P7: Document Clustering, Summarization and Visualization

CSE 573 Spring 23 Semantic Web Mining - Group 12

Team Members:

- Akashkiran Shivakumar (1222183248)
- Jayavardhan Karampudi (1222872339)
- Prateek Pandey (1224105467)
- Raviram Mamidi (122307268)
- Sundaravadivel CP (1222352703)
- Tejesh Andhavarapu (1225589664)

Problem Definition

- With the evolution of the internet, many documents are available online and it has been difficult to find out and extract important information.
- Large-scale text summarization is difficult and time-consuming. Extensive text processing and calculations are required.
- Document clustering is grouping a set of documents based on a similarity score. Integrated with any search engine, clustering allows us to see the overall structure of the document set and browse as deep into it as you want.
- Document summarization saves a lot of time and helps in gaining a subjective understanding of the articles.
- The main goal of the project is to

    1. Cluster the articles and provide a short summary
    2. Apply visualization techniques to showcase relevancy
    3. Document summarization

Algorithms & Techniques

- **Clustering:** Latent Dirichlet Allocation(LDA), Hierarchical Density Based Spatial Clustering(HDBScan), Agglomerative Clustering
- **Latent Dirichlet Allocation (LDA):** A probabilistic generative model used for topic modeling that assigns topic distributions to documents and word distributions to topics.
- **Hierarchical Density Based Spatial Clustering (HDBScan)**: A density-based clustering algorithm that can discover clusters of varying shapes and sizes in a dataset and also identify noise and outliers.
- **Agglomerative Clustering:** A bottom-up hierarchical clustering algorithm that starts with each data point as its own cluster and iteratively merges clusters based on a distance metric until a stopping criterion is met.

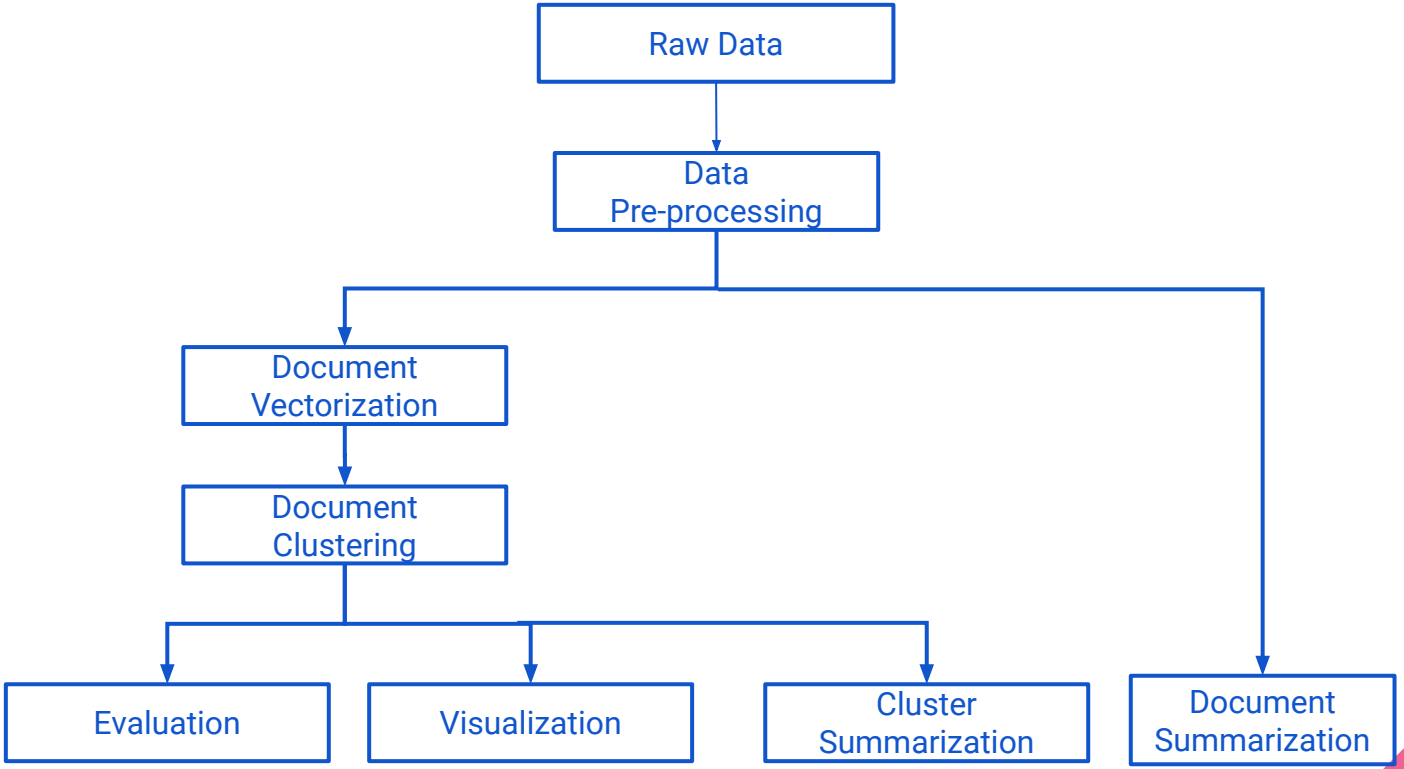
Algorithms & Techniques Contd.

