Assignment No: - 3

Name :- Isha Ghorpade

Enrollnment :- 22420020

Roll :- 381066

Problem Statement

Perform text cleaning, perform lemmatization (any method), remove stop words (any method), label encoding. Create representations using TF-IDF. Save outputs

Objectives

- To understand the importance of text cleaning in Natural Language Processing (NLP).
- To perform lemmatization to obtain meaningful base words.
- To remove stop words to reduce noise in text data.
- To apply label encoding for converting categorical labels into numeric form.
- To create numerical text representations using the TF-IDF technique.
- To store the processed outputs for further analysis.

Software Packages Used

- Operating System: Windows / Linux / macOS
- Programming Language: Python 3.x
- Development Tools: Jupyter Notebook / Google Colab / VS Code
- Libraries Used:
  - NLTK
  - Scikit-learn
  - Pandas

Hardware Apparatus Used

- Computer or Laptop
- Processor: Intel / AMD or equivalent
- Minimum 4 GB RAM
- Internet connection (for downloading libraries and resources)

Theory

Text Cleaning

Text cleaning is the process of removing unwanted and irrelevant parts of text so that the data becomes suitable for analysis.

Text Cleaning Includes

- Converting text to lowercase
- Removing punctuation
- Removing numbers
- Removing extra spaces
- Removing special characters

Diagram

Raw Text

↓

Cleaning (lowercase, remove symbols)

↓

Clean Text

Stop Word Removal

Stop words are commonly used words that do not add much meaning to the text.

Examples:
is, the, and, in, on, at, of

Removing stop words:

- Reduces data size
- Improves model performance

Diagram

Text: This is a simple example

↓

Remove stop words

↓

simple example

Lemmatization

Lemmatization converts words into their dictionary base form (lemma) by considering grammar and meaning.

- More accurate than stemming
- Produces meaningful words

Examples:

- running → run
- better → good

Diagram

Words

↓

Lemmatization

↓

Meaningful root words

Label Encoding

Label encoding converts categorical labels into numerical values so that machine learning models can understand them.

Example:

Positive → 1

Negative → 0

Neutral → 2

Diagram

Text Labels

↓

Label Encoding

↓

Numeric Labels

TF-IDF Representation

TF-IDF (Term Frequency – Inverse Document Frequency) converts text into numerical vectors based on word importance.

- TF measures word frequency in a document
- IDF measures word rarity across documents

TF-IDF Flow Diagram

Cleaned Text

↓

Stop Word Removal

↓

Lemmatization

↓

TF-IDF Vector

Saving Outputs

After preprocessing and feature extraction:

- Cleaned text is stored in files or dataframes
- TF-IDF vectors are saved for model training
- Encoded labels are stored for supervised learning

Applications

- Sentiment analysis
- Text classification
- Spam detection
- Document clustering
- Machine learning and NLP projects

Conclusion

Text cleaning, lemmatization, and stop word removal improve text quality and reduce noise.Label encoding helps convert text labels into machine-readable form.TF-IDF provides an effective numerical representation of text.Saving processed outputs allows efficient reuse of data for training and evaluation in NLP applications.