Clustering documents and visualization of Embedding Vector Space

Presented By:

- Aditya Shidhaye (1566213)
- Pradeep Patwa (1564727)
- Adithya Ramesh (1567434)

Examined by:
Prof. Dr. Damir Dobric
Prof. Dr. Andreas Pech

Content

- Introduction
- Problem Statement
- Methodology
- Implementation
- Results
- Conclusion

Introduction

- Document clustering is an essential technique for organizing large volumes of textual data.
- Traditional clustering (TF-IDF, BoW) approaches often struggle with capturing semantic meaning.
- This project leverages OpenAI's text embeddings for efficient and meaningful document grouping.

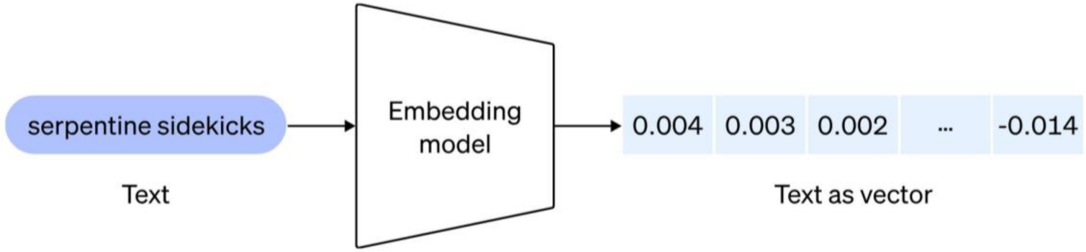
OpenAI

Problem Statement

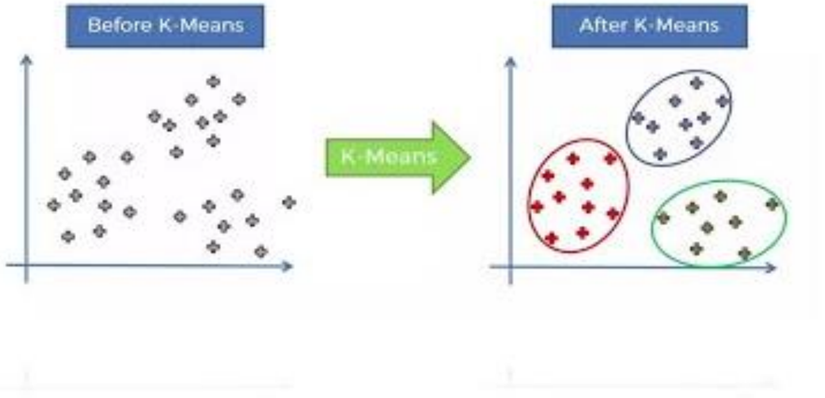
- Large datasets contain unstructured textual data that is difficult to categorize manually.
- Keyword-based approaches fail to capture semantic relationships between documents.
- The goal is to create an AI-powered clustering model that groups similar documents efficiently.

Methodology

Text Embedding Process



- OpenAI's **text-embedding-3-large**
- Captures semantic relationships between words and phrases.
  Eg. Apple(fruit)/Apple(company), not feeling good / sick
- Allows for more accurate clustering compared to traditional term frequency methods.



