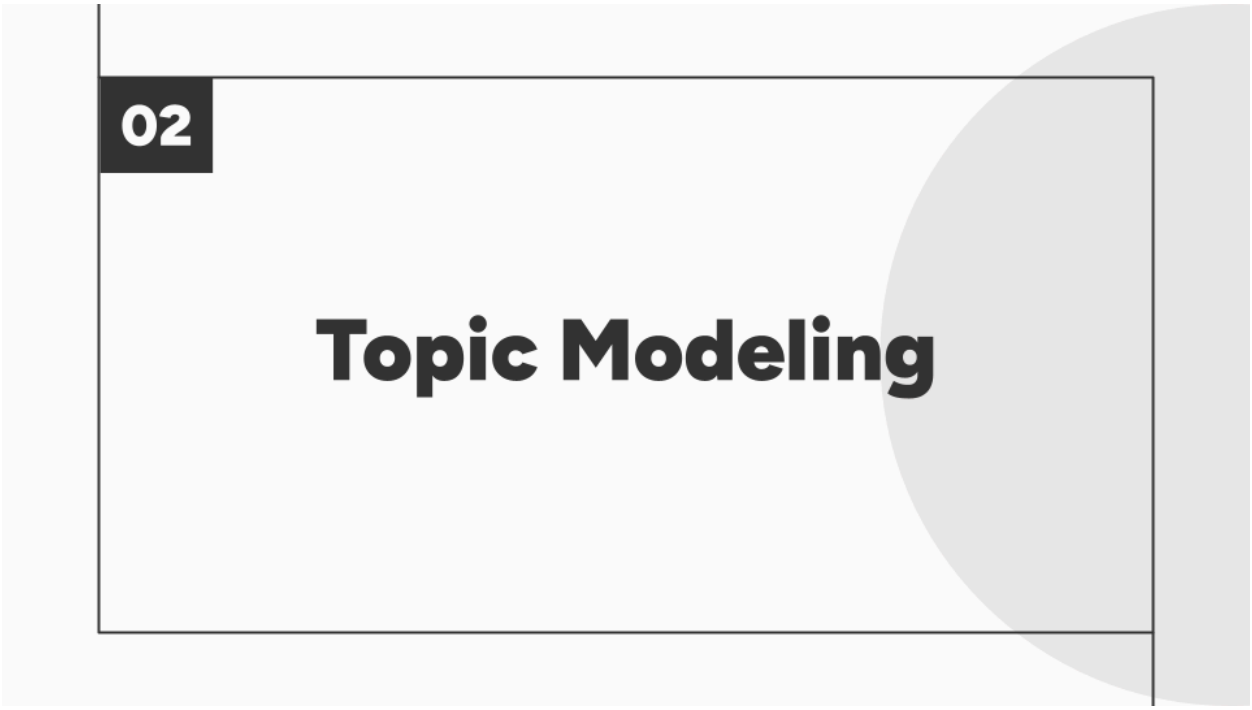
- Struggled with the high dimensionality of the TF-IDF vectorized data.
- Required extensive parameter tuning (e.g., epsilon and min_samples), and results were inconsistent across different subsets of the data.
- Failed to identify meaningful clusters due to the sparse nature of the dataset.

Spectral Clustering:

Strengths: Suitable for data with complex, non-convex cluster structures.

Limitations:

- Computationally intensive for large datasets due to eigenvalue decomposition.
- Did not perform well after dimensionality reduction, as it relies heavily on the affinity matrix, which can degrade in high-dimensional spaces.
- Clusters formed needed to be more well-defined and interpretable for the given features.



Topic modeling is an unsupervised machine-learning technique used to extract hidden themes or topics from large collections of textual data. This approach is widely used in applications such as information retrieval, document classification, and recommendation systems.

By grouping similar documents under common topics, topic modeling enables efficient navigation and personalized content recommendations, making it a cornerstone for text-based applications. All topic models are based on the same basic assumption:

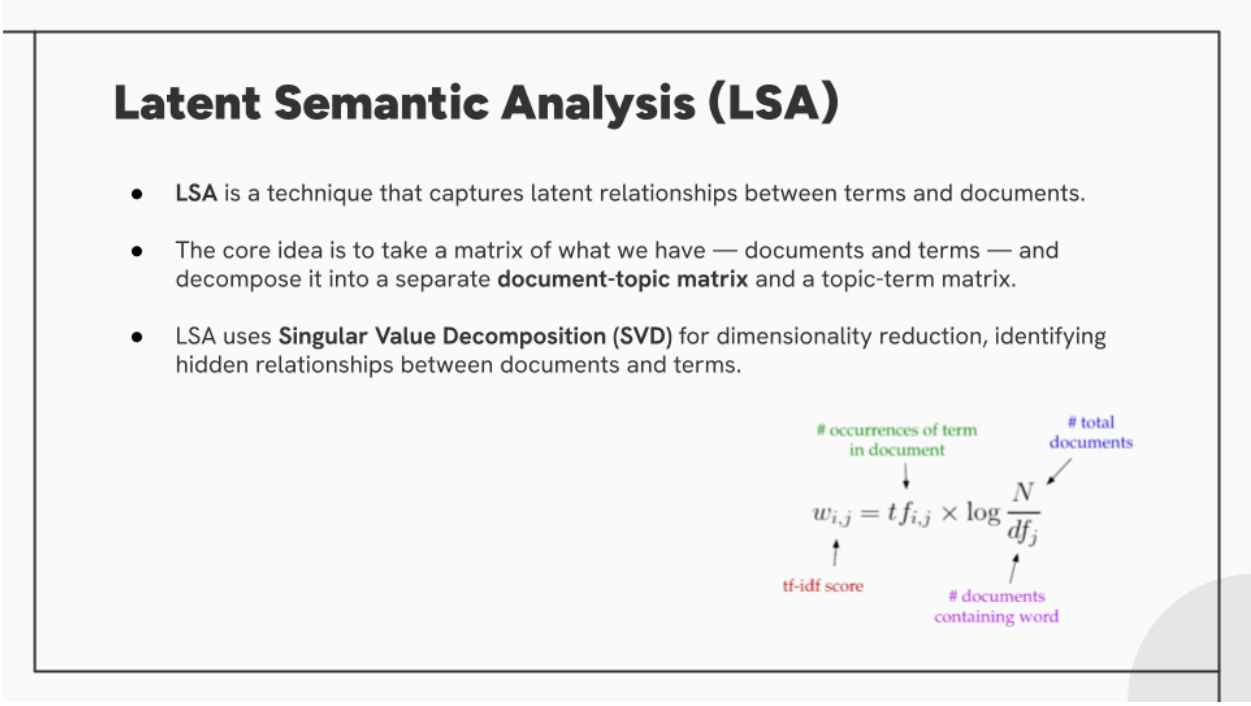
- Each document consists of a mixture of topics, and
- Each topic consists of a collection of words.

