

TITLE : Text Preprocessing - Implement text preprocessing techniques, including tokenization and normalization, and apply it to any text data.

NAME : Shinde Shubham Dnyandev,

ROLL NO : 23107121,

BATCH : B.

```
In [1]: import nltk
import re
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, SnowballStemmer, WordNetLemmatizer
```

```
In [3]: nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to /home/admin1/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /home/admin1/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /home/admin1/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

```
Out[3]: True
```

```
In [5]: text_data = "Hello! This is an example sentence, demonstrating text preprocessing in nlp."
```

```
In [7]: text_data = text_data.lower()
print("Lowercase Text : \n\n", text_data)
```

Lowercase Text :

hello! this is an example sentence, demonstrating text preprocessing in nlp. running is a good exercise.

```
In [9]: text_data = re.sub(r'[^a-zA-Z0-9\s]', '', text_data)
print("Text Without Special Characters : \n\n", text_data)
```

Text Without Special Characters :

hello this is an example sentence demonstrating text preprocessing in nlp running is a good exercise

```
In [11]: text_data = re.sub(r'\s+', ' ', text_data).strip()
print("Text Without Extra Spaces : \n\n", text_data)
```

Text Without Extra Spaces :

hello this is an example sentence demonstrating text preprocessing in nlp running is a good exercise

```
In [13]: tokens = word_tokenize(text_data)
print(f"Tokens : \n\n {tokens}")
```

Tokens :

```
['hello', 'this', 'is', 'an', 'example', 'sentence', 'demonstrating', 'text', 'preprocessing', 'in', 'nlp', 'running', 'is', 'a', 'good', 'exercise']
```

```
In [15]: stop_words = set(stopwords.words('english'))
tokens = [token
          for token in tokens]
```

```
    if token not in stop_words]:
print(tokens)

['hello', 'example', 'sentence', 'demonstrating', 'text', 'preprocessing', 'nlp', 'running', 'good', 'exercise']
```

```
In [17]: PR_stemmer = PorterStemmer()
tokens_stem = [PR_stemmer.stem(token)
               for token in tokens]
print(tokens_stem)
```

```
['hello', 'exampl', 'sentenc', 'demonstr', 'text', 'preprocess', 'nlp', 'run', 'good',
'exercis']
```

```
In [19]: SB_stemmer = SnowballStemmer('english')
tokens_stem = [SB_stemmer.stem(token)
               for token in tokens]
print(tokens_stem)
```

```
['hello', 'exampl', 'sentenc', 'demonstr', 'text', 'preprocess', 'nlp', 'run', 'good',
'exercis']
```

```
In [21]: lemmatizer = WordNetLemmatizer()
tokens_lem = [lemmatizer.lemmatize(token)
              for token in tokens]
print(tokens_lem)
```

```
['hello', 'example', 'sentence', 'demonstrating', 'text', 'preprocessing', 'nlp', 'running',
'good', 'exercise']
```

```
In [ ]:
```