1. **Write a Python program for the following preprocessing of text in NLP:**

* **Tokenization**
* **Filtration**
* **Script Validation**
* **Stop Word Removal**
* **Stemming**

**Program:**

import nltk

import re

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

from nltk.stem import PorterStemmer

# Download necessary NLTK resources

nltk.download('punkt')

nltk.download('stopwords')

def preprocess\_text(text):

text = text.replace('\u00A0', ' ')

    # Step 1: Tokenization

    tokens = word\_tokenize(text)

    print("Tokens:", tokens)

    # Step 2: Filtration (remove special characters, numbers, etc.)

    filtered\_tokens = [word for word in tokens if re.match(r'^[a-zA-Z]+$', word)]

    print("Filtered Tokens:", filtered\_tokens)

    # Step 3: Script Validation (ensure all tokens are in English script)

    # Assuming the text is already in English, no further action is needed.

    # If not, you can use a language detection library like `langdetect`.

    # Step 4: Stop Word Removal

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

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

    print("Tokens without Stopwords:", tokens\_without\_stopwords)

    # Step 5: Stemming

    stemmer = PorterStemmer()

    stemmed\_tokens = [stemmer.stem(word) for word in tokens\_without\_stopwords]

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

    return stemmed\_tokens

# Example Usage

text = "This is an example text! It includes different words, numbers like 123, and punctuation."

processed\_text = preprocess\_text(text)

print("Processed Tokens:", processed\_text)

**Output:**

Tokens: ['This', 'is', 'an', 'example', 'text', '!', 'It', 'includes', 'different', 'words', ',', 'numbers', 'like', '123', ',', 'and', 'punctuation', '.']

Filtered Tokens: ['This', 'is', 'an', 'example', 'text', 'It', 'includes', 'different', 'words', 'numbers', 'like', 'and', 'punctuation']

Tokens without Stopwords: ['example', 'text', 'includes', 'different', 'words', 'numbers', 'like', 'punctuation']

Stemmed Tokens: ['exampl', 'text', 'includ', 'differ', 'word', 'number', 'like', 'punctuat']

Processed Tokens: ['exampl', 'text', 'includ', 'differ', 'word', 'number', 'like', 'punctuat']