In other words, topic models are built around the idea that the semantics of our document are governed by some hidden, or “latent,” variables that we are not observing. As a result, the goal of topic modeling is to uncover these latent variables — topics — that shape the meaning of our document and corpus. In this project, we discover 2 types of topic models - LDA and LSA.

Why LSA and LDA?

- LSA was chosen for its efficiency in grouping papers into initial clusters.
- LDA was used as a follow-up to refine and diversify the topic distributions for better recommendations.

By combining LSA and LDA, we leveraged their respective advantages to achieve a balance between computational efficiency and topic modeling accuracy.



LSA Overview

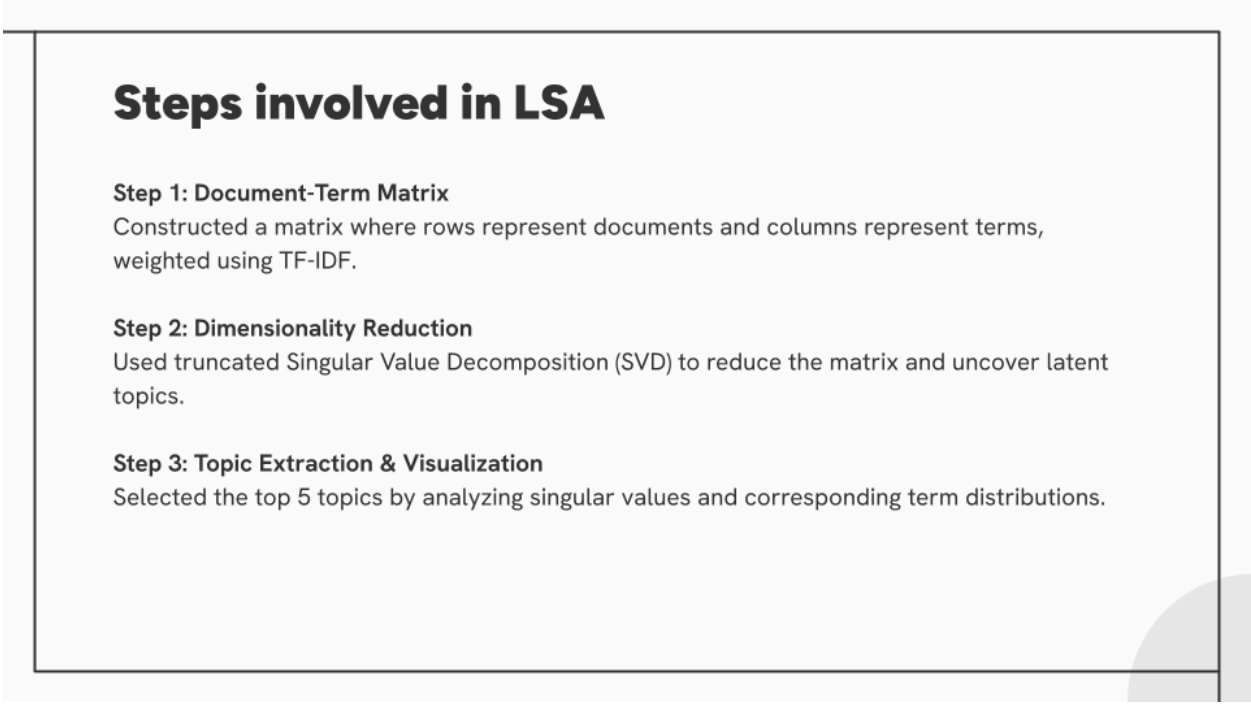
Latent Semantic Analysis (LSA) is a linear algebra-based method for topic modeling that leverages Singular Value Decomposition (SVD). It reduces high-dimensional text data into a lower-dimensional latent space. It identifies patterns in the matrix to discover hidden structures and relationships between terms and documents.

Characteristics:

- Efficient for dimensionality reduction and initial exploration of themes.
- Provides interpretable results by ranking words and documents based on their contributions to topics.

