Project Milestone Report

Project Overview: This project investigates a dataset of 1 million research papers spanning various academic fields, such as computer science, physics, and mathematics. By leveraging structured metadata and abstracts, the project aims to uncover thematic clusters, interdisciplinary trends, and impactful research themes through advanced data exploration and analysis. This milestone report outlines the progress made, including successes, challenges, new ideas, and the refined scope for the final report.

Ideas Considered from Abstract in this Phase:

- **DBScan** for Clustering: DBScan was applied for clustering the dataset; however, a significant challenge was the high number of noise points detected. This resulted in a large proportion of the data being considered as outliers, which affected the overall clustering process.
- **KMeans** for Clustering: KMeans was utilized to segment the dataset into clusters. The optimal number of clusters, derived using the elbow method, was found to be 5. This provided a stable and interpretable clustering solution. The clusters seemed to correspond well to distinct groups in the dataset, offering useful segmentation.

New Ideas Not Mentioned in the Abstract:

LSA for Topic Modeling: Initially, Latent Semantic Analysis (LSA) was explored for topic modeling. The primary concern with LSA was that it tended to lean towards a dominant cluster in the dataset, limiting the diversity of the topics discovered. This issue was noted during the phase when analyzing the top words associated with the topics.

After analyzing the behavior of different topic modeling techniques, it became evident that LSA could be further optimized. The dataset revealed that LSA was underperforming in capturing diverse topics.

How the Data Led to New Ideas:

Upon review of initial clustering results, it was clear that while KMeans clustering provided good results in terms of segmentation, LSA's topic modeling approach was dominated by a few major clusters, limiting the depth of analysis. The prevalence of this issue led to a more focused exploration of LSA and its impact on the quality of topic modeling. A refinement of the LSA model, including the adjustment of the number of topics emerged as an essential next step.

Proposed Scope of the Work for the Final Report:

The final report will focus on building a Recommendation Engine based on Latent Dirichlet Allocation (LDA) as a more robust alternative to LSA. While LSA has shown promise, LDA's probabilistic nature and ability to handle more complex topic structures make it a better fit for recommendations. The proposed scope includes:

1. **Topic Modeling Using LDA:** Moving from LSA to LDA to better capture the structure of topics in the dataset.
2. **Recommendation System**: Developing a recommendation engine that leverages topic distributions derived from LDA to suggest relevant academic papers based on the topics they cover.
3. **Visualization**: Continuing the visualization of topics over time and their relationships to other variables like publication year, author affiliations, and citations to provide deeper insights.

Progress So Far:

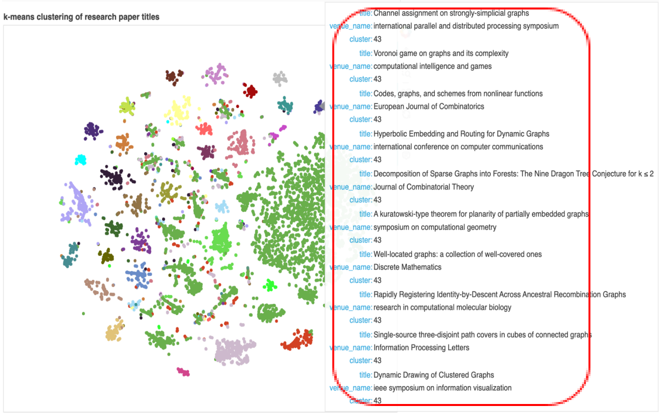
- LSA Implementation**:** The abstracts were cleaned and preprocessed using standard NLP techniques such as tokenization, stopword removal, and lemmatization.
- A TF-IDF matrix was constructed and passed through LSA to extract 5 topics.
- Reduce high-dimensional data using t-SNE to project it into a 2D / 3D space, and then perform K-Means clustering on the reduced data to identify distinct groups or patterns.
- The dominant topic was assigned to each paper based on the highest score in the topic matrix.
- Topic-word visualizations and heatmaps were created to illustrate the relationships between topics and their top words.

