TABLE OF CONTENTS

1.SUBTOPIC CLUSTERING AND ANALYSIS USING TF-IDF AND K-MEANS CLUSTERING

A comprehensive analysis of subtopics is performed using advanced text mining and machine learning techniques. Initially, subtopics are transformed into numerical vectors through TF-IDF (Term Frequency-Inverse Document Frequency) vectorization, which quantifies the importance of words within the corpus. This vectorized data is then subjected to K-Means clustering, where subtopics are grouped into a specified number of clusters based on their content similarity. To facilitate visualization, the high-dimensional TF-IDF vectors are reduced to two principal components using PCA (Principal Component Analysis). The resulting clusters are visualized in a scatter plot, where each subtopic is color-coded according to its assigned cluster. Finally, the subtopics within each cluster are listed, providing a clear understanding of the thematic groupings. This process effectively organizes the subtopics into meaningful clusters, aiding in the analysis and interpretation of thematic content within the data.

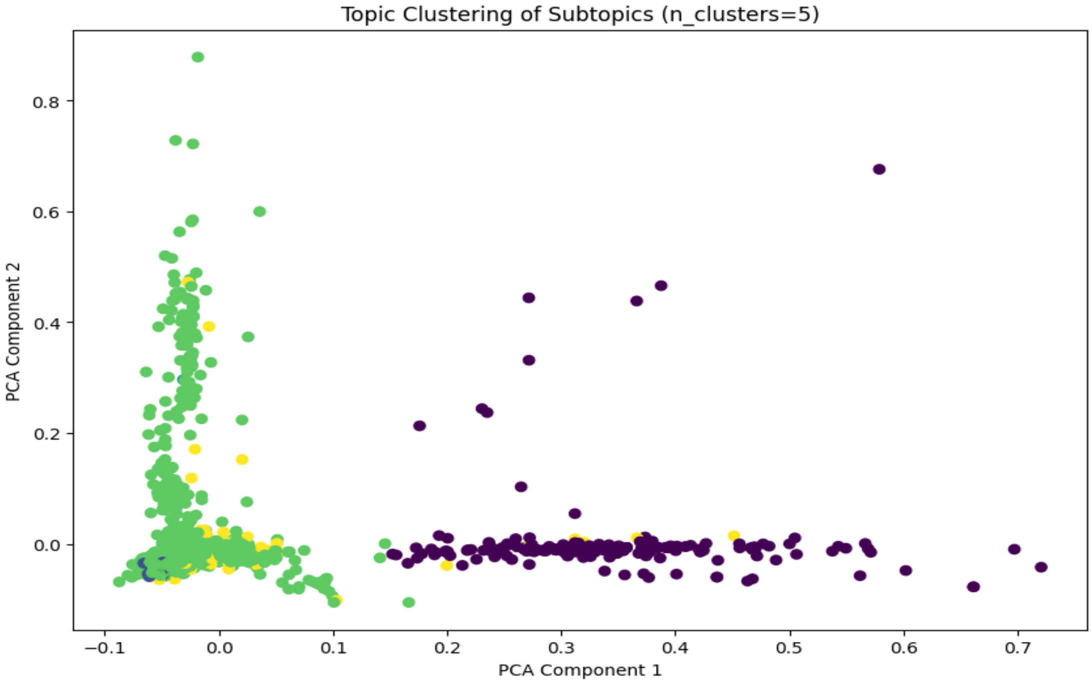
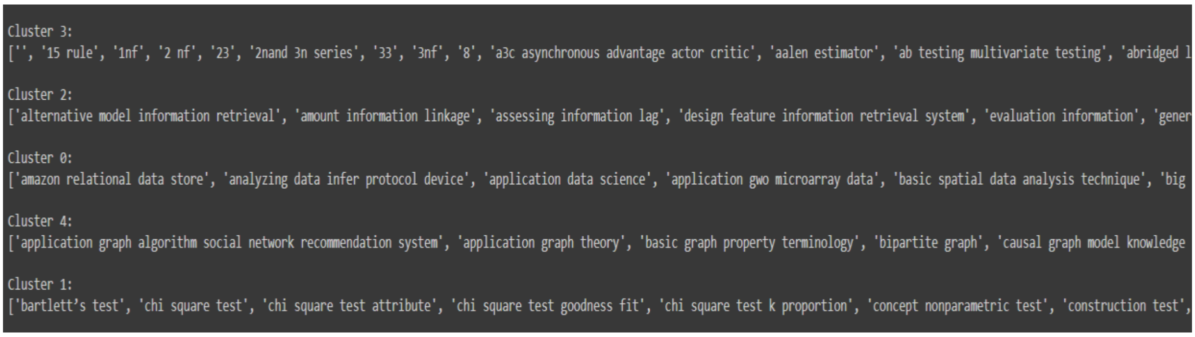


