Internship Final Project Report

Document Clustering for Topic Modeling on Twenty Newsgroups Dataset

Name: Rosy Mondal
Internship Organization: Celebal Technologies
Duration: Summer Internship 2025
Project Title: Document Clustering for Topic Modeling on Twenty Newsgroups Dataset

1. Introduction

This internship project at Celebal Technologies involved unsupervised text analysis using advanced machine learning techniques. The primary goal was to extract meaningful insights from large text corpora by applying topic modeling and clustering methods.

The project used the **Twenty Newsgroups dataset**, a popular benchmark dataset in Natural Language Processing (NLP) comprising approximately 20,000 documents categorized into 20 distinct newsgroups.

---

2. Objective

To leverage unsupervised learning algorithms, specifically Latent Dirichlet Allocation (LDA) and K-Means clustering, for extracting topics and grouping similar documents. The objective was to explore, visualize, and understand the inherent structure of textual data.

---

3. Tools and Technologies Used

- Python
- Jupyter Notebook
- Scikit-learn
- NLTK
- Gensim
- Matplotlib, Seaborn
- Principal Component Analysis (PCA)

4. Workflow

a. Data Preprocessing

- Loaded the Twenty Newsgroups dataset
- Removed headers, footers, and quotes
- Cleaned and tokenized the text data (stop word removal, punctuation cleaning, lemmatization)

b. Vectorization

- Transformed text into numerical format using TF-IDF Vectorizer and Count Vectorizer

c. Topic Modeling (LDA)

Applied Latent Dirichlet Allocation to uncover hidden thematic structures. Top keywords were extracted for each topic. Example topics:

- **Topic 0:** god, bible, jesus, christian, religion
- **Topic 1:** windows, graphics, image, driver, file
- **Topic 2:** team, game, season, win, player

d. Clustering (K-Means)

- Implemented KMeans clustering on TF-IDF vectors
- Used PCA to reduce dimensions for 2D visualization
- Visualized and analyzed document clusters

e. Evaluation

- Compared LDA topic distributions with KMeans clusters
- Identified dominant topics per document
- Summarized and counted documents per cluster and topic

5. Summary and Output

- Document-topic associations and dominant topic mappings were exported to CSV
- KMeans cluster labels were saved and visualized
- PCA plots were generated to illustrate the separation between clusters

The results demonstrated how unsupervised learning can uncover semantic groupings and hidden themes within large text corpora. Topic modeling offered thematic insight, while clustering grouped related documents efficiently.

6. Conclusion

This internship project offered substantial hands-on experience with unsupervised machine learning techniques applied to real-world NLP tasks. By working with topic modeling and clustering, I gained deeper understanding and practical proficiency in text mining and data analysis. Successfully completing this project has strengthened both my confidence and readiness for future data science opportunities.

---

7. Acknowledgements

I extend my heartfelt gratitude to **Celebal Technologies** for this enriching opportunity. I also thank my mentor and team for their continuous support and guidance throughout the internship.

Rosy Mondal
B.Tech in Information Technology, Jalpaiguri Government Engineering College
B.Sc. (Hons) in Data Science and Artificial Intelligence, IIT Guwahati