Visualization: Uniform Manifold Approximation and Projection (UMAP), t-Distributed Stochastic Neighbor Embedding (t-SNE), Compression Variational Autoencoder (CVAE)

Summarization:

- Extractive text summarization using Spacy & Word frequencies
- Abstractive text summarization using Facebook BART Large CNN

System Architecture



Data Set

- The 20 Newsgroups dataset is a collection of 20,000 documents from 20 different newsgroups.
- The documents are evenly distributed among the newsgroups, meaning that each newsgroup has an equal number of documents.
- The dataset is available for download at http://qwone.com/~jason/20Newsgroups/.

Data Preprocessing

- We are taking all the subset of fetch_20newsgroups and removed headers, footers etc
- Then we toned it down to text and label (18846, 2)
- Converting the text to lowercase and tokenizing the sentences
- Removing whitespaces, punctuation and stop words and normalizing the sentence
- Tokens to digits and lemmatization
- Data is then converted to vector form and removed null char to preprocessed data
- This data of shape (18846, 2) is then used for our clustering algorithms

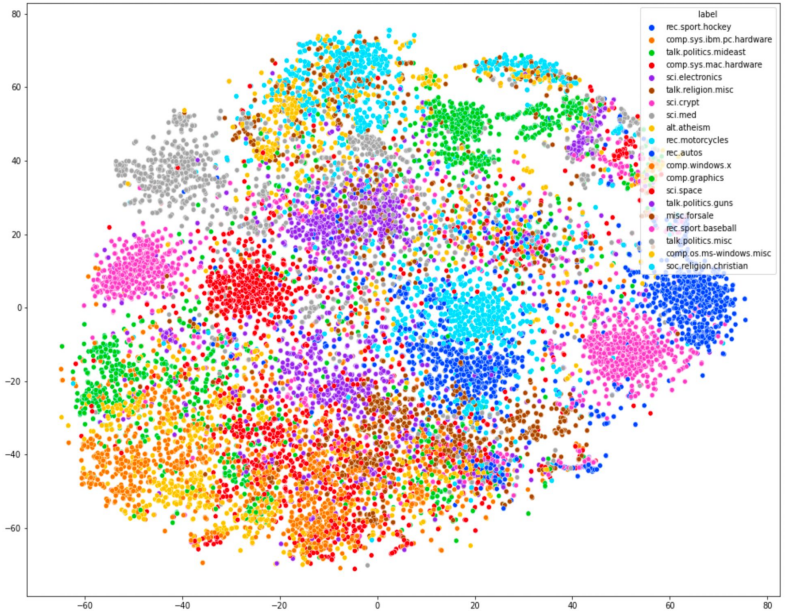
Evaluation metrics

- **Homogeneity**: Measures how much each cluster contains only samples from a single class.
- **Completeness**: Measures how much all samples from a given class are assigned to the same cluster.
- **V-measure:** Computes the harmonic mean between Homogeneity and Completeness, giving equal importance to both measures.
- **Adjusted Rand-Index:** Measures the similarity between the true labels and the predicted labels, taking into account chance agreement.
- **Silhouette Coefficient:** Measures how similar an object is to its own cluster compared to other clusters, ranging from -1 to 1.

