NLP Assignment: Text Analysis and Topic Discovery

Introduction & Objective

This project demonstrates an end-to-end workflow for text analysis and topic discovery using Natural Language Processing (NLP). The main objective is to extract meaningful insights from text documents using various NLP techniques such as text preprocessing, TF-IDF analysis, Word2Vec embeddings, and Latent Dirichlet Allocation (LDA) topic modeling.

Methodology

The project is implemented in the following steps:

1. Text Preprocessing

- Lowercasing: Converts all text to lowercase.
- Punctuation & number removal: Removes special characters and digits.
- Tokenization: Splits sentences into words.
- Stopword removal: Eliminates common words like 'the', 'is', 'and'.

2. TF-IDF Analysis

TF-IDF (Term Frequency – Inverse Document Frequency) is applied to identify the most important words in each document. Using TfidfVectorizer, the top 10 words with highest TF-IDF scores are extracted per document, highlighting unique and meaningful terms.

3. Word2Vec Embeddings

A Word2Vec model is trained on the preprocessed corpus to learn semantic word representations. It allows us to find similar words and visualize embeddings in 2D using PCA. Words with similar meanings appear close together in the plot.

4. Topic Modeling (LDA)

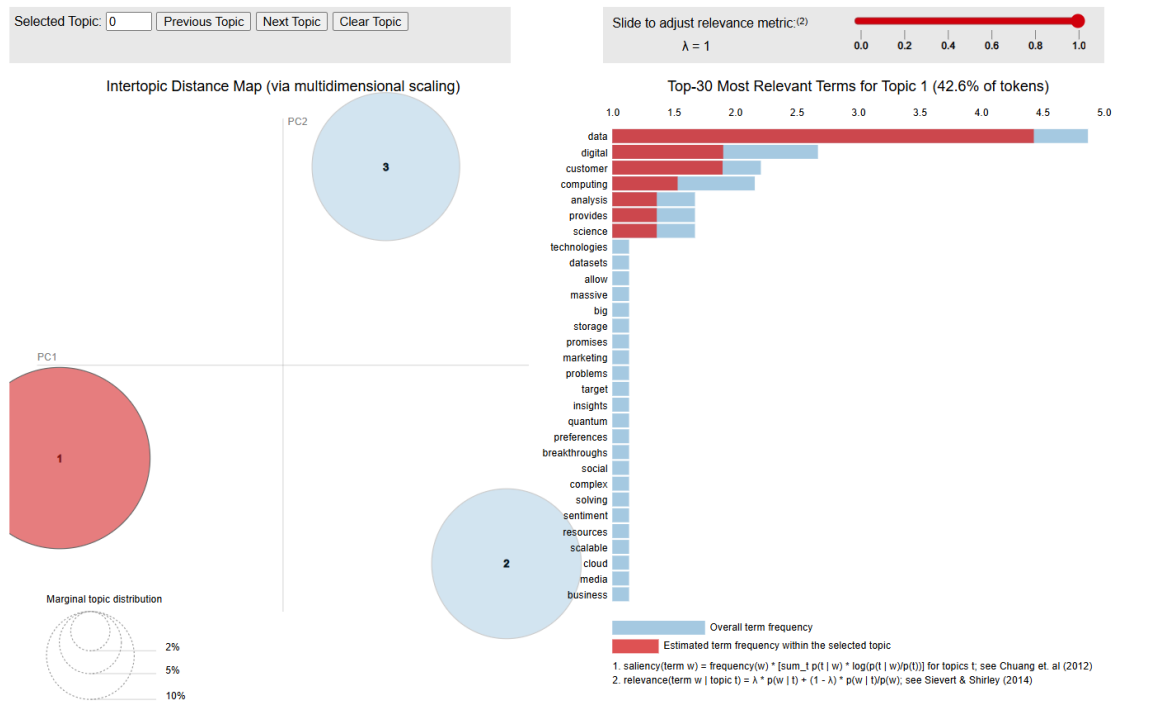
Latent Dirichlet Allocation (LDA) is used to discover hidden topics in the corpus. Each topic is represented by its top keywords, and documents are assigned the most dominant topic. This reveals major themes such as AI/ML, Data Analysis, Cybersecurity, IoT, etc.

