# Text Preprocessing and Visualization with TF-IDF

## 1. Overview

This project explores text preprocessing and feature extraction using TF-IDF on a collection of textual data. The primary focus is on cleaning text, converting it into numerical format using TF-IDF, and visualizing important terms using WordCloud.

## 2. Objective

To clean raw text, extract term importance using TF-IDF vectorization, and visualize the most frequent or significant terms to support future classification tasks.

## 3. Techniques Used

- Text Preprocessing: Lowercasing, punctuation removal, stopword filtering  
- TF-IDF vectorization for converting text into numerical features  
- Word Frequency Analysis  
- WordCloud for visualizing key terms

## 4. Tools & Libraries

- Python  
- Pandas, NumPy  
- Scikit-learn (TF-IDF)  
- WordCloud, Matplotlib  
- NLTK / spaCy for text cleaning

## 5. Outcome

The notebook successfully demonstrates the use of TF-IDF for numerical feature extraction and word frequency visualization. Key terms were identified and visualized, offering insights into common themes and vocabulary. This serves as a foundational step for future classification models.

## 6. Future Work

- Train and evaluate text classification models (e.g., Logistic Regression, Naive Bayes)  
- Compare TF-IDF with CountVectorizer and n-gram extensions  
- Integrate model into a FastAPI or Streamlit app for deployment