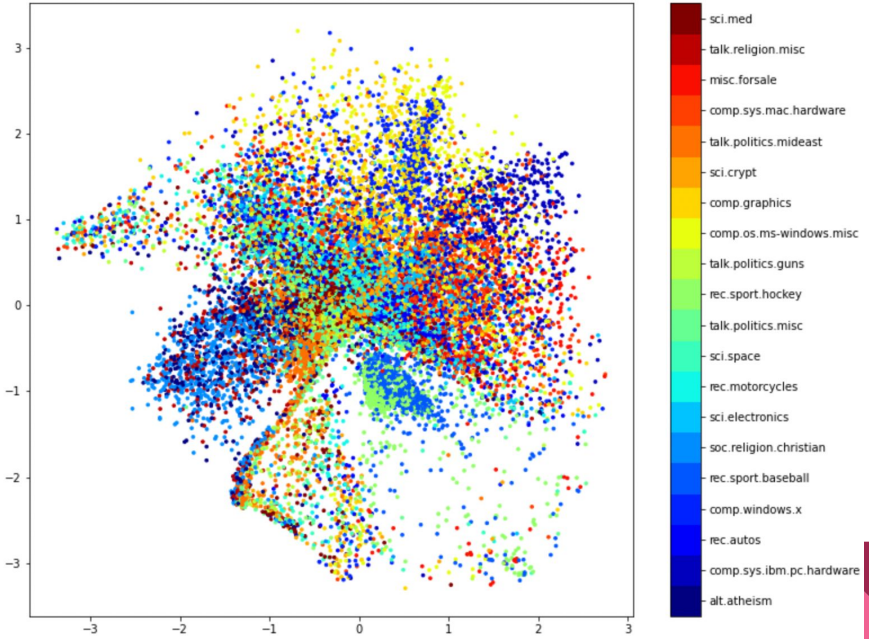
Evaluation

Clustering Technique	Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Silhouette Coefficient
LDA	0.583	0.584	0.584	0.491	0.014
HDBScan	0.317	0.493	0.385	0.132	0.343
Agglomerative Clustering	0.379	0.396	0.387	0.206	0.004