Challenges:

- LSA’s Dominance of a Single Cluster: The LSA approach tended top dominate one particular cluster, limiting the diversity of topics.
- DBScan's Noise: High noise levels in DBScan made it challenging to extract meaningful clusters, leading to a pivot towards KMeans for clustering analysis.
- KMeans Performance: KMeans clustering with 5 clusters provided an optimal solution, but the lack of semantic depth in the clustering led to the exploration of more advanced techniques like LDA for recommendation systems.
- Dataset size leads to memory allocation size errors leading to temporary subset experiments.

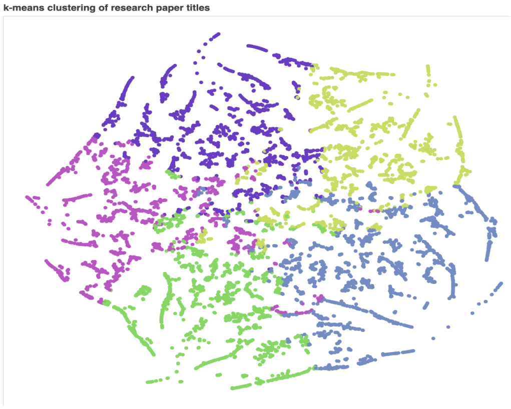
The next phase will involve implementing LDA and leveraging its benefits over LSA to build a more accurate recommendation engine.

Key Visualizations

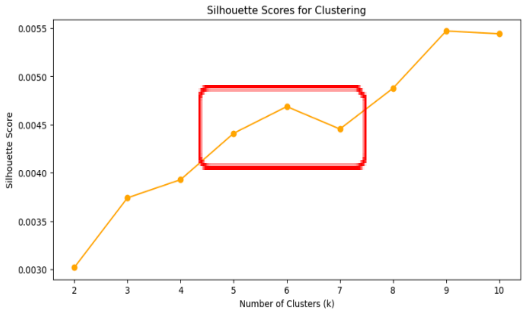


```
     title: Channel assignment on strongly-simplicial graphs
venue_name: international parallel and distributed processing symposium
   cluster: 43
     title: Voronoi game on graphs and its complexity
venue_name: computational intelligence and games
   cluster: 43
     title: Codes, graphs, and schemes from nonlinear functions
venue_name: European Journal of Combinatorics
   cluster: 43
     title: Hyperbolic Embedding and Routing for Dynamic Graphs
venue_name: international conference on computer communications
   cluster: 43
     title: Decomposition of Sparse Graphs into Forests: The Nine Dragon Tree Conjecture for k ≤ 2
venue_name: Journal of Combinatorial Theory
   cluster: 43
     title: A kuratowski-type theorem for planarity of partially embedded graphs
venue_name: symposium on computational geometry
   cluster: 43
     title: Well-located graphs: a collection of well-covered ones
venue_name: Discrete Mathematics
   cluster: 43
     title: Rapidly Registering Identity-by-Descent Across Ancestral Recombination Graphs
venue_name: research in computational molecular biology
   cluster: 43
     title: Single-source three-disjoint path covers in cubes of connected graphs
venue_name: Information Processing Letters
   cluster: 43
     title: Dynamic Drawing of Clustered Graphs
venue_name: ieee symposium on information visualization
   cluster: 43
```

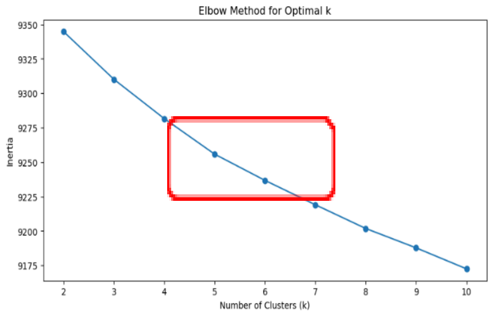
[Figure 1: K-Means with 50 clusters (without tuning optimal cluster size)]



