NLP: Topic Modeling on COVID-19 Research Papers

Project Overview

Topic modeling is an **unsupervised machine learning technique** that helps uncover hidden thematic structures in large textual datasets. In this project, I explored **Latent Dirichlet Allocation (LDA)** and compared it with **Latent Semantic Analysis (LSA)** and **Top2Vec**, using research abstracts from the **NIH COVID-19 Portfolio** (as of November 21, 2020). The objective was to extract meaningful topics from 65,292 research abstracts related to COVID-19, helping us answer: **What themes are prevalent in scientific research on COVID-19?**

Why Topic Modeling?

Topic modeling allows us to:

- Identify dominant topics in a collection of documents
- Automatically classify new documents
- Perform exploratory data analysis in massive text corpora

Each document is represented as a **probability distribution over topics**, and each topic is a **distribution over words**.

Methodology

Data Source:

- **NIH COVID-19 Portfolio**, containing titles and abstracts of scientific papers.

Preprocessing Steps:

- **Stopword Removal**: Using nltk and spaCy
- **Tokenization**: Break text into meaningful units
- **Bigram & Trigram Modeling**: Using Gensim's Phrases to identify multi-word expressions
- **Lemmatization**: Convert words to their base form (e.g., "running" → "run")
- **Noise Removal**: Email addresses, punctuation, and irrelevant characters

Tools & Libraries Used

Tool	**Purpose**
Gensim	LDA, Bigrams/Trigrams
Mallet LDA	Alternate LDA implementation
pyLDAvis	Topic visualization
spaCy	NLP preprocessing
pandas/numpy	Data manipulation
re	Regex-based cleaning
matplotlib	Data visualization

Modeling Techniques

1 Base LDA Model

Initial implementation using Gensim's LDA with 11 pre-defined topics.

2 Tuning Hyperparameters (Alpha)

We experimented with different alpha values to explore topic sparsity and distribution.

3 Mallet LDA

A Gibbs sampling-based alternative to standard LDA with often better topic separation.

4 pyLDAvis Visualization

Each topic appears as a bubble. Larger bubbles indicate more prevalent topics. Ideal models show large, **non-overlapping bubbles** dispersed across the chart.

Experimental Comparison

We went beyond just building LDA. The project extended into comparing:

Method	Strengths
LDA	Well-known, widely supported; interpretable
LSA	Fast; requires fewer computational resources
Top2Vec	No need to predefine topic count; minimal preprocessing

Correlation Analysis

Using **Spearman correlation**, we analyzed topic alignment between titles and abstracts. Key findings:

- **LDA and Top2Vec** showed the highest topic similarity
- **LSA and LDA** shared moderate similarities
- **Top2Vec and LSA** had lower similarity but less preprocessing dependency

Key Insights & Recommendations

- **LDA vs Top2Vec:** High similarity. Use Top2Vec for flexibility and low preprocessing. Use LDA for better interpretability with human input.
- **LDA vs LSA:** Comparable, but LSA is faster and ideal for smaller setups or low-resource environments.
- **Top2Vec:** Automatically determines the number of topics. Great for generating both major and minor themes. However, without constraints, it can produce too many topics for interpretability.

Resource Note:

- Running LDA or Top2Vec on large corpora (65k+ documents) requires **research computing resources**. Desktop-based tools may crash or timeout.

Conclusion

This project serves as a **practical guide** for researchers looking to apply topic modeling to scientific corpora. We recommend:

- Use **LDA** for controlled, interpretable modeling when you can afford high preprocessing.
- Use **Top2Vec** for speed, low setup, and automatic topic generation — especially when resources and time are limited.
- Consider **LSA** for lightweight modeling on local machines or where interpretability is less critical

Citation

If you reference this work, feel free to cite:

Zengul, Ferhat & Bulut, Aysegul & Oner, Nurettin & Ahmed, Abdulaziz & Yadav, Manju & Gray, Hope & Ozaydin, Bunyamin. (2023). A Practical and Empirical Comparison of Three Topic Modeling Methods Using a COVID-19 Corpus: LSA, LDA, and Top2Vec. 10.24251/HICSS.2023.116