K Means Clustering

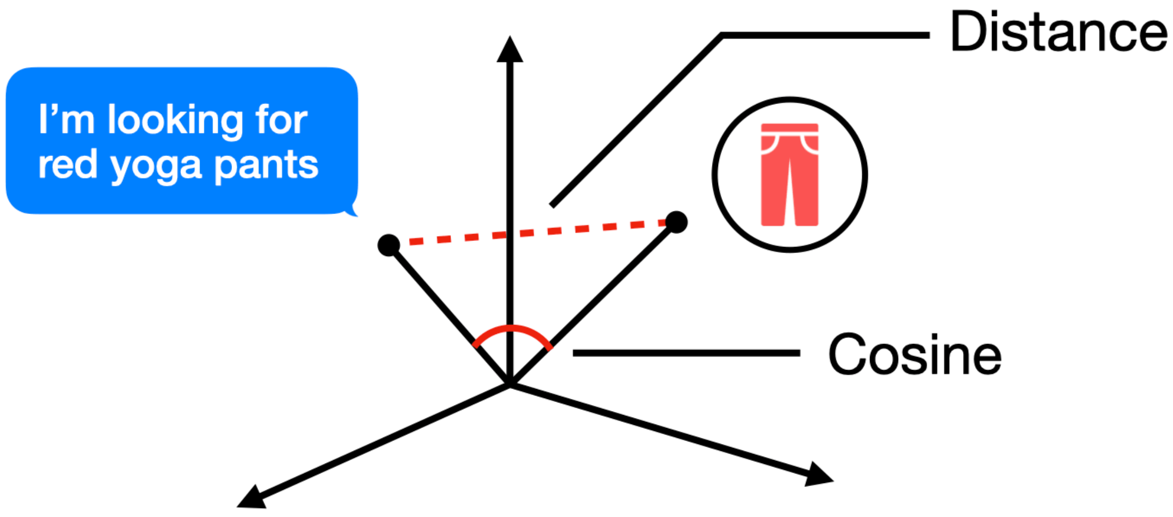
Clustering Algorithm

- **K-Means Clustering:** An unsupervised learning algorithm that partitions data into K clusters.
- **Objective:** Minimize intra-cluster distance and maximize inter-cluster separation.
- The embeddings are clustered based on their numerical similarities.

Cosine Similarity Analysis

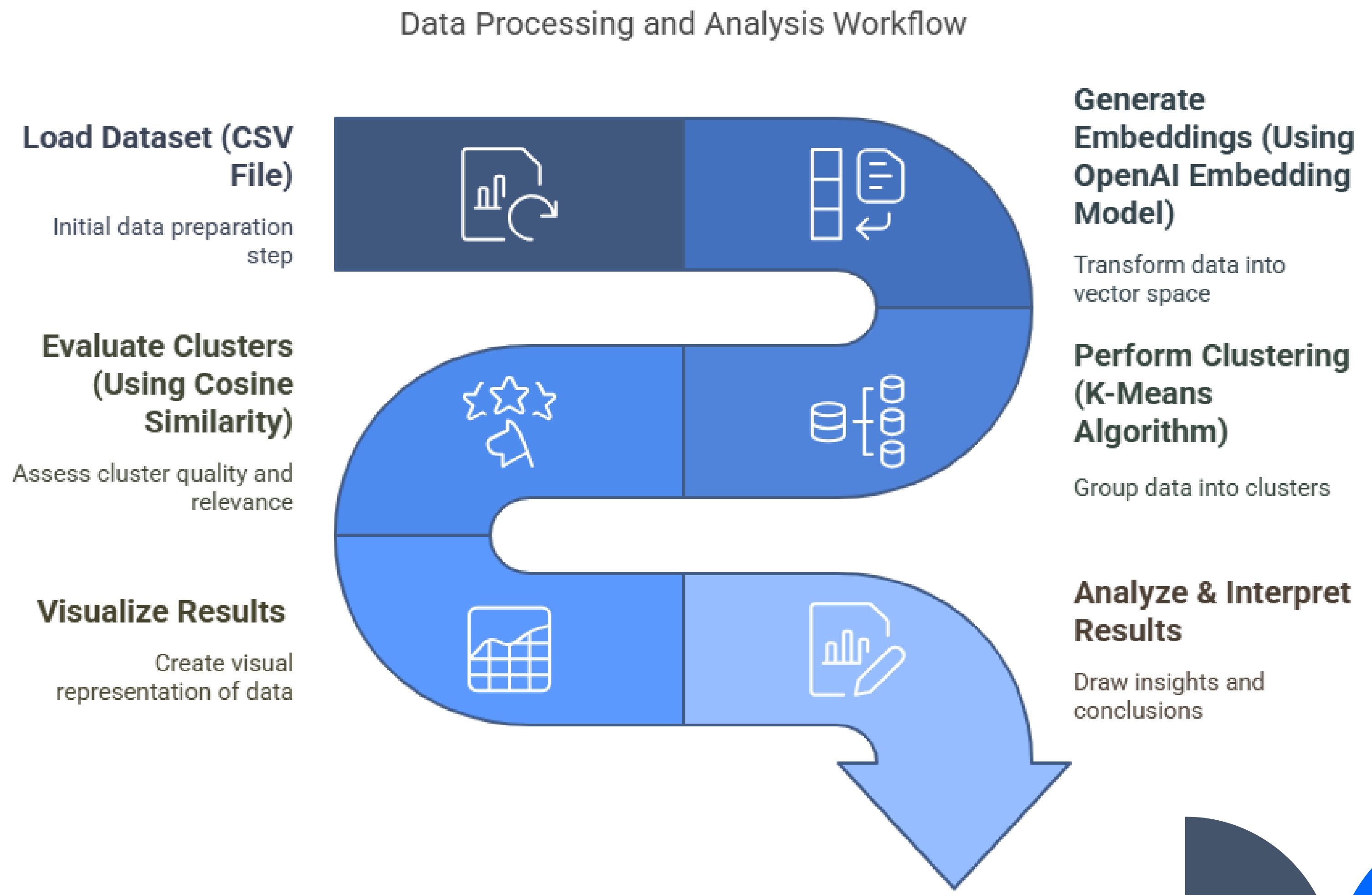
- measure simililarity within the same cluster.
- **Intra-cluster similarity**: Higher similarity within a cluster indicates well-defined groups.
  Similarity = 0.95 → "Artificial Intelligence" and "Machine Learning" (very related)
- **Inter-cluster similarity**: Lower similarity between different clusters ensures better separation.
  Similarity = 0.10 → "Artificial Intelligence" and "Cooking Recipes" (almost unrelated)