Visualization Results for LDA Clustering

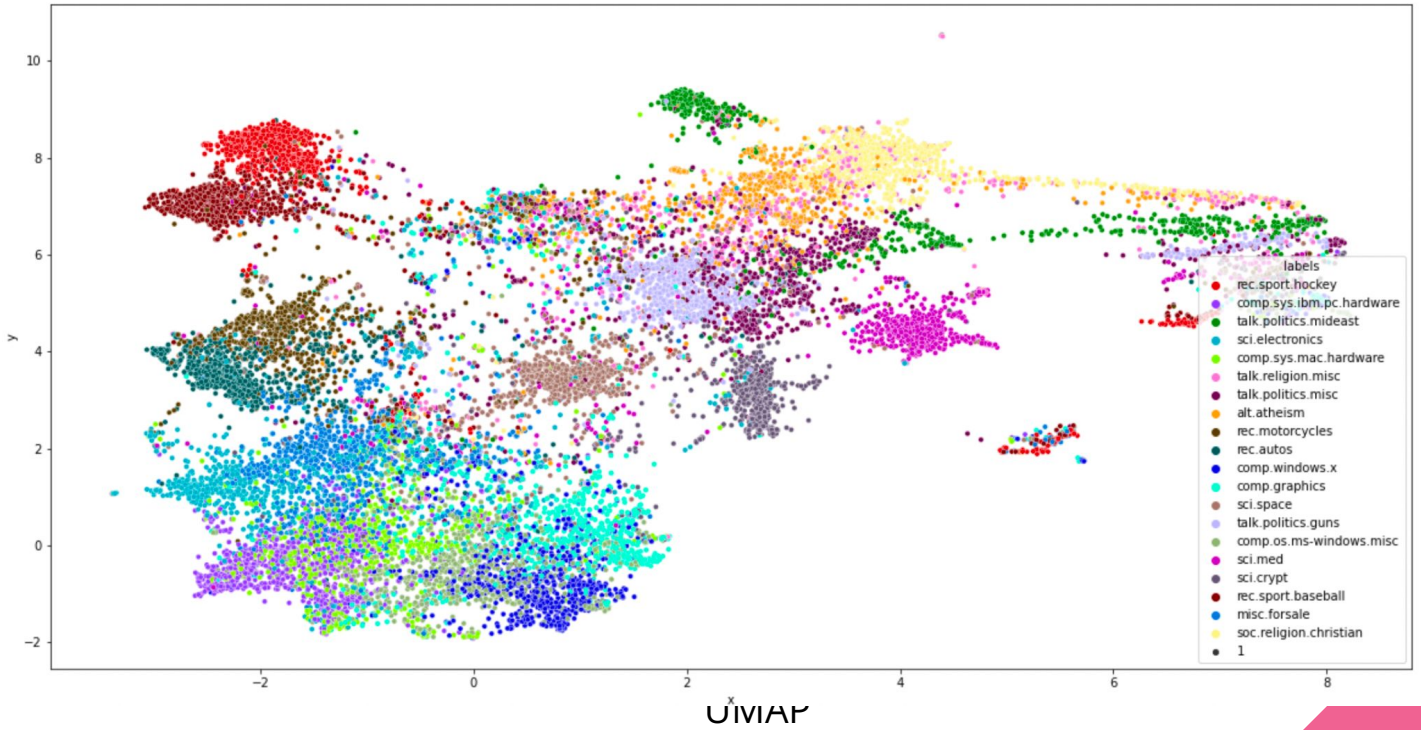


t-SNE



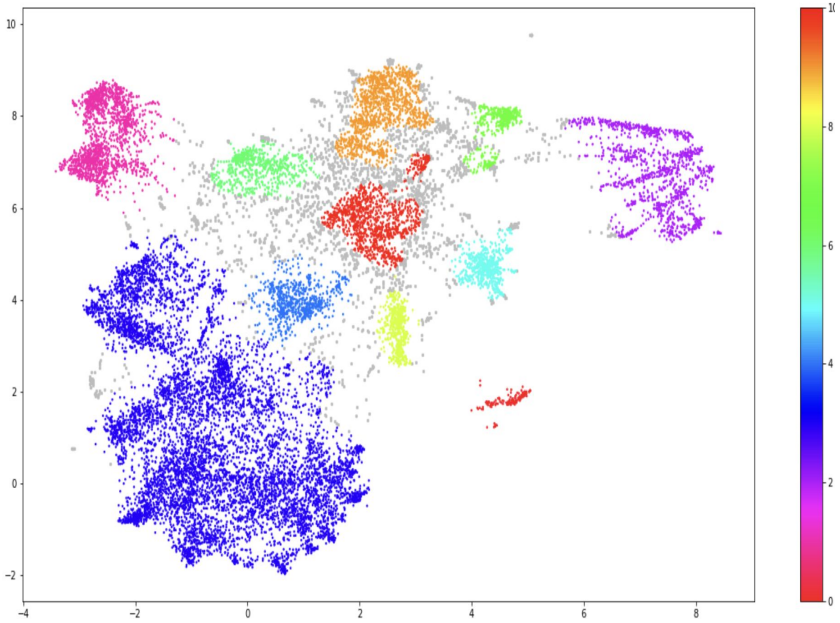
CVAE

Visualization Results for LDA Clustering

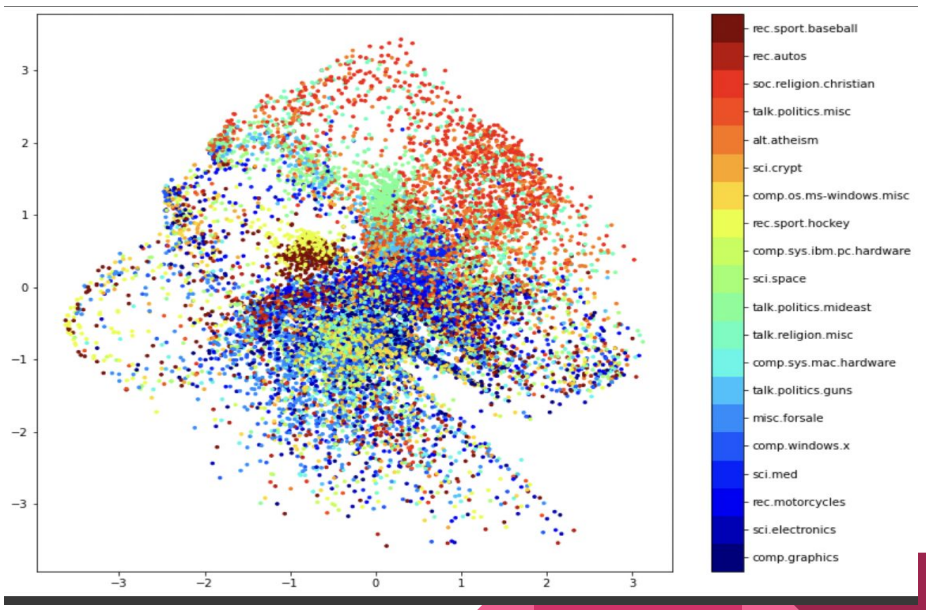


UMAP

Visualization results for HDBScan

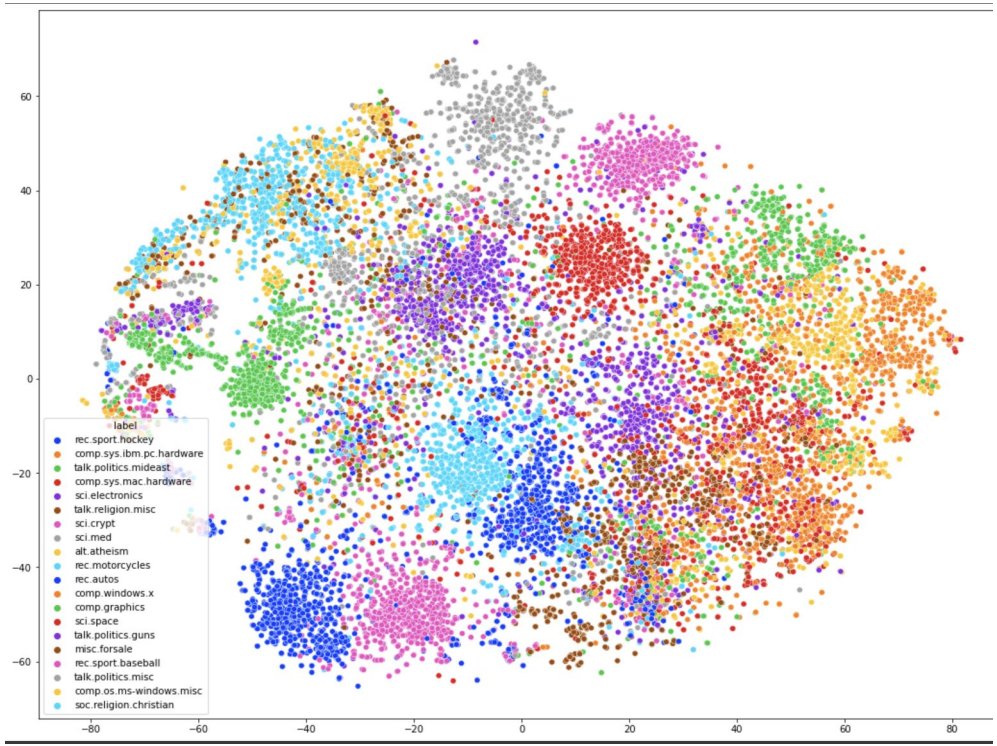


UMAP



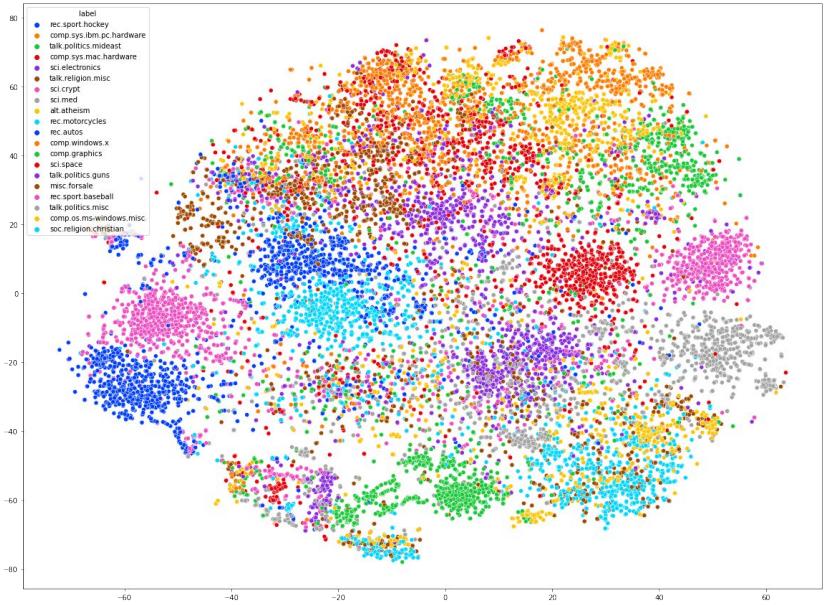
CVAE

Visualization results for HDBScan

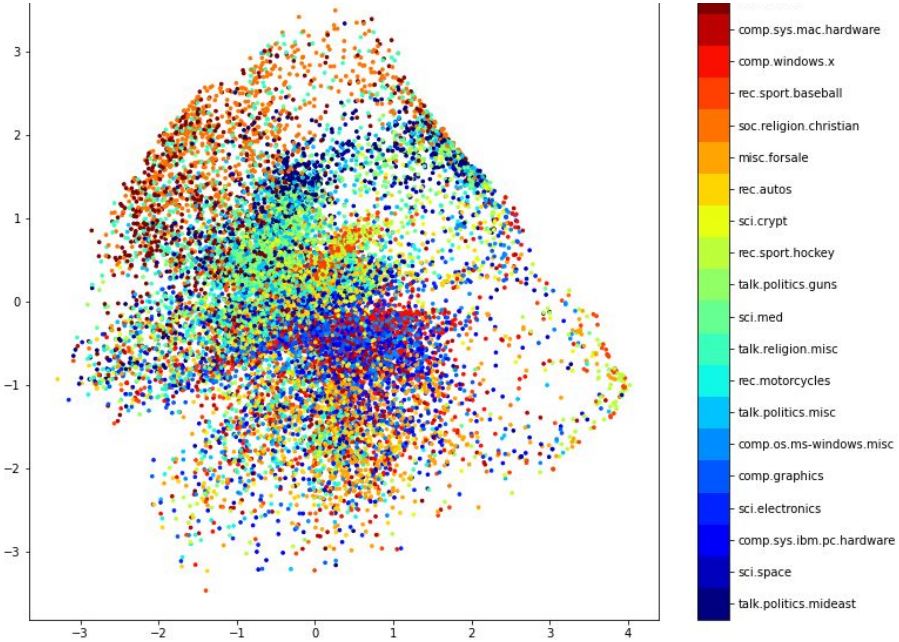


t-SNE

Visualization Results for Agglomerative Clustering

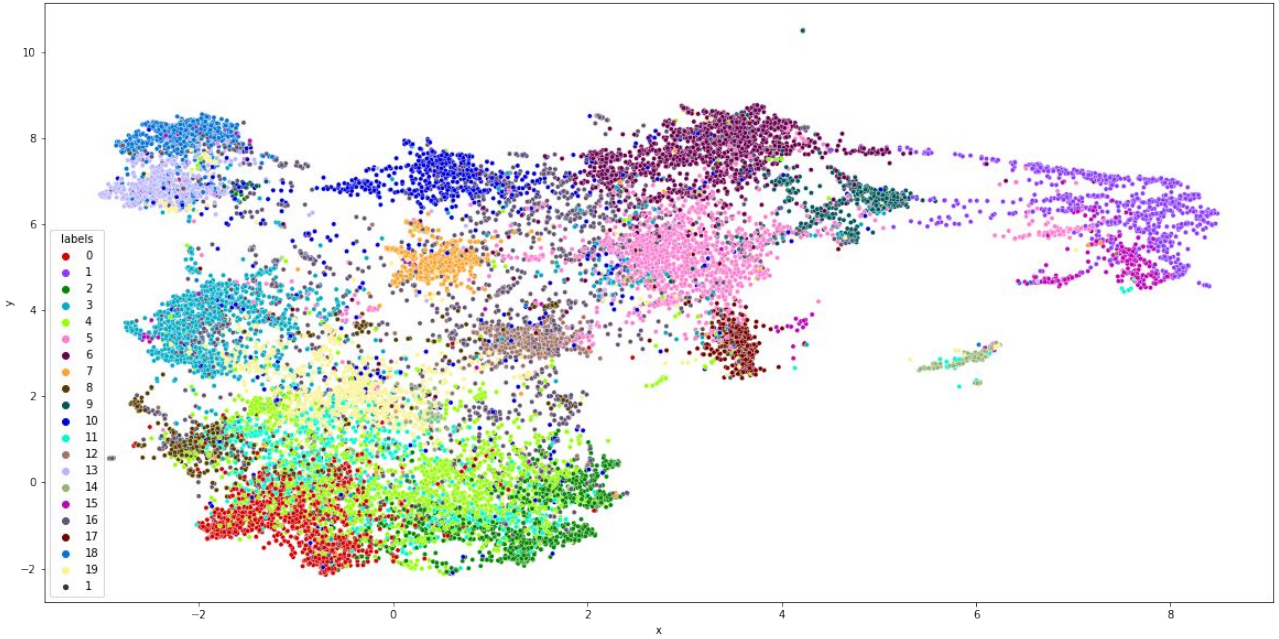


t-SNE



CVAE

Visualization Results for Agglomerative Clustering



