Experiment - 4

1. Algorithm:   
   **Import NLTK and Download Resources:**
   * Import the necessary modules from NLTK, such as **word\_tokenize**, **PorterStemmer**, **WordNetLemmatizer**, and **stopwords**.
   * Download required NLTK resources using **nltk.download()**.
2. **Define Text:**
   * Input the text you want to process.
3. **Tokenization:**
   * Use **sent\_tokenize** to split the text into sentences.
   * For each sentence, use **word\_tokenize** to split it into words.
4. **Stemming:**
   * Create an instance of **PorterStemmer**.
   * For each word token, apply stemming using the **stem** method of the stemmer.
5. **Lemmatization:**
   * Create an instance of **WordNetLemmatizer**.
   * For each word token, apply lemmatization using the **lemmatize** method of the lemmatizer.

import nltk

from nltk.tokenize import word\_tokenize

from nltk.stem import PorterStemmer, WordNetLemmatizer

from nltk.corpus import stopwords

# Download NLTK resources if not already installed

nltk.download('punkt')

nltk.download('stopwords')

nltk.download('wordnet')

# Sample text

text = "The quick brown foxes are jumping over the lazy dogs. The dogs are not amused."

# Tokenization

tokens = word\_tokenize(text)

# Remove stop words

stop\_words = set(stopwords.words('english'))

filtered\_tokens = [word for word in tokens if word.lower() not in stop\_words]

# Stemming

stemmer = PorterStemmer()

stemmed\_tokens = [stemmer.stem(word) for word in filtered\_tokens]

# Lemmatization

lemmatizer = WordNetLemmatizer()

lemmatized\_tokens = [lemmatizer.lemmatize(word) for word in filtered\_tokens]

# Display the results

print("Original Text:", text)

print("\nTokenization:", tokens)

print("\nFiltered Tokens (without stop words):", filtered\_tokens)

print("\nStemmed Tokens:", stemmed\_tokens)

print("\nLemmatized Tokens:", lemmatized\_tokens)

Output :

[nltk\_data] Downloading package punkt to /root/nltk\_data...

[nltk\_data] Unzipping tokenizers/punkt.zip.

[nltk\_data] Downloading package stopwords to /root/nltk\_data...

[nltk\_data] Unzipping corpora/stopwords.zip.

[nltk\_data] Downloading package wordnet to /root/nltk\_data...

Original Text: The quick brown foxes are jumping over the lazy dogs. The dogs are not amused.

Tokenization: ['The', 'quick', 'brown', 'foxes', 'are', 'jumping', 'over', 'the', 'lazy', 'dogs', '.', 'The', 'dogs', 'are', 'not', 'amused', '.']

Filtered Tokens (without stop words): ['quick', 'brown', 'foxes', 'jumping', 'lazy', 'dogs', '.', 'dogs', 'amused', '.']

Stemmed Tokens: ['quick', 'brown', 'fox', 'jump', 'lazi', 'dog', '.', 'dog', 'amus', '.']

Lemmatized Tokens: ['quick', 'brown', 'fox', 'jumping', 'lazy', 'dog', '.', 'dog', 'amused', '.']