Cosine similarity is a metric used to determine how similar two vectors are



Closer angle = more similar | Wider angle = less similar

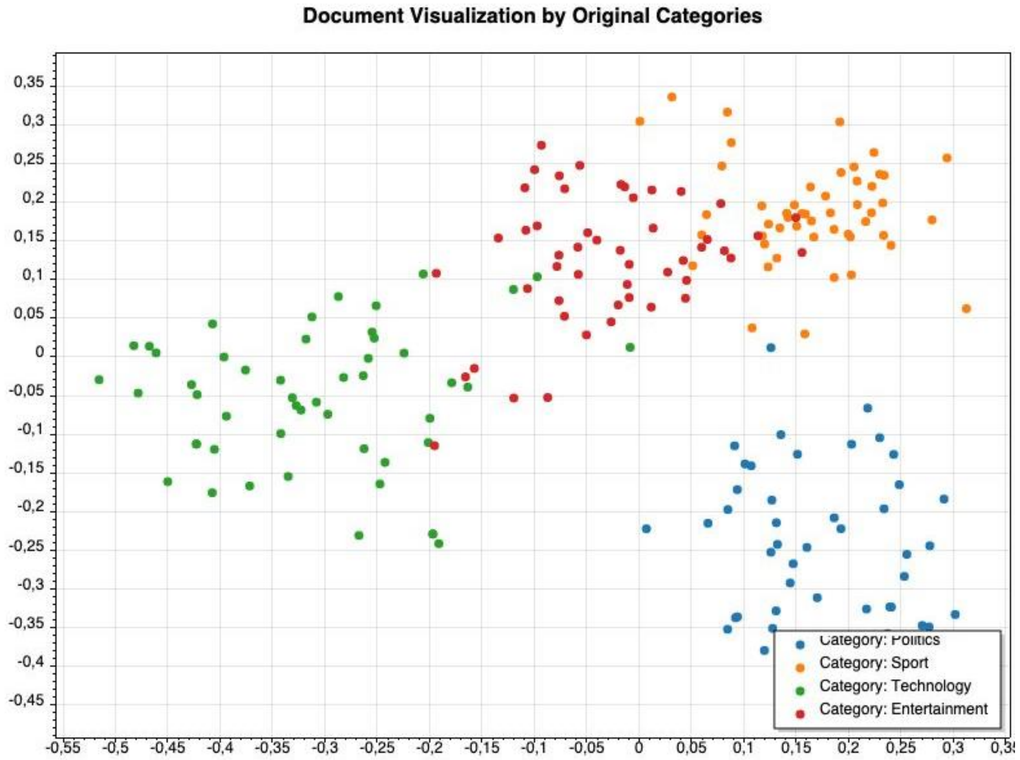
Implementation



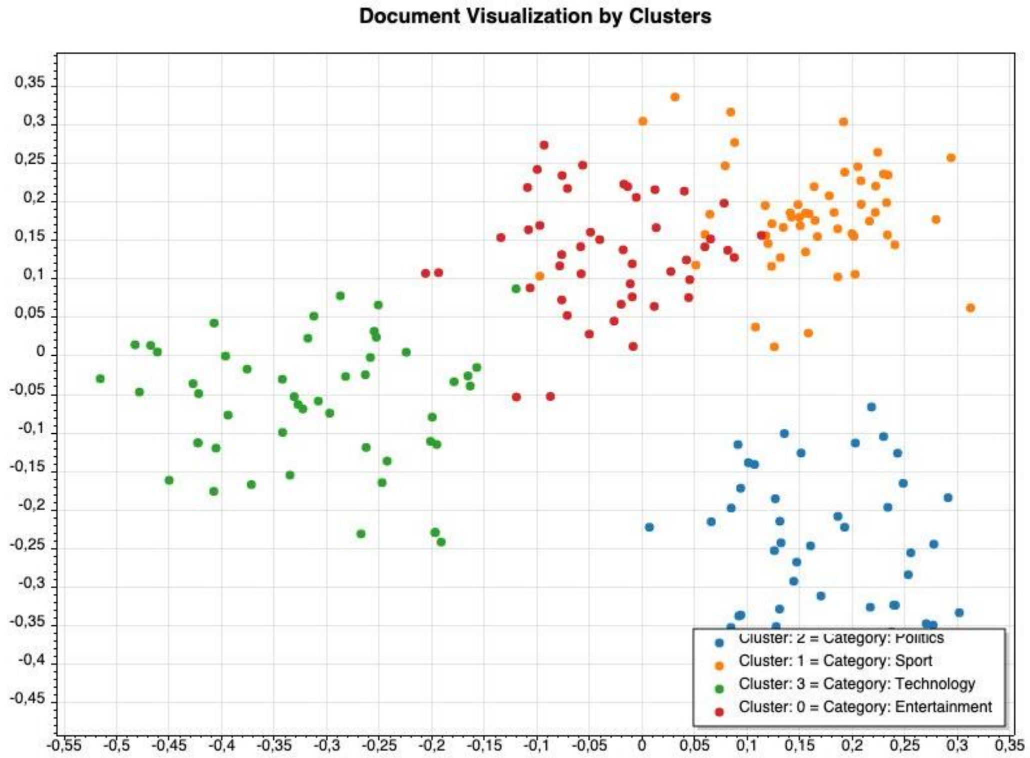
Results

- Clusters were successfully formed based on the semantic meaning of documents.
- Cosine similarity confirmed well-defined groups with minimal overlap.
- Visualization showed clear separation between clusters.
- Challenges: Optimal selection of K-value and processing time for large datasets.

