Assignment 2

Name: Adhija Satish Bachhav

Roll No: 381053

PRN: 22311656

Problem Statement

To perform text preprocessing on a given textual dataset by applying text cleaning, lemmatization, stop word removal, and label encoding. Further, to generate numerical representations of the processed text using the TF-IDF technique and save the obtained outputs for further analysis.

Objective

The objective of this assignment is to understand and implement the fundamental steps involved in Natural Language Processing (NLP) preprocessing. These steps help convert raw text data into a structured and machine-readable format suitable for machine learning models.

Terminologies Used

1. Text Cleaning

Text cleaning is the process of removing unwanted elements such as punctuation, numbers, and special characters from raw text. It also involves converting text to lowercase to maintain uniformity.

2. Tokenization

Tokenization refers to splitting text into smaller units called tokens, usually words. It is a crucial step for further text processing operations.

3. Lemmatization

Lemmatization is the process of converting words into their base or root form. For example, “running” becomes “run”. This helps reduce vocabulary size while preserving meaning.

4. Stop Words

Stop words are commonly used words such as “is”, “the”, “and”, etc., which do not add significant meaning to the text. Removing them helps improve model efficiency.

5. Label Encoding

Label encoding converts categorical labels into numerical form. Each unique label is assigned a numeric value, making it suitable for machine learning algorithms.

6. TF-IDF (Term Frequency – Inverse Document Frequency)

TF-IDF is a numerical representation technique that reflects how important a word is to a document in a collection. It

assigns higher values to words that are frequent in a document but rare across documents.

Methodology

The following steps were performed sequentially:

1. **Raw text data was loaded into a structured format.**
2. **Text cleaning was applied to remove noise and standardize text.**
3. **The cleaned text was tokenized and lemmatized to obtain root words.**
4. **Stop words were removed to retain meaningful information.**
5. **Labels were converted into numerical form using label encoding.**
6. **TF-IDF vectorization was applied to generate numerical feature representations.**
7. **The processed dataset and TF-IDF features were saved as output files.**

Output Description

- **Processed Text Dataset: Contains original text, cleaned text, lemmatized text, final processed text, and encoded labels.**
- **TF-IDF Feature File: Contains numerical vectors representing the importance of each word in the documents.**

These outputs can be directly used for classification, clustering, or other machine learning tasks.

Applications

- **Text classification**
- **Sentiment analysis**
- **Information retrieval**
- **Spam and fraud detection**
- **Document similarity analysis**

Conclusion

This assignment demonstrates the complete pipeline of text preprocessing and feature extraction using TF-IDF. Proper preprocessing significantly improves the performance of NLP models by reducing noise and enhancing meaningful patterns in text data. The generated outputs serve as a strong foundation for further machine learning and NLP-based applications.

Result

Text preprocessing was successfully performed, and TF-IDF representations were generated and saved. The objective of the assignment was achieved effectively.