

```

import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer

# Download stopwords once
nltk.download('stopwords')

def clean_text(text):
    """Remove special chars, numbers & convert to lowercase"""
    text = text.lower()
    text = re.sub(r'[^a-zA-Z\s]', " ", text)      # Keep only alphabets
    return text

def tokenize(text):
    """Split into words"""
    return text.split()

def remove_stopwords(words):
    """Remove common English stopwords"""
    stop_words = set(stopwords.words('english'))
    return [word for word in words if word not in stop_words]

def stemming(words):
    """Apply stemming"""
    ps = PorterStemmer()
    return [ps.stem(word) for word in words]

def word_frequency(words):
    """Count frequency using dictionary"""
    freq = {}
    for w in words:
        freq[w] = freq.get(w, 0) + 1
    return freq

# ----- Main Program -----
news_articles = [
    "The government has introduced new policies to improve quality education in rural areas.",
    "Digital learning platforms are supporting students with remote education opportunities.",
    "Education for all is essential to achieve Sustainable Development Goal 4 by the United Nations."
]

```

```
print("\n--- RAW ARTICLES ---")
for article in news_articles:
    print(article)

combined_text = " ".join(news_articles)

# Step-by-step preprocessing
cleaned = clean_text(combined_text)
tokens = tokenize(cleaned)
no_stopwords = remove_stopwords(tokens)
stemmed_words = stemming(no_stopwords)
frequency = word_frequency(stemmed_words)

print("\n\n--- AFTER PREPROCESSING ---")
print("Cleaned Text:", cleaned)
print("\nTokens:", tokens)
print("\nAfter Removing Stopwords:", no_stopwords)
print("\nAfter Stemming:", stemmed_words)

print("\n--- WORD FREQUENCY ---")
for word, count in frequency.items():
    print(f"{word} : {count}")
```