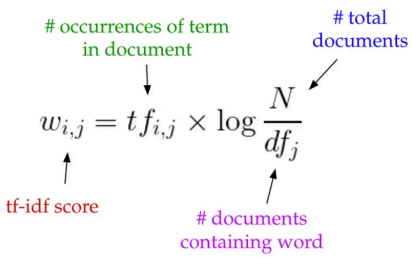
The core idea is to take a matrix of what we have — documents and terms — and decompose it into a separate document-topic matrix and a topic-term matrix.

LSA is particularly useful for dimensionality reduction, clustering, and improving the interpretability of text data. However, its reliance on linear relationships and inability to capture contextual nuances limit its applicability to complex semantic data.



LSA was used to extract topics and organize papers thematically. Here's how it was applied:

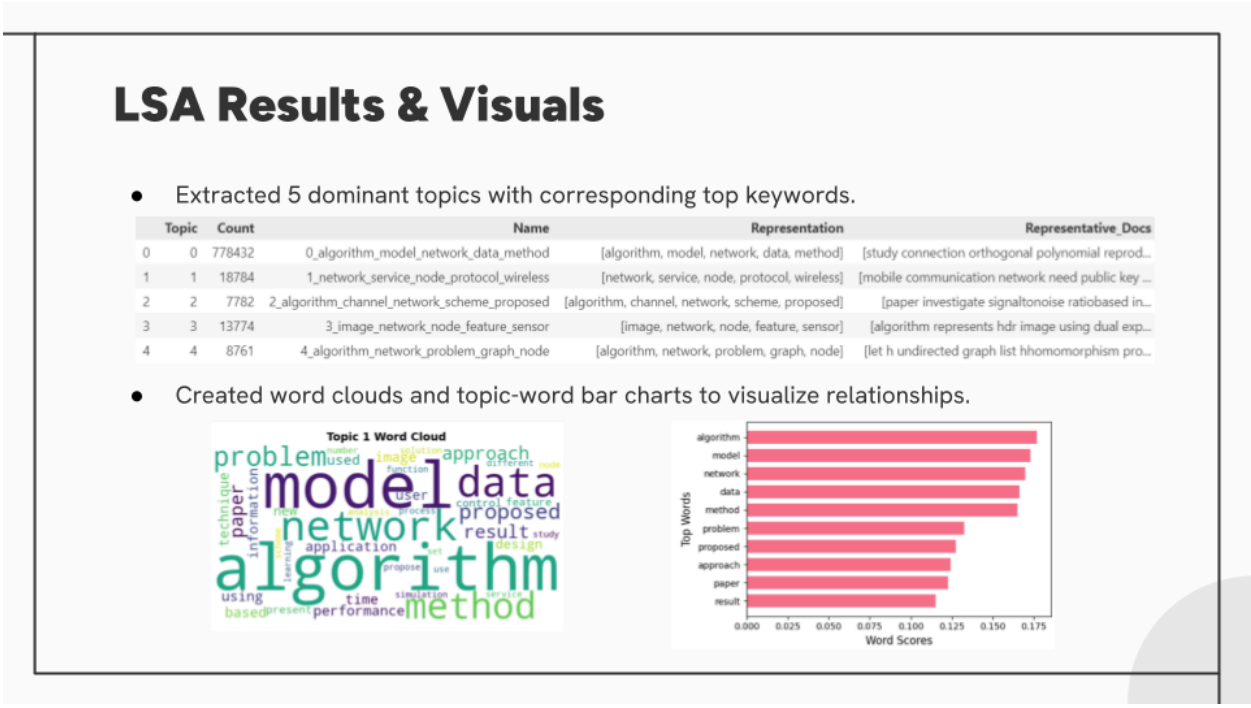
Step 1) Document-Term matrix:
The first step is generating our document-term matrix. Given m documents and n words in our vocabulary, we can construct an m × n matrix A in which each row represents a document and each column represents a word. In the simplest version of LSA, each entry can simply be a raw count of the number of times the j-th word appeared in the i-th document. In practice, however, raw counts do not work particularly well because they do not account for the significance of each word in the document. Hence, our LSA model replaces raw counts in the document-term matrix with a tf-idf score. **Tf-idf, or term frequency-inverse document frequency**, assigns a weight for term j in document I as shown:



Step 2) Dimensionality Reduction:
Since our document matrix A is very sparse, dimensionality reduction can be performed using truncated SVD. **SVD, or singular value decomposition,** is a technique in linear algebra that factorizes any matrix M into the product of 3 separate matrices:

- U: Document-topic matrix.
- S: Singular values representing topic importance.
- V: Term-topic matrix.

Step 3) Topic Extraction and Visualization:
The final step is to select the top 5 topics by analyzing singular values and corresponding term distributions and visualize each topic in detail using word clouds and bar charts.



