Text Preprocessing Exercise (Using Basic Python)

Objective

In this exercise, you will perform basic text preprocessing steps on a given text sample using only built-in Python functions. This will help you understand the fundamental concepts of text processing in NLP.

Task Description

You are provided with a raw text sample. Your task is to preprocess this text by performing the following steps:

1. Tokenization: Split the text into individual words (tokens).
2. Lowercasing: Convert all tokens to lowercase.
3. Punctuation Removal: Remove all punctuation marks from the tokens.
4. Stop Word Removal: Remove common stop words (e.g., "the", "is", "and").
5. Stemming: Reduce words to their root form using a simple algorithm.

Data

Use the following text for your preprocessing task:

Natural Language Processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language. It's used to analyze text, allowing machines to understand, interpret, and manipulate human language. NLP has many real-world applications, including machine translation, sentiment analysis, and chatbots.

Expected Output

Provide the following:

1. The list of tokens after step 1.
2. The list of tokens after steps 2 and 3.
3. The list of tokens after step 4 (stop word removal).
4. The final list of tokens after step 5 (simple stemming).

Bonus (Optional)

Implement a simple lemmatization function that uses a dictionary to map common words to their base form (e.g., "is" -> "be", "are" -> "be"). Compare the results with your stemming function.

Submission

Submit your Python code along with the output at each step. Include comments explaining your approach for each step.

Hints

- For stop words, start with a small list like: ["the", "a", "an", "in", "on", "at", "for", "to", "of", "and", "is", "are"]
- For stemming, focus on removing common suffixes. Don't worry about handling all possible cases.