Practical No:1

Title:

Convert the text into tokens. Find the word frequency.

Aim:

To write a program that converts a given text into tokens and computes the frequency of each word.

Pre-requisites:

1. Understanding of Natural Language Processing (NLP) concepts, such as tokenization and word frequency.
2. Familiarity with basic string manipulation techniques in programming.
3. Knowledge of basic data structures like lists, dictionaries, and sets.
4. Understanding of how to handle text data, such as preprocessing and cleaning.

Theory:

Natural Language Processing (NLP) is a field of Artificial Intelligence (AI) that enables computers to understand, interpret, and manipulate human language. One of the foundational tasks in NLP is **tokenization** and **word frequency analysis**. These steps are essential for understanding the structure and meaning of the text.

1. **Tokenization:**
    - Tokenization is the process of breaking down text into smaller units called **tokens**. Tokens can be

words, sentences, or even characters, depending on the task.
    - **Word Tokenization** refers to splitting the text into individual words. This is typically done by splitting on whitespace and punctuation marks.
2. **Word Frequency:**
    - **Word frequency** refers to the number of times each word appears in a given text. Finding word frequency is useful in understanding the importance of words in a text, detecting common themes, and supporting further text analysis.
    - A common representation of word frequency is a dictionary where each unique word is a key, and its frequency is the value.

Steps for Tokenization and Word Frequency Calculation:

1. **Input Text:**
    - A string of text is provided as input.
2. **Text Preprocessing:**
    - **Lowercasing:** Convert all characters in the text to lowercase to ensure that word frequency is case-insensitive.
    - **Removing Punctuation:** Strip out punctuation marks such as periods, commas, and question marks to treat words with the same meaning equally (e.g., "word" and "word," are considered the same).
    - **Tokenization:** Split the preprocessed text into individual words.
3. **Word Frequency Calculation:**
    - Traverse the list of tokens.

    - For each token, increment its count in a dictionary (or hash map) if it already exists, otherwise, initialize the count to 1.
4. **Output:**
    - Print the tokens.
    - Display the word frequency as a list of words with their corresponding counts.

Conclusion:

Tokenization and word frequency calculation are fundamental tasks in Natural Language Processing. Tokenization breaks the text into meaningful units, and word frequency helps to understand the distribution and importance of words. By preprocessing the text (lowercasing, removing punctuation), we ensure accurate results, regardless of variations in word forms. These tasks form the foundation for more advanced NLP techniques, such as sentiment analysis, topic modeling, and machine translation.