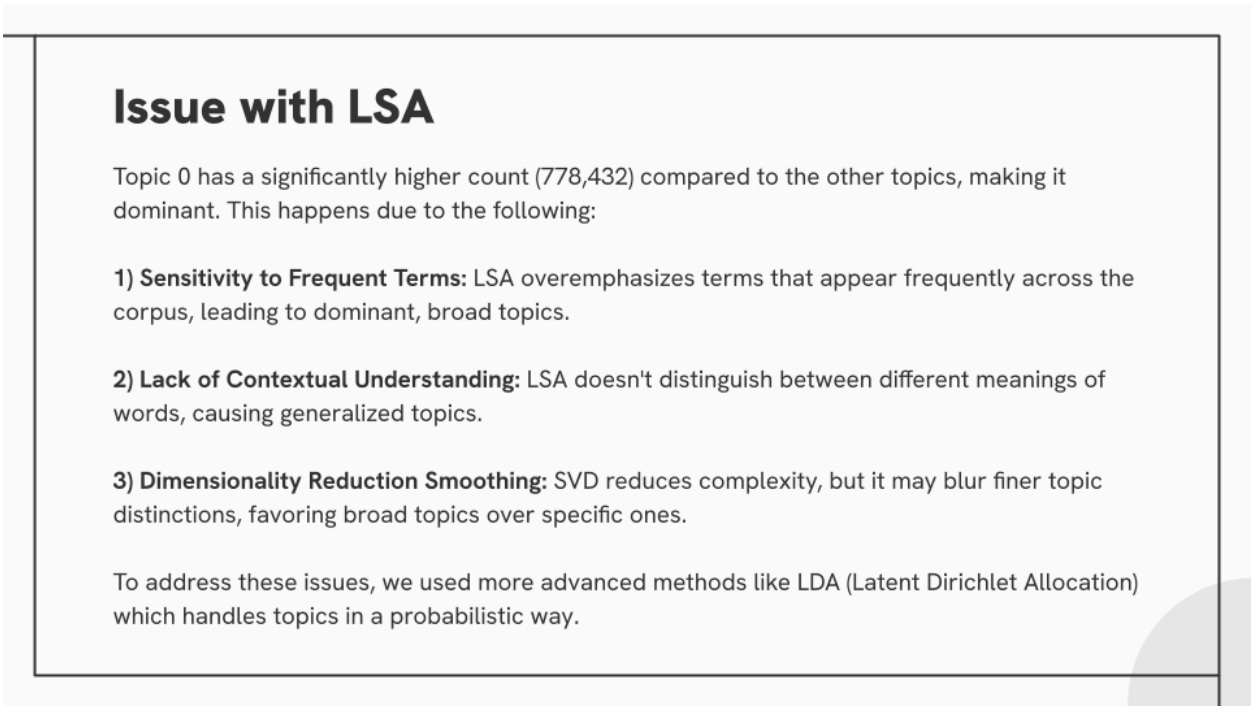
Here is a summary of the results obtained by performing LSA:

1. **Summary Dataframe:** Provides an overview of the topics identified through LSA.
    - **Topics:** Sequentially numbered unique identifiers (e.g., Topic 0, Topic 1).
    - **Count**: Number of documents associated with each topic.
    - **Name**: Top 5 dominant words defining the core theme of each topic.
    - **Representation**: A broader set of words providing context about the topic.
    - **Representative Documents:** Sample documents illustrating the content linked to each topic.

2. **Word Clouds:** The word clouds help visualize the most dominant terms in each topic, showcasing the keywords that define the central themes of the respective topics. Larger words in the cloud correspond to terms that are more strongly associated with the respective topic.

3. **Bar Chart:** The bar charts represent the top words associated with each of the five identified topics in the dataset. Each chart highlights the most significant terms that define a topic, based on their scores derived from the term-topic matrix. The height of each bar reflects the 'weight' or importance of a word in describing the respective topic. A higher score indicates that the word contributes more to the semantic meaning of the topic. For example,

    - **Topic 1:** Dominated by terms like 'algorithm,' 'model,' and 'network.' Suggests a focus on computational methods or systems.
    - **Topic 2:** Includes words like 'network,' 'service,' and 'protocol.' Indicates this topic revolves around communication technologies or services.



Topic Dominance: Topic 0, with a count of 778432, is significantly larger than other topics, highlighting its prominence in the dataset. The dominance of Topic 0 in LSA can be attributed to some inherent drawbacks of the LSA method:

1. **Sensitivity to Frequent Terms (High Term Frequency):**

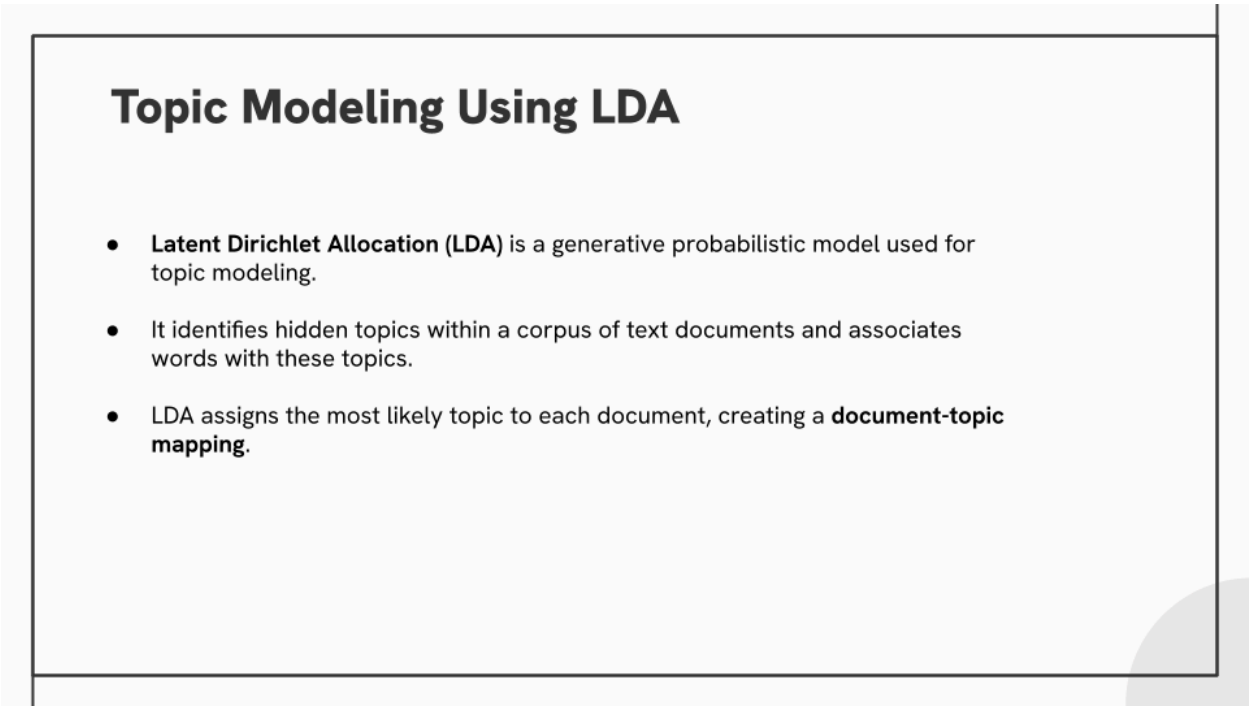
LSA is heavily influenced by the frequency of terms across the entire corpus. In cases where certain terms like "algorithm," "network," and "data" appear frequently, they dominate the singular value decomposition (SVD) process, leading to a topic that is overly broad and dominant. This is because LSA relies on singular values that capture global patterns, and frequently occurring terms often dominate these patterns, overshadowing more specific topics.