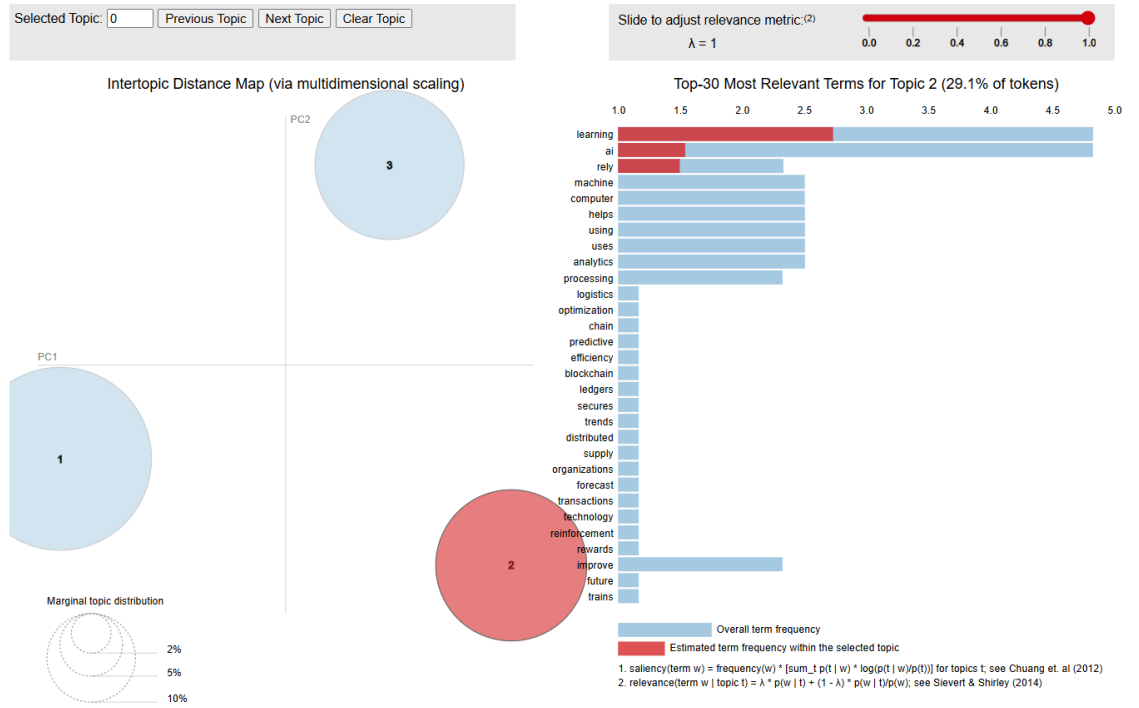
5. Visualization

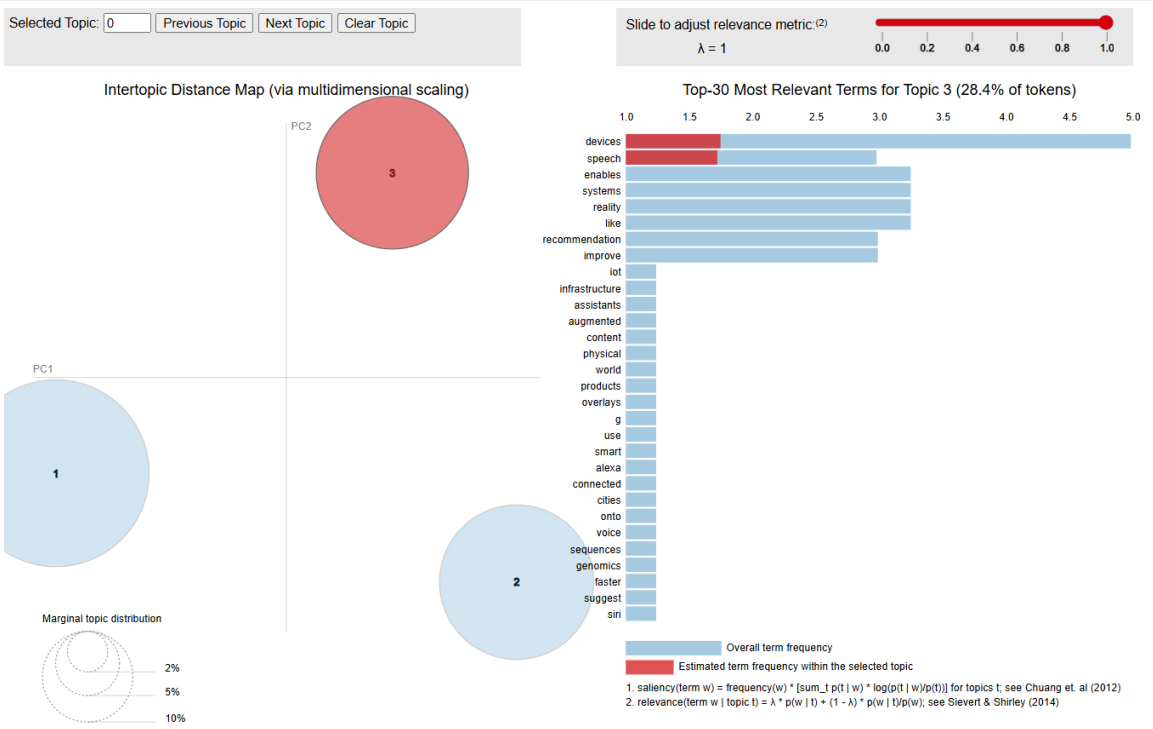
pyLDAvis is used for interactive visualization of topics. The output is saved as an HTML file (`lda_topics.html`) which can be opened in a browser. Additionally, Matplotlib is used to generate static bar plots of top words per topic.

Results & Observations

1. TF-IDF identified the most important and unique words in each document.
2. Word2Vec successfully grouped semantically similar words together.
3. LDA revealed 3–5 meaningful topics across the documents.
4. Visualization clearly showed topic distributions and top keywords.







Discussion & Conclusion

This project highlights the power of NLP techniques in extracting insights from unstructured text. TF-IDF helps in keyword extraction, Word2Vec captures semantic similarities, and LDA provides document-level themes. Together, they enable a deeper understanding of text data, making them useful for applications such as sentiment analysis, recommendation systems, and knowledge discovery.