Assignment No: - 1

Name :- Isha Ghorpade

Enrollnment :- 22420020

Roll :- 381066

Problem Statement

Perform tokenization (Whitespace, Punctuation-based, Treebank, Tweet, MWE) using NLTK library.

Use porter stemmer and snowball stemmer for stemming. Use any technique for lemmatization.

Objectives

- To understand the concept of text preprocessing in Natural Language Processing (NLP).
- To study different tokenization techniques using the NLTK library.
- To learn stemming using Porter Stemmer and Snowball Stemmer.
- To understand lemmatization and its importance in obtaining meaningful root words.
- To analyze how preprocessing improves the performance of NLP applications

S/W Packages and H/W apparatus used

- **Operating System:** Windows / Linux / macOS
- **Programming Language:** Python 3.x
- **Development Tools:** Jupyter Notebook / Google Colab / VS Code
- **Library Used:** NLTK (Natural Language Toolkit)

Hardware Apparatus Used

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

Theory

Text Preprocessing

Text preprocessing is the process of cleaning and transforming raw text into a structured format that can be easily processed by machines.
Important steps include tokenization, stemming, and lemmatization.

Tokenization

Tokenization is the process of dividing text into smaller units called tokens. These tokens help machines understand the structure of language.

Types of Tokenization

1. Whitespace Tokenization

Splits text based on spaces.

Example Diagram:

Example Diagram:

Text: I love NLP

↓

Tokens: I | love | NLP

---

2. Punctuation-based Tokenization

Separates punctuation marks along with words.

Example Diagram:

Text: Hello, world!

↓

Tokens: Hello | , | world | !

---

3. Treebank Tokenization

- Rule-based tokenizer in NLTK
- Handles punctuation and contractions

Example Diagram:

Text: Don't worry!

↓

Tokens: Do | n't | worry | !

---

4. Tweet Tokenization

Used for social media text.
Handles hashtags, mentions, emojis, and informal words.

Example Diagram:

Text: Loving #AI @OpenAI

↓

Tokens: Loving | #AI | @OpenAI

---

5. Multi-Word Expression (MWE) Tokenization

Treats a group of words as a single token.

Example Diagram:

machine learning

↓

[machine_learning]

Stemming

Stemming removes suffixes from words to get the root form.
The output may not always be a valid dictionary word.

---

1. Porter Stemmer

- Most commonly used
- Rule-based

Diagram:

Running → Run

Studies → Studi

2. Snowball Stemmer

- Improved and faster version of Porter Stemmer
- Supports multiple languages

Diagram:

Happiness → Happi

Flying → Fly

Lemmatization

Lemmatization converts a word into its **dictionary base form (lemma)** by considering grammar and meaning.

Example Diagram:

Running → Run

Better → Good

- More accurate than stemming
- Uses **WordNet Lemmatizer** in NLTK

Difference Between Stemming and Lemmatization (Diagram)

Word: Running

|

|-- Stemming → run

|

|-- Lemmatization → run (meaningful word)

Conclusion

Tokenization, stemming, and lemmatization are essential steps in NLP.Different tokenizers are used based on text type.Stemming is fast but less accurate, while lemmatization provides meaningful results.Using NLTK makes text preprocessing easy and efficient for real-world applications.