2. **Lack of Semantic Context (No Word Sense Disambiguation):**

LSA does not account for multiple meanings of a word (polysemy). For example, the word "network" can refer to various concepts such as computer networks, social networks, or biological networks. LSA treats these words as a single, global term without distinguishing between their different senses, which can result in overly general topics that do not capture domain-specific nuances.

3. **Dimensionality Reduction Limitations:**

In LSA, dimensionality reduction via SVD is used to find latent structures in the data. While this is effective in identifying broad patterns, it can also smooth out finer, domain-specific distinctions. As a result, general topics that have broad term distributions (like Topic 0) can dominate, while niche or specialized topics with more concentrated term distributions may fail to emerge clearly.

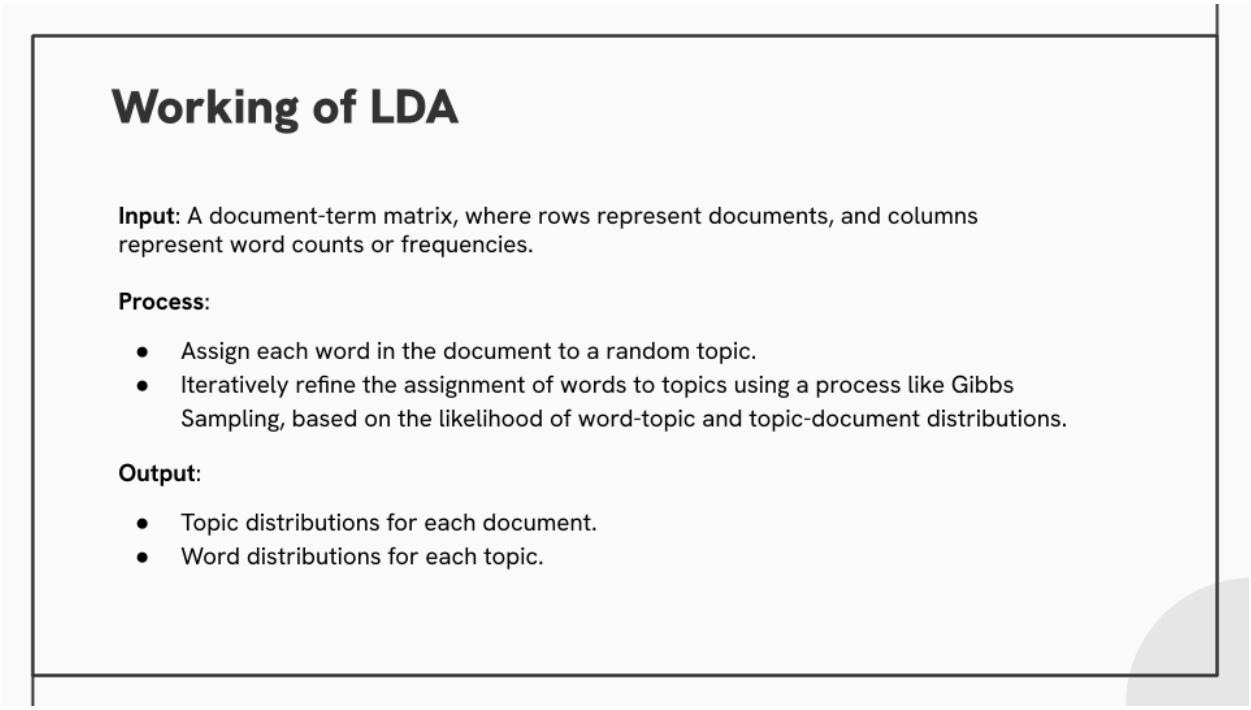


The use of topic modeling and Latent Dirichlet Allocation (LDA) was driven by the objective of uncovering latent topics in a collection of document abstracts and segmenting them based on shared topics. This approach provides valuable insights into the hidden motifs and trends within the data.

LDA was chosen for several reasons:

1. **Probabilistic Modeling**:
   LDA analyzes the words in each document and assigns probabilities to various topics, enabling the automatic discovery of themes without requiring predefined labels.
2. **Soft Clustering**:
   This method allows documents to belong to multiple topics, offering a more flexible and precise understanding of the data.
3. **Unsupervised Technique**:
   LDA does not rely on prior knowledge of the topics. The model independently deduces the topics from the dataset.

