NLP Preprocessing

This notebook demonstrates basic preprocessing tasks in NLP, including lowercasing, tokenization, stopword removal, stemming, and lemmatization.

```
import nltk
import spacy
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk import WordNetLemmatizer
```

```
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('punkt_tab')
```

```
[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...
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
True
```

```
# Sample text with multiple sentences
text = """Natural Language Processing tasks include tokenization, stemming, and lemmatization.
It is a field of Artificial Intelligence. Natural Language Processing is used in chatbots, sentiment analysis, and machine t
print(text)
```

```
Natural Language Processing tasks include tokenization, stemming, and lemmatization.
It is a field of Artificial Intelligence. Natural Language Processing is used in chatbots, sentiment analysis, and machi
```

```
# 1. Lowercasing
text_lower = text.lower()
print("Lowercased Text:", text_lower)
```

```
Lowercased Text: natural language processing tasks include tokenization, stemming, and lemmatization.
it is a field of artificial intelligence. natural language processing is used in chatbots, sentiment analysis, and machi
```

```
# 2. Sentence Tokenization using NLTK
sentences = sent_tokenize(text)
print("Sentence Tokenization using NLTK:")
print("Sentences:", sentences)
```

```
Sentence Tokenization using NLTK:
Sentences: ['Natural Language Processing tasks include tokenization, stemming, and lemmatization.', 'It is a field of Ar
```

```
# 3. Word Tokenization
tokens = word_tokenize(text)
print("Tokens:", tokens)
```

```
Tokens: ['Natural', 'Language', 'Processing', 'tasks', 'include', 'tokenization', ',', 'stemming', ',', 'and', 'lemmatiz
```

```
# 4. Stopword Removal
stop_words = set(stopwords.words('english'))
print("All standard Stop words:", stop_words)
print("Number of standard Stop words:", len(stop_words))
filtered_tokens = [word for word in tokens if word not in stop_words]
print("Tokens after Stopword Removal:", filtered_tokens)
```

```
All standard Stop words: {'she', 'own', 'which', 'what', 'does', 'by', 'most', 'am', 'them', 'yours', 'down', 'm', 'out'
Number of standard Stop words: 198
Tokens after Stopword Removal: ['Natural', 'Language', 'Processing', 'tasks', 'include', 'tokenization', ',', 'stemming'
```

```
# 5. Stemming
ps = PorterStemmer()
stemmed_tokens = [ps.stem(word) for word in filtered_tokens]
print("Stemmed Tokens:", stemmed_tokens)
```

```
Stemmed Tokens: ['natur', 'languag', 'process', 'task', 'includ', 'token', ',', 'stem', ',', 'lemmat', '.', 'it', 'field
```

```
# 6. Lemmatization
lemmatizer = WordNetLemmatizer()
lemmas = [lemmatizer.lemmatize(t) for t in tokens]
print("Lemmatized Tokens:", lemmas)
```

```
Lemmatized Tokens: ['Natural', 'Language', 'Processing', 'task', 'include', 'tokenization', ',', 'stemming', ',', 'and',
```