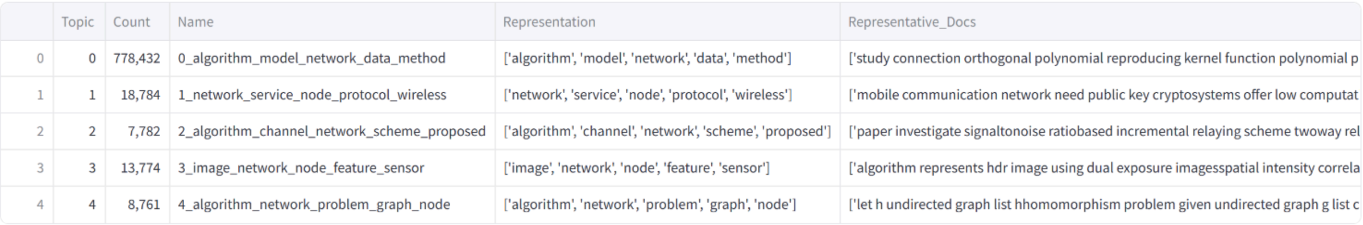
[Figure 2: K-Means with 5 clusters]



[Figure 3: Silhouette Scores for clustering]

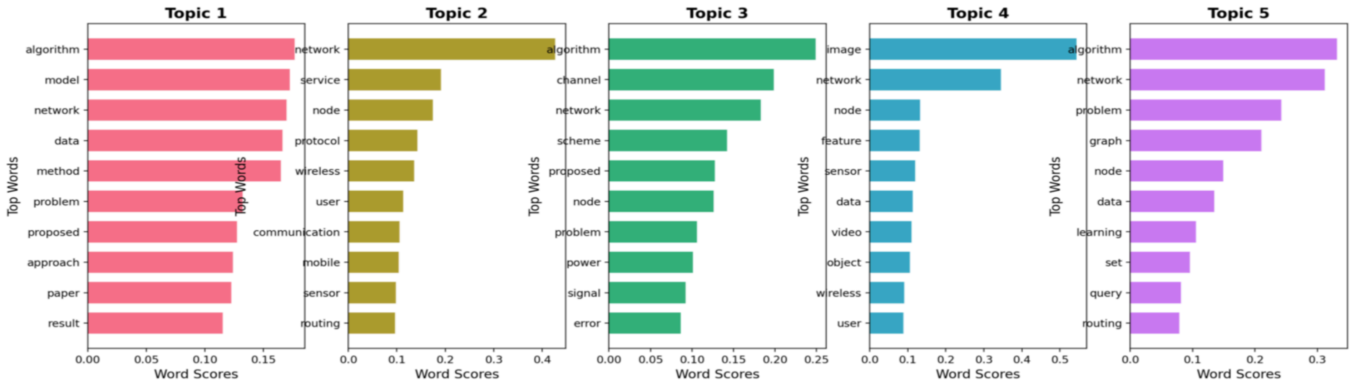


[Figure 4:Elbow Method For Optimal K]



	Topic	Count	Name	Representation	Representative_Docs
0	0	778,432	0_algorithm_model_network_data_method	['algorithm', 'model', 'network', 'data', 'method']	['study connection orthogonal polynomial reproducing kernel function polynomial p
1	1	18,784	1_network_service_node_protocol_wireless	['network', 'service', 'node', 'protocol', 'wireless']	['mobile communication network need public key cryptosystems offer low computat
2	2	7,782	2_algorithm_channel_network_scheme_proposed	['algorithm', 'channel', 'network', 'scheme', 'proposed']	['paper investigate signaltonoise ratiobased incremental relaying scheme twoway rel
3	3	13,774	3_image_network_node_feature_sensor	['image', 'network', 'node', 'feature', 'sensor']	['algorithm represents hdr image using dual exposure imagesspatial intensity correla
4	4	8,761	4_algorithm_network_problem_graph_node	['algorithm', 'network', 'problem', 'graph', 'node']	['let h undirected graph list hhomomorphism problem given undirected graph g list c

[Figure 5 & 6: Topics Summary Table with Word Scores]


Topic 1
algorithm
model
network
data
method
problem
proposed
approach
paper
result
Top Words
Word Scores
Topic 2
network
service
node
protocol
wireless
user
communication
mobile
sensor
routing
Top Words
Word Scores
Topic 3
algorithm
channel
network
scheme
proposed
node
problem
power
signal
error
Top Words
Word Scores
Topic 4
image
network
node
feature
sensor
data
video
object
wireless
user
Top Words
Word Scores
Topic 5
algorithm
network
problem
graph
node
data
learning
set
query
routing
Top Words
Word Scores