By employing the LDA technique, documents were organized into logical groups, revealing underlying topics. This automated approach yielded insights that would be speculative or challenging to achieve through manual analysis.



The functioning of Latent Dirichlet Allocation (LDA) can be described in three major steps:

1. **Input**:
   LDA begins with a document-term matrix, where each row represents a document, and each column represents the frequency of words within those documents.
2. **Process**:
   Initially, each word in the document is assigned to a topic at random. The algorithm then refines these initial assignments using Gibbs Sampling, a Bayesian technique. This iterative process updates the word-topic and topic-document assignments along with their probabilities, improving the results over time.
3. **Output**:
   After several iterations, the output consists of:
   - **Topic Distributions**: Indicating the proportion of each topic within every document.
   - **Word Distributions**: Representing the probabilities of each word occurring within a specific topic.

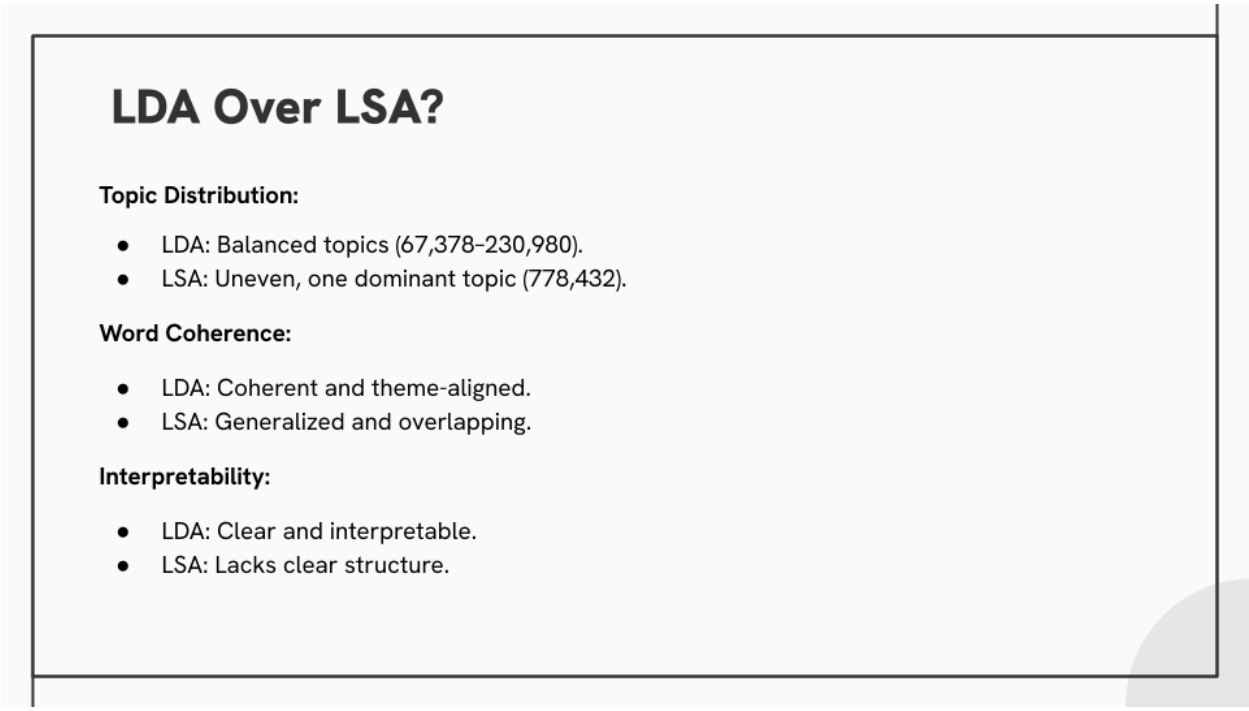
In summary, LDA unveils the underlying themes in the text and establishes connections between words and topics.

The project involved analyzing and visualizing the content to create a descriptive and visual representation of topic modeling results using LDA (Latent Dirichlet Allocation) and LSA (Latent Semantic Analysis):

- **Table**: A table was created for each topic, including the count of documents, keywords describing the theme, and representative documents which are a sample of documents that best exemplify each topic.

- **Word Cloud**: The **word cloud** is a visual representation of the most frequent keywords within **Topic 5**. Larger words indicate higher frequencies, emphasizing their importance in defining the topic. Key terms such as **"user," "data,"** and **"information"** are dominant, highlighting their central role in this topic. Other significant words like **"technology," "service,"** and **"research"** suggest sub-themes or related contexts, giving a broad view of the topic's focus.

- **Bar Graph**: The **bar plot** displays the **top word scores** for Topic 5, reflecting the relevance of individual words to the topic. The **x-axis** represents the word scores, which indicate how strongly each word is associated with the topic. The **y-axis** lists the top words, such as **"user," "data," "information,"** and others. This quantitative representation complements the word cloud by providing a measurable perspective on keyword importance.