Outputs

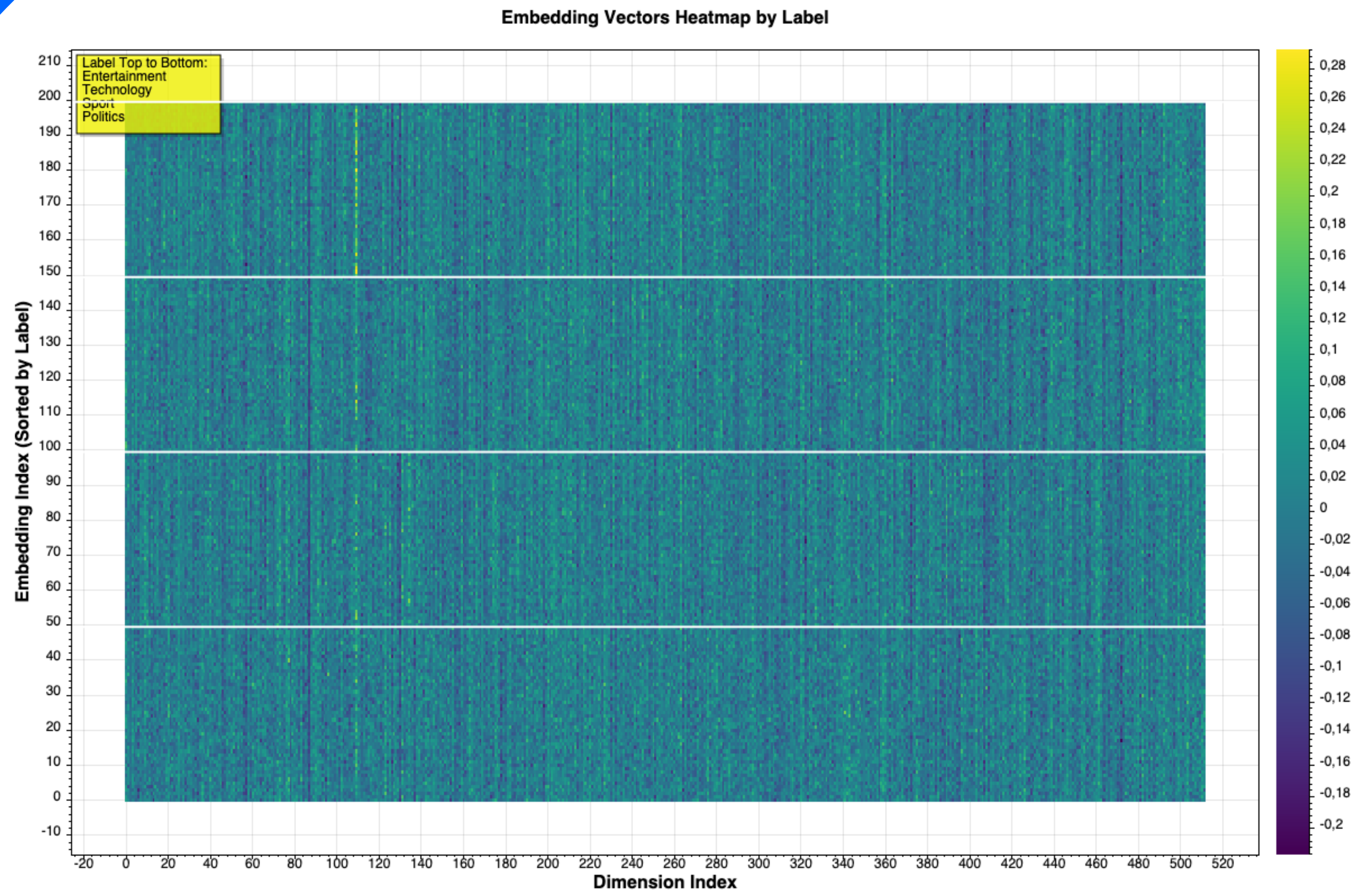


K - means algorithm



Cosine Similarity values

```
=== Document Clustering Evaluation ===

Overall Metrics:
Average Intra-Cluster Similarity: 0,2940
Average Inter-Cluster Similarity: 0,1868
Average Category Similarity: 0,2961

Silhouette Coefficient: 0,3647 (higher is better)

Cluster to Category Mapping:
Cluster 2 -> Category '0' (Purity: 100,00 %)
Cluster 0 -> Category '1' (Purity: 94,34 %)
Cluster 1 -> Category '2' (Purity: 97,78 %)
Cluster 3 -> Category '3' (Purity: 88,68 %)

Cluster Details:
Cluster 0 Intra-Similarity: 0,3535
Cluster 1 Intra-Similarity: 0,2761
Cluster 2 Intra-Similarity: 0,3143
Cluster 3 Intra-Similarity: 0,2465

Category Details:
Category 0 Intra-Similarity: 0,3483
Category 1 Intra-Similarity: 0,2811
Category 2 Intra-Similarity: 0,3025
Category 3 Intra-Similarity: 0,2526
```



Heatmap

representation of the embeddings based on the category they belong.

Conclusion

- AI-driven document clustering provides a more efficient and scalable approach.
- OpenAI embeddings enhance clustering accuracy by capturing semantic meaning.
- Future improvements can lead to even more refined document classification models.
- Trade-off between accuracy and efficiency

Thank you !