## Assignment 1

## Name : Adhija Satish Bachhav

## Roll No.: 381053

## PRN: 22311656

## Perform tokenization (Whitespace, Punctuation-based, Treebank, Tweet, MWE) using NLTK library. Use porter stemmer and snowball stemmer for stemming. Use any technique for lemmatization.

---

# Text Preprocessing in NLP

## Tokenization

Tokenization is the process of breaking text into smaller units called tokens. These tokens can be words, symbols, or meaningful phrases.

### Types of Tokenization

- **Whitespace Tokenization:** Splits text based on spaces.
- **Punctuation-based Tokenization:** Separates words and punctuation marks.
- **Treebank Tokenization:** Handles punctuation and contractions accurately.
- **Tweet Tokenization:** Processes social media text including hashtags and mentions.
- **MWE Tokenization:** Treats multi-word expressions as a single token.

---

# Stemming

Stemming is the process of reducing words to their root form by removing suffixes. The resulting stem may not be a valid dictionary word.

### Types of Stemming

- **Porter Stemmer:** Rule-based stemmer for English.
- **Snowball Stemmer:** Improved version of Porter stemmer with better accuracy.

---

# Lemmatization

Lemmatization is the process of converting words into their meaningful dictionary form called a lemma. It uses grammatical rules and part-of-speech information to preserve meaning.

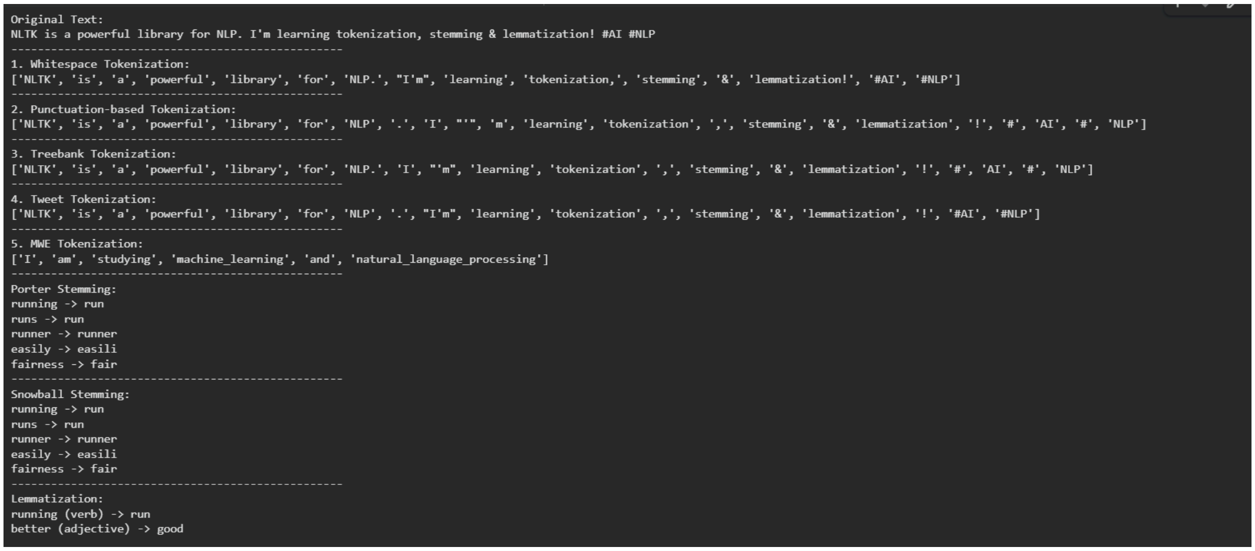
---

# Difference Between Stemming and Lemmatization

| Feature | Stemming | Lemmatization |
|---|---|---|
| Method | Rule-based | Dictionary-based |
| Output | Not always meaningful | Meaningful word |
| Accuracy | Lower | Higher |

---

# Output



```
Original Text:
NLTK is a powerful library for NLP. I'm learning tokenization, stemming & lemmatization! #AI #NLP
--------------------------------------------------
1. Whitespace Tokenization:
['NLTK', 'is', 'a', 'powerful', 'library', 'for', 'NLP.', "I'm", 'learning', 'tokenization,', 'stemming', '&', 'lemmatization!', '#AI', '#NLP']
--------------------------------------------------
2. Punctuation-based Tokenization:
['NLTK', 'is', 'a', 'powerful', 'library', 'for', 'NLP', '.', 'I', "'", 'm', 'learning', 'tokenization', ',', 'stemming', '&', 'lemmatization', '!', '#', 'AI', '#', 'NLP']
--------------------------------------------------
3. Treebank Tokenization:
['NLTK', 'is', 'a', 'powerful', 'library', 'for', 'NLP.', 'I', "'m", 'learning', 'tokenization', ',', 'stemming', '&', 'lemmatization', '!', '#', 'AI', '#', 'NLP']
--------------------------------------------------
4. Tweet Tokenization:
['NLTK', 'is', 'a', 'powerful', 'library', 'for', 'NLP', '.', "I'm", 'learning', 'tokenization', ',', 'stemming', '&', 'lemmatization', '!', '#AI', '#NLP']
--------------------------------------------------
5. MWE Tokenization:
['I', 'am', 'studying', 'machine_learning', 'and', 'natural_language_processing']
--------------------------------------------------
Porter Stemming:
running -> run
runs -> run
runner -> runner
easily -> easili
fairness -> fair
--------------------------------------------------
Snowball Stemming:
running -> run
runs -> run
runner -> runner
easily -> easili
fairness -> fair
--------------------------------------------------
Lemmatization:
running (verb) -> run
better (adjective) -> good
```