The comparison between Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) highlights their effectiveness as follows:

Balance of Topics:

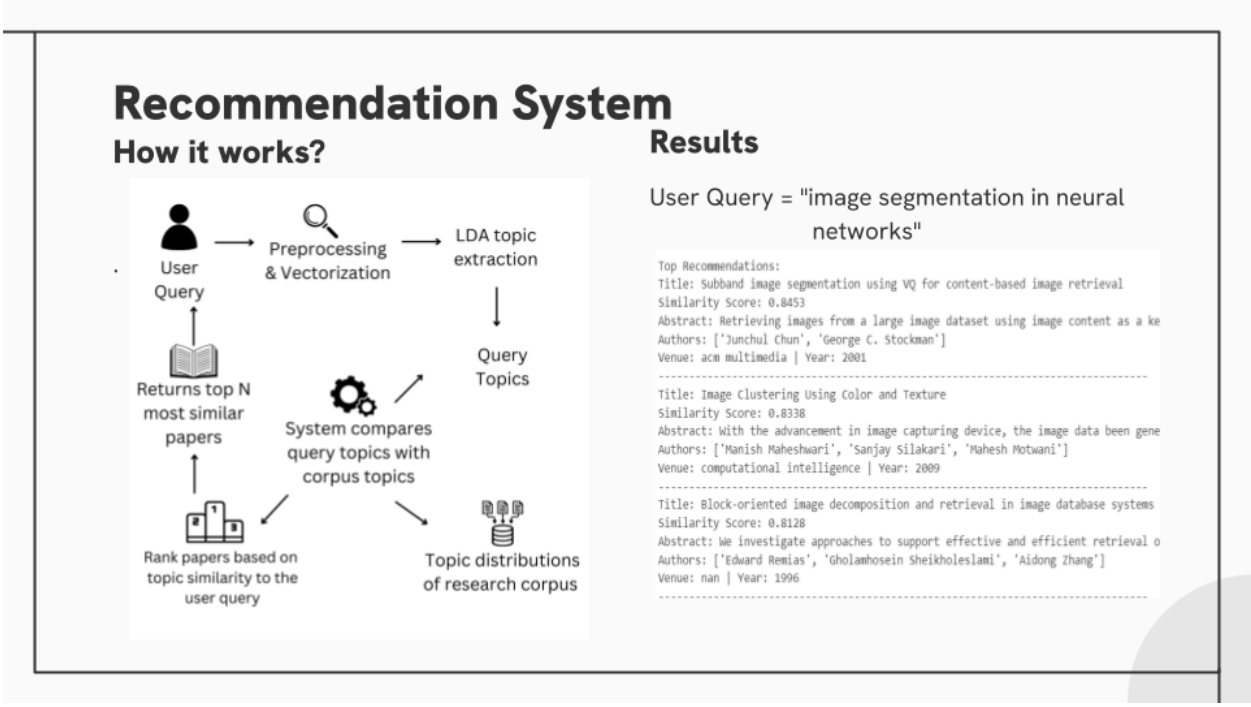
- **LDA**: Demonstrates a uniform distribution of topics, with documents ranging in size from approximately 67,000 to 231,000, ensuring balanced topic representation.
- **LSA**: Displays a skewed distribution, with one dominant topic (~778,000 documents) and other significantly smaller topics (~7,800 documents), resulting in an imbalance.

Word Representations:

- **LDA**: Provides clear and expressive relationships between words and topics. Words are associated with distinct themes and exhibit minimal overlap, improving interpretability and topic coherence.
- **LSA**: Produces non-specific associations of words, leading to overlapping themes and reduced clarity, making it harder to differentiate between topics.

Conclusion:

LDA offers superior topic balance and clearer topic representations, making it more effective for clustering and discriminating topics in text data. These attributes ensure better visualization and interpretability compared to LSA.

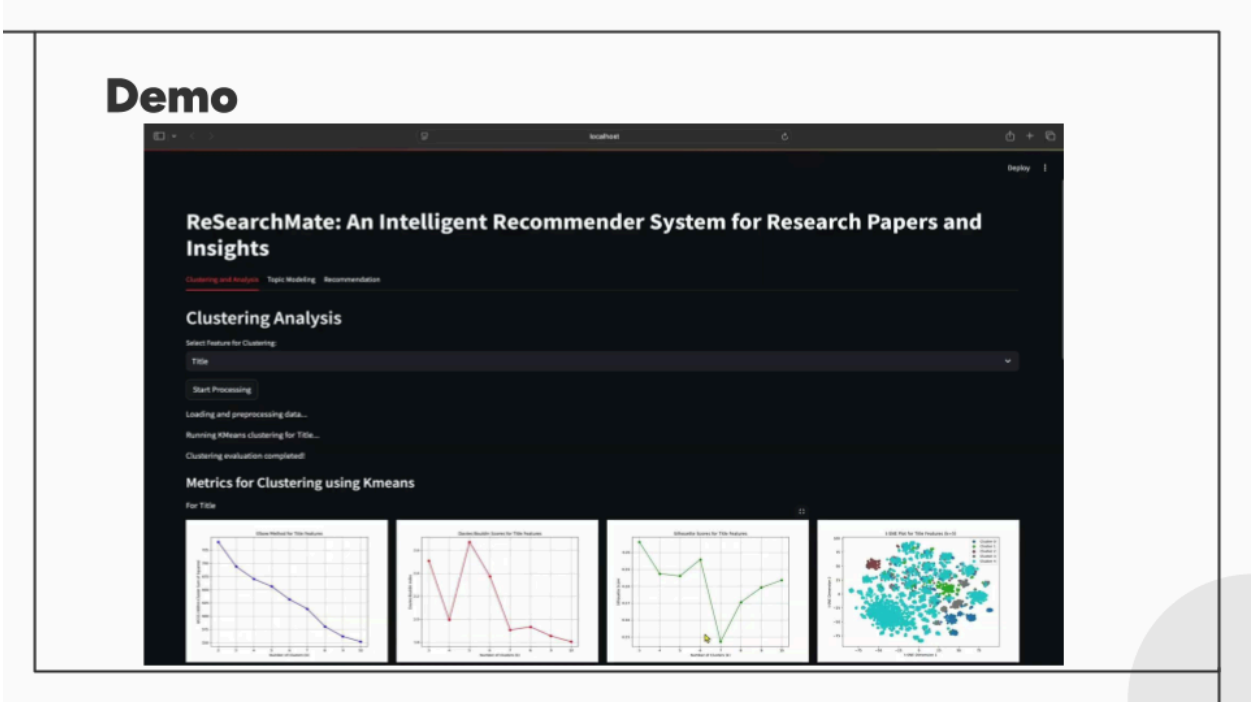


A recommendation system was built on top of the extracted LDA topics to suggest relevant research papers based on user input.