Fig 1.a ***Scatter Plot of Subtopic Clustering Using K-Means (n_clusters=5) and PCA Component Analysis***

The scatter plot visualizes the clustering of subtopics into five distinct groups, as determined by the K-Means clustering algorithm. Each point represents a subtopic, and its color indicates the cluster to which it has been assigned. The two axes represent the first two principal components derived from a PCA, which reduce the dimensionality of the TF-IDF features to facilitate visualization.

The plot shows that subtopics are distributed into clusters, with some clusters displaying tighter grouping (less spread), indicating a higher degree of similarity among the subtopics within those clusters. In contrast, other clusters are more spread out, suggesting more diversity among the subtopics. The variability in the spread and overlap between clusters suggests that while some subtopics are distinctly grouped, others may share characteristics with multiple clusters, reflecting the complexity and potential overlap in topics.



```
Cluster 3:
['', '15 rule', '1nf', '2 nf', '23', '2nand 3n series', '33', '3nf', '8', 'a3c asynchronous advantage actor critic', 'aalen estimator', 'ab testing multivariate testing', 'abridged l

Cluster 2:
['alternative model information retrieval', 'amount information linkage', 'assessing information lag', 'design feature information retrieval system', 'evaluation information', 'gener

Cluster 0:
['amazon relational data store', 'analyzing data infer protocol device', 'application data science', 'application gwo microarray data', 'basic spatial data analysis technique', 'big

Cluster 4:
['application graph algorithm social network recommendation system', 'application graph theory', 'basic graph property terminology', 'bipartite graph', 'causal graph model knowledge

Cluster 1:
['bartlett’s test', 'chi square test', 'chi square test attribute', 'chi square test goodness fit', 'chi square test k proportion', 'concept nonparametric test', 'construction test',
```

Fig 1.b Clusters of subtopics (10-15 showed in image per cluster)

2.VISUALISING AND ANALYZING SUBUNIT SIMILARITIES USING NODE2VEC AND T-SNE

1. **Embedding Computation with Node2Vec**:

   I applied the Node2Vec algorithm to embed the nodes (subunits) into a continuous vector space. Node2Vec captures the structure of the graph by learning embeddings that reflect node similarities and relationships. The result was an embedding matrix where each row represented a node in the graph, encoded in a high-dimensional vector space.

2. **Dimensionality Reduction and Visualization**:

   To visualize the embeddings, I employed t-Distributed Stochastic Neighbor Embedding (t-SNE), a dimensionality reduction technique. t-SNE projects the high-dimensional embeddings into a two-dimensional space, making it easier to observe the clustering and distribution of nodes.

3. **Color Coding for Visualization**:

   In the t-SNE visualization, nodes were color-coded based on their program affiliation:

   a. **Blue** for 'MDS' subunits
   b. **Red** for 'MST' subunits
   c. **Green** for 'MDA' subunits

4. This color-coding allowed us to visually differentiate and analyze how subunits from various programs were distributed and clustered in the embedding space.

   Cluster Analysis:

   The t-SNE plot revealed distinct clusters for each program, indicating that subunits within the same program were more closely related in the embedding space compared to those from other programs. These clusters provide insights into how similar subunits are grouped together and highlight potential overlaps between different programs.

