# **Topic Modeling Project (English Version)**

## **1. Introduction**

Topic modeling is an unsupervised machine learning method used to discover abstract topics within a collection of documents. In this project, we aim to extract and analyze the main themes from English news articles using Natural Language Processing (NLP) and Non-negative Matrix Factorization (NMF).

## **2. Data Preprocessing**

To prepare the text for modeling:

* Converted text to lowercase.
* Removed special characters using regular expressions.
* Tokenized words using nltk.word\_tokenize.
* Removed stopwords in English.
* Lemmatized tokens using WordNetLemmatizer.
* Joined tokens into clean strings for vectorization.
* Filled missing author values with 'Unknown' and dropped the url column.

## **3. Exploratory Data Analysis (EDA)**

* **Most Frequent Words**: Visualized with a bar chart using Counter.
* **Word Cloud**: Provided an intuitive visualization of the top words across all articles.

## **4. Feature Extraction**

Two approaches were used:

* **Bag of Words (BOW)** with CountVectorizer
* **TF-IDF** with TfidfVectorizer

Parameters:

* max\_features=10000
* min\_df=5
* max\_df=0.7

## **5. Hyperparameter Tuning**

## To determine the optimal number of topics:

## A range of topic counts from 10 to 30 (in steps of 10) was evaluated.

## Reconstruction error from the NMF model was plotted against topic count.

## The topic count with the lowest reconstruction error was selected as optimal.

## **6. Topic Modeling with NMF**

* **Model**: Non-negative Matrix Factorization (NMF)
* **Input**: TF-IDF vectors
* **Evaluation**: Reconstruction error used to find the best number of topics
* **Topic Interpretation**: Top 4 keywords per topic were printed for interpretability.

## **7. Results Visualization**

* **Word Cloud** of all articles.
* **Bar Chart** showing the number of documents per topic.
* **Top Keywords per Topic** were printed to provide insight into the nature of each topic.

## **8. Model Evaluation**

* Evaluated using **Silhouette Score** with cosine distance to assess how well documents fit into the learned topics.  
    
    
   **Silhouette Score**: 0.030

## **9. Saving the Model**

Both the trained **NMF model** and the **TF-IDF vectorizer** were saved using pickle for future inference.

Files saved:

* nmf\_model.pkl
* tfidf\_vectorizer.pkl

## **10. Conclusion**

* Successfully implemented topic modeling using NMF on English news articles.
* Tuned and selected the optimal number of topics based on reconstruction error.
* Extracted meaningful topics with keyword summaries and document distributions.