Process:

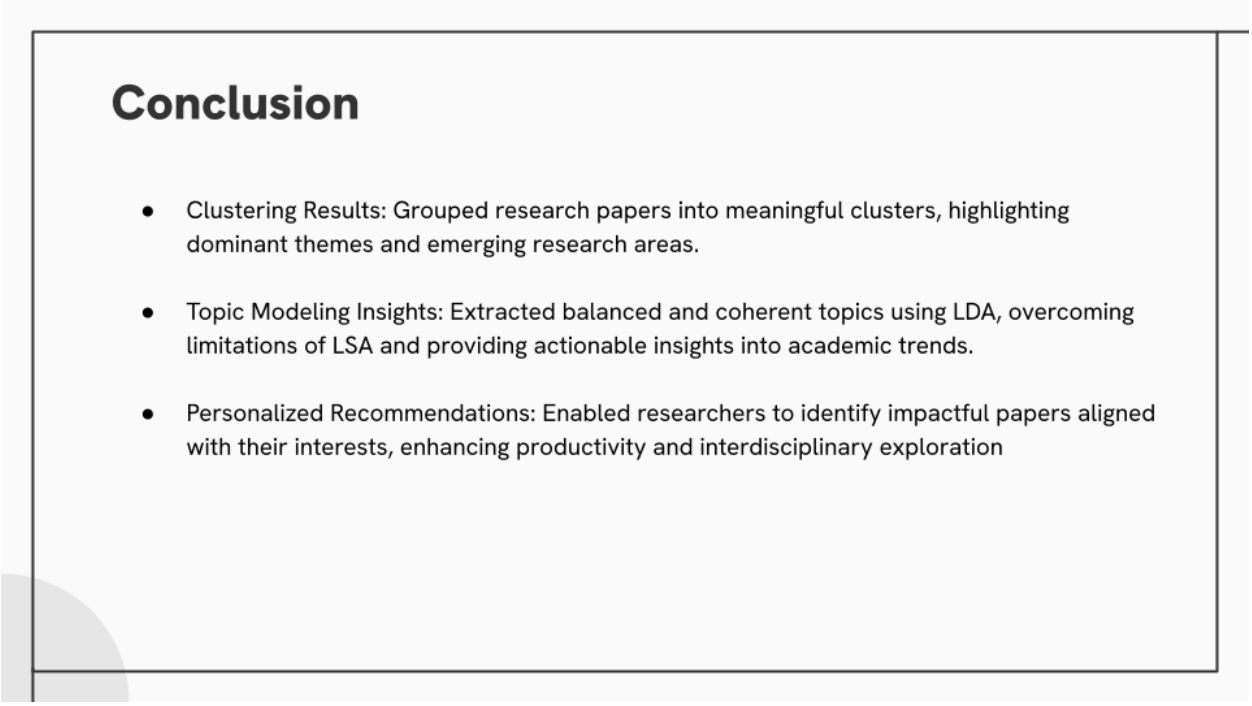
1. **User Query:** Users input a query, such as keywords or a paper title.
2. **Vectorization and Topic Extraction**: The query is preprocessed, vectorized using TF-IDF, and mapped into the LDA topic space.
3. **Similarity Search:** The system compares the user's topic vector to the topic vectors of the research papers using a FAISS index which uses cosine similarity for fast similarity computation.
4. **Recommendation Generation**: Papers with the highest similarity scores are ranked and returned to the user.
5. **Outputs**: Recommendations include paper titles, abstracts, authors, and similarity scores.

This approach leverages LDA's ability to capture diverse and nuanced research topics, potentially improving upon the LSA implementation by providing more accurate and contextually relevant recommendations.



The demo is the project in action. It shows the results from Clustering Analysis, Topic Modeling and personalized recommendation system. Here is the link to the demo:

Link to demo **(at Minute 8:47)**



To summarize the project report:

Clustering Results:
The clustering process successfully grouped research papers into distinct, meaningful clusters based on thematic similarities. These clusters highlighted dominant themes in the dataset, such as advancements in algorithms, sensor networks, or image processing. The use of techniques like t-SNE for dimensionality reduction and K-Means for clustering ensured that the papers were visually and semantically well-organized.

Topic Modeling Results:
By transitioning from Latent Semantic Analysis (LSA) to Latent Dirichlet Allocation (LDA), the topic modeling process overcame key challenges, improving the representation of overlapping and nuanced themes while avoiding the dominance of generic topics often seen with LSA. LDA's probabilistic approach resulted in higher coherence scores and more balanced topics, offering a clearer understanding of the corpus.

Personalized Recommendations:
The recommendation system, built on top of the topic modeling outputs, offered personalized paper suggestions to researchers. By analyzing the topics most relevant to a user's query or interests, the system identified papers that were closely aligned with their research focus. This personalized approach not only saves time but also empowers researchers to explore impactful papers across disciplines.

Overall, the system enhanced productivity by tailoring recommendations to the user's academic needs and providing seamless access to relevant research.