## **Project Overview**

This project focuses on **Topic Modelling** using Python to uncover hidden topics within text documents. The goal is to extract meaningful topics from a collection of articles using **Natural Language Processing (NLP)** techniques and machine learning algorithms. Specifically, **Latent Dirichlet Allocation (LDA)** is employed to assign topic labels to documents based on word patterns.

## **Data Processing**

1. **Loading Data**:
   * A dataset of articles is loaded from a CSV file.
2. **Text Preprocessing**:
   * **Lowercasing**: All text is converted to lowercase to ensure uniformity.
   * **Punctuation Removal**: Removes special characters to avoid noise in the data.
   * **Tokenization**: Splits text into individual words.
   * **Stopword Removal**: Removes common words (like "the", "and") that do not add much meaning.
   * **Lemmatization**: Converts words to their base form (e.g., "running" → "run").

## **Methods Used**

1. **Text Vectorization**:
   * **TF-IDF Vectorizer**: Converts text into numerical vectors, capturing word importance across documents.
2. **Topic Modelling Algorithm**:
   * **Latent Dirichlet Allocation (LDA)**:
     + Identifies topics by finding patterns in word usage.
     + The number of topics (n\_components) is set to 5, meaning the algorithm will group the documents into 5 distinct topics.

## **Results**

* **Topic Assignment**:
  + The LDA algorithm assigns topic labels to each article.
  + Example output shows which topic label each article belongs to based on the patterns identified.

**Sample Output**:

| **Title** | **Topic Label** |
| --- | --- |
| Best Books to Learn Data Analysis | 2 |
| Assumptions of Machine Learning Algorithms | 3 |
| News Classification with Machine Learning | 1 |
| Multiclass Classification Algorithms | 3 |
| Multinomial Naive Bayes in Machine Learning | 1 |

## **Summary**

This project demonstrates the use of **NLP and machine learning techniques** to assign topic labels to articles. LDA successfully uncovers hidden topics by analyzing word patterns and relationships. With proper preprocessing and text vectorization, topic modeling becomes a powerful tool for extracting insights from large textual datasets.