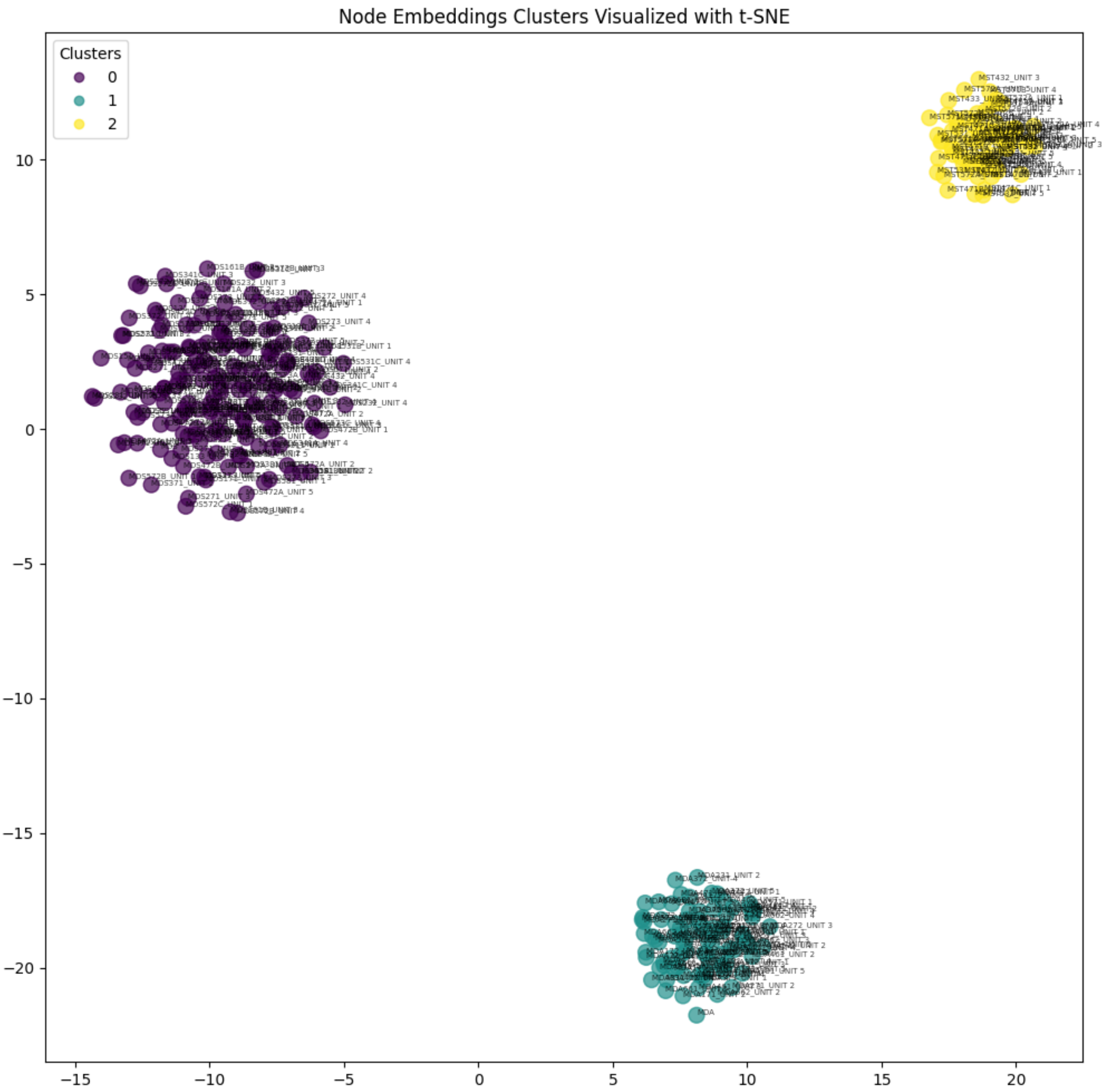


Fig 2.a Node Embedding clusters using t-SNE

Insights and Interpretation:

The analysis demonstrated that subunits from the same program tend to form cohesive clusters, while subunits from different programs are more dispersed. This clustering confirms the effectiveness of the Node2Vec embeddings in capturing the program-specific characteristics and relationships of subunits.

