PREPROCESSING TEXT DATASET ON THE GIVEN CASE STUDY

NAME: TAMANNA VAIKKATH

```
!pip install -U spacy --quiet
!python -m spacy download en_core_web_sm
import nltk
nltk.download('punkt')
nltk.download('stopwords')


import spacy
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize


nlp = spacy.load("en_core_web_sm")
stemmer = PorterStemmer()
stop_words = set(stopwords.words('english'))


case_study_text = """
In this project, we aimed to build a simple sentiment analysis tool using Natural Language Processing (NLP) techniques.
The goal was to classify user reviews as positive or negative. We started by collecting a sample dataset of product reviews.
Text preprocessing was done using tokenization, stopword removal, stemming, and lemmatization.
The processed data was then used to train a basic machine learning classifier using the scikit-learn library.
Our results showed that preprocessing significantly improved the accuracy of sentiment classification.
This practical demonstrated how AI techniques can be applied to real-world textual data effectively.
"""


doc = nlp(case_study_text)

# Tokenization
tokens = [token.text for token in doc]
print(" Tokens:\n", tokens)

# Stopword Removal
filtered = [token for token in doc if not token.is_stop and token.is_alpha]
filtered_words = [token.text for token in filtered]
print("\n After Stopword Removal:\n", filtered_words)

# Lemmatization
lemmatized = [token.lemma_ for token in filtered]
print("\n Lemmatized Tokens:\n", lemmatized)

# Stemming (using nltk)
stemmed = [stemmer.stem(token.text) for token in filtered]
print("\n Stemmed Tokens:\n", stemmed)
```

```
Collecting en-core-web-sm==3.8.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.8.0/en_core_web_sm-3.8.0-py3-none-any.whl (12
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.8/12.8 MB 92.6 MB/s eta 0:00:00
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
⚠ Restart to reload dependencies
If you are in a Jupyter or Colab notebook, you may need to restart Python in
order to load all the package's dependencies. You can do this by selecting the
'Restart kernel' or 'Restart runtime' option.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
 Tokens:
 ['\n', 'In', 'this', 'project', ',', 'we', 'aimed', 'to', 'build', 'a', 'simple', 'sentiment', 'analysis', 'tool', 'using', 'Natural',

 After Stopword Removal:
 ['project', 'aimed', 'build', 'simple', 'sentiment', 'analysis', 'tool', 'Natural', 'Language', 'Processing', 'NLP', 'techniques', 'goa

 Lemmatized Tokens:
 ['project', 'aim', 'build', 'simple', 'sentiment', 'analysis', 'tool', 'Natural', 'Language', 'Processing', 'NLP', 'technique', 'goal',

 Stemmed Tokens:
 ['project', 'aim', 'build', 'simpl', 'sentiment', 'analysi', 'tool', 'natur', 'languag', 'process', 'nlp', 'techniqu', 'goal', 'classif
```