UMAP

Project Timeline

Task	Description	Team Members	Deadline
Study of clustering and visualization techniques	Perform research on different clustering and visualization techniques to apply on datasets.	All team members	Jan 31 - Feb 16
Data Pre-processing	Pre-process the data to remove noise and convert it to process for data embedding.	Sundar,Tejesh, Prateek	Feb 17 - Feb 28
Data Embedding	Perform sentence embedding to represent the data in vector form.	Akashkiran, Jayavardhan, Raviram	Feb 28 - March 15
Clustering	Implement LDA, HDBScan, Agglomerative clustering	All team members	Mar 15 - Mar 30
Document summarization	Individual Documents & Cluster Documents	All team members	April 1 - April 5
Visualization	t-SNE,UMAP,Compression VAE	Prateek,Raviram, Jayavardhan	April 5 - April 10
Summary, Final Evaluation and Analysis	Evaluate and analyze the implemented clustering techniques. Documentation of methods, evaluation techniques and results	All team members	April 10– April 15

References

- Giri. (2021, May 2). Is Latent Dirichlet Allocation (LDA) A clustering algorithm? HDS; High Demand Skills. https://highdemandskills.com/lda-clustering/
- http://qwone.com/ jason/20Newsgroups/
- Millar, Jeremy R. et al. “Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps.” FLAIRS Conference (2009)
- Cao,Tuan-Dungetal.“Hot Topic Detection on Newspaper”Conference: the Ninth International Symposium (2018)
- Karmakar, Saurav. “Syntactic and Semantic Analysis and Visualization of Unstructured English Texts.” (2011)
- https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection
- https://albertauyeung.github.io/2020/06/19/bert-tokenization.html/

Code

- Link: https://github.com/jayavardhan3112/SWM573_Document_Clustering_Summarization_and_Visualization