**Document Clustering and Topic Modelling**

**Introduction**

Document clustering and topic modelling are essential techniques in the field of Natural Language Processing (NLP) and Machine Learning. These methods help in organizing, summarizing, and understanding large text corpora by identifying patterns, themes, and underlying structures within the data. This report presents an approach to apply clustering algorithms like Latent Dirichlet Allocation (LDA) and K-means to group similar documents together, providing a detailed methodology, implementation, and visualization.

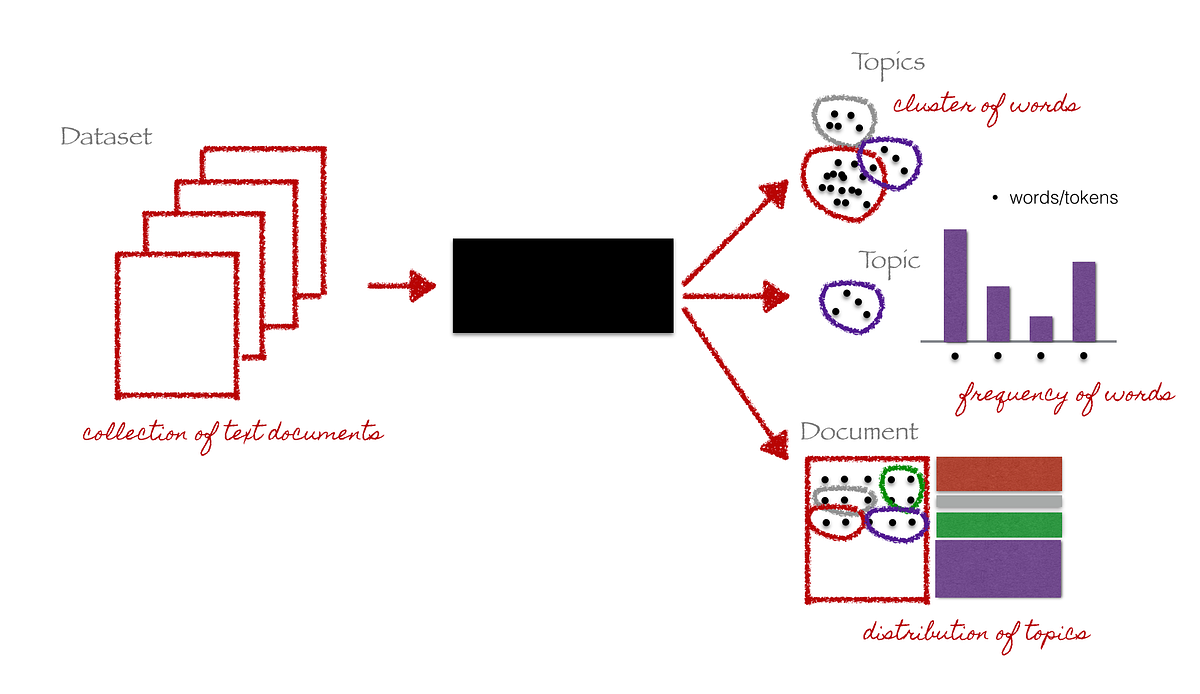
**Aim**

The aim of this project is to:

1. Apply document clustering and topic modelling techniques to a collection of text documents.
2. Use Latent Dirichlet Allocation (LDA) to identify and understand the latent topics within the documents.
3. Employ K-means clustering to group documents into clusters based on their content.
4. Visualize the results using t-SNE for dimensionality reduction to facilitate the understanding of the document clusters and topics.

**Methodology**

1. **Data Collection**:
   * A sample dataset of documents is created for this project. Each document represents a different topic or theme.
2. **Text Preprocessing**:
   * Convert the raw text documents into numerical representations using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. This step includes tokenization, removing stop words, and converting text to lowercase.
3. **Latent Dirichlet Allocation (LDA)**:
   * Apply LDA to the TF-IDF matrix to identify the underlying topics in the documents. LDA is a generative probabilistic model that assumes each document is a mixture of topics, and each topic is a mixture of words.
4. **K-means Clustering**:
   * Use the K-means algorithm to cluster the documents into a predefined number of clusters based on their TF-IDF representations. K-means is a centroid-based algorithm that partitions data into clusters by minimizing the variance within each cluster.
5. **Visualization using t-SNE**:
   * Apply t-SNE (t-Distributed Stochastic Neighbour Embedding) for dimensionality reduction to visualize the high-dimensional TF-IDF matrix in a two-dimensional space. This helps in visualizing the clustering results and the distribution of topics.
6. **Gensim LDA**:
   * Additionally, use Gensim's LDA implementation to confirm the topic modelling results and to provide an alternative approach.



**[Topic modelling using Latent Dirichlet Allocation (LDA)](https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045" \t "_blank)**

**Results**

1. **LDA Topic Modeling**:
   * The LDA model identified two distinct topics within the documents. Each topic is represented by a set of top words that characterize it.
2. **K-means Clustering**:
   * The K-means algorithm successfully clustered the documents into two groups based on their content.
3. **Visualization**:
   * The t-SNE plots provide a visual representation of the document clusters and the distribution of topics, making it easier to interpret the results.
4. **Gensim LDA**:
   * The Gensim LDA model confirmed the topics identified by the scikit-learn LDA implementation, providing additional validation of the results.

**Conclusion**

This project demonstrated the application of document clustering and topic modeling techniques to a small corpus of text documents. By using Latent Dirichlet Allocation (LDA) and K-means clustering, we were able to identify and visualize the underlying topics and group similar documents together. These techniques are valuable for organizing and understanding large text corpora, and the methodology can be extended to larger and more complex datasets.

The modular code provided allows for easy adaptation and scalability, making it a robust foundation for further exploration in the fields of NLP and machine learning.