[ ]:

```
Practical No.: 7

Contents for Theory:
1. Basic concepts of Text Analytics
2. Text Analysis Operations using natural language
toolkit
3. Text Analysis Model using TF-IDF.
4. Bag of Words (BoW)
```

[1]:

```
import nltk
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\SSOS21\AppData\Roaming\nltk_data…
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\SSOS21\AppData\Roaming\nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\SSOS21\AppData\Roaming\nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\SSOS21\AppData\Roaming\nltk_data…
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

[1]:

```
True
```

[2]:

```
text = """Tokenization is the first step in text analytics. The
process of breaking down a text paragraph into smaller
chunks such as words or sentences is called Tokenization."""
```

[3]:

```
from nltk.tokenize import sent_tokenize

# Sentence Tokenization
```

```
tokenized_text = sent_tokenize(text)