The goal of this project is to automatically identify and group hidden topics within a collection of textual data such as news or research articles. Using unsupervised machine learning (Topic Modeling), the system discovers common themes without any manual labeling.

Tools and Libraries Used

- **Python**
- **Google Colab**
- **Libraries:**
  - nltk – for text preprocessing (stopwords, lemmatization)
  - gensim – for implementing Latent Dirichlet Allocation (LDA)
  - pyLDAvis – for interactive topic visualization
  - pandas, re, warnings – for data handling and cleaning

Steps Performed in the Project

1. Library Installation

Installed required Python libraries in Colab using:

```
!pip install nltk gensim pyLDAvis
```

These libraries were used for text processing, topic modeling, and visualization.

2. Importing Libraries

All necessary modules were imported — pandas for data management, nltk for preprocessing, gensim for model creation, and pyLDAvis for visualization.

Deprecation warnings were safely ignored to maintain a clean output.

3. Loading Dataset

A small custom dataset was created consisting of short news/research-style sentences such as:

- AI in healthcare

- Stock market updates
- Deep learning in cancer analysis
- Political tensions and global trade

The dataset was stored in a pandas DataFrame named df.

4. Text Preprocessing

To make the text ready for modeling:

- All text was **converted to lowercase**.
- **Special characters and punctuation** were removed using regex.
- **Stopwords** (like “the”, “is”, “in”) were removed using NLTK’s English stopword list.
- **Lemmatization** was performed using WordNetLemmatizer to convert words to their root form (e.g., “diagnostics” → “diagnostic”).

Each cleaned article was stored in a new column Cleaned.

5. Dictionary and Corpus Creation

- A **dictionary** was created mapping each word to a unique ID.
- A **corpus** (bag-of-words representation) was generated — this is what the LDA model uses to identify topics.

```
dictionary = corpora.Dictionary(df['Cleaned'])

corpus = [dictionary.doc2bow(text) for text in df['Cleaned']]
```

6. Training the LDA Model

The **Latent Dirichlet Allocation (LDA)** model was trained using:

```
models.LdaModel(corpus=corpus, id2word=dictionary, num_topics=3, passes=15, random_state=42)
```

- num_topics=3 means the model tries to find 3 distinct themes in the dataset.
- passes=15 ensures good convergence for topic distribution.

7. Viewing Discovered Topics

The model displayed 3 topics with the most important words (keywords) in each.

Output Example:

Topic 0: medical, model, accuracy, machine, ai, improved

Topic 1: sector, stock, market, gain, technology, sustainable, future

Topic 2: learning, pattern, trade, inflation, global, rate, economic

8. Visualization of Topics

An **interactive visualization** was generated using pyLDAvis, where:

- Each bubble represents a **topic**.
- The **distance between bubbles** shows how distinct topics are.
- The **bar charts** show the most frequent keywords in each topic.

This visualization allows for deeper understanding and validation of topic separation.

9. Final Topic Interpretation

Topic	Top Keywords	Interpretation
Topic 0	medical, model, ai, machine, healthcare	Technology & Healthcare
Topic 1	market, sector, stock, gain, sustainable, future	Economy & Trade
Topic 2	learning, inflation, global, trade, economic	*Research & Global Economy*

Results Summary

- **Model Used:** Latent Dirichlet Allocation (LDA)
- **Number of Topics Extracted:** 3
- **Dataset Used:** 7 short news/research-like articles

- **Output Type:** List of discovered topics and an interactive visualization

Conclusion

This project successfully demonstrated how **unsupervised topic modeling** can uncover hidden themes from unstructured text data.
Using **LDA**, we identified 3 clear topics from a small dataset — Technology & Healthcare, Economy & Trade, and Global Research Themes.

The combination of **NLTK for preprocessing**, **Gensim for modeling**, and **PyLDAvis for visualization** provided a complete and interpretable topic modeling workflow.

Key Learnings

- Gained hands-on experience in Natural Language Processing (NLP).
- Learned how to preprocess and clean textual data effectively.
- Understood the working of LDA for topic extraction.
- Learned how to visualize and interpret topic distributions in a dataset.

Final Remarks

Project Title: Topic Modeling (News or Research Articles)
Developed by: *Shivaya*
Tools Used: Python, Google Colab, NLTK, Gensim, PyLDAvis
Status: *Successfully Completed*