3.KEYWORD FREQUENCY ANALYSIS

In this analysis, a systematic approach was employed to identify and visualize the most frequently occurring keywords across the subtopics within the course structure. Initially, the subtopics were compiled into a single text corpus, where each word was extracted and counted to determine its frequency. Utilizing the `Counter` class from Python's `collections` module, the analysis quantified the occurrence of each word, allowing for the identification of the top 20 most frequent keywords. These keywords were then visualized through a horizontal bar plot, generated using the `seaborn` library, with a focus on clarity and visual appeal. The resulting plot, styled with the 'viridis' color palette, effectively highlights the predominant themes and concepts emphasized within the course material, providing valuable insights into the core topics covered. This methodical approach ensures that the analysis is both comprehensive and visually informative, making it a key component of the overall course content review.

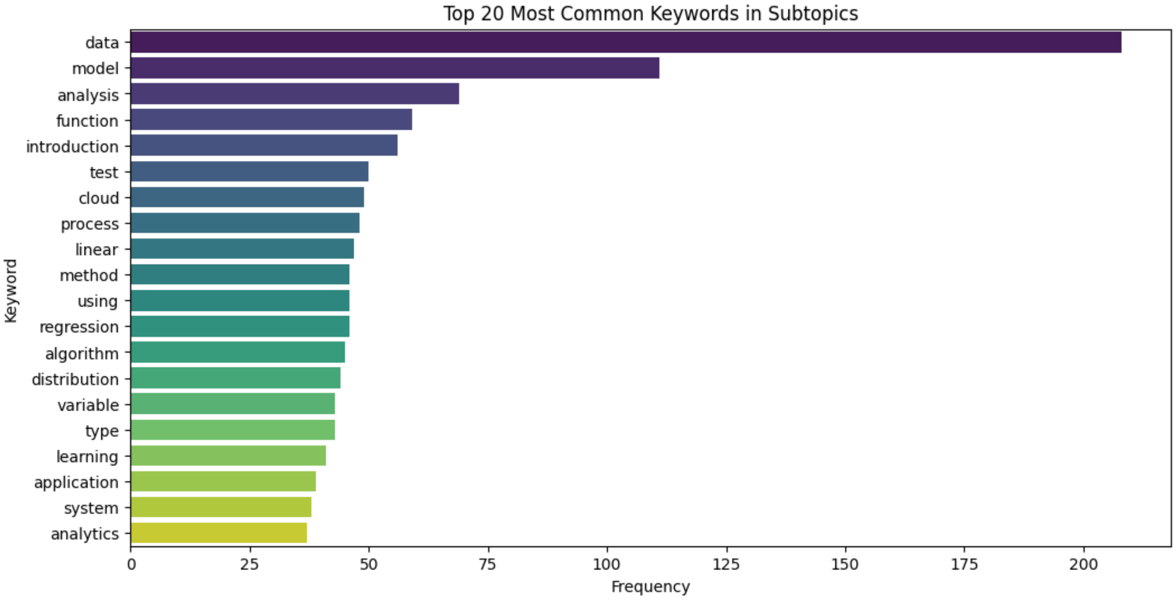


Fig 3.a Top 20 most common keywords used in subtopics

Interpretation: "Data" emerges as the most dominant keyword, appearing nearly 200 times, followed by "Model," which occurs over 150 times. Keywords such as "Analysis," "Function," and "Introduction" also feature prominently, each with frequencies ranging from 50 to 100. This distribution indicates a strong emphasis on topics related to data analysis and modeling, suggesting a focus on fields like data science, machine learning, and statistics.

4.REFERENCES

[1] Y. J. Kim, J. Y. Lee, and C. H. Lee, "Cosine Similarity and Term Frequency–Inverse Document Frequency for Text Analysis," *Journal of Physics: Conference Series*, vol. 154, no. 1, p. 012012, 2020.

[2] A. Grover and J. Leskovec, "Node2Vec: Scalable Feature Learning for Networks," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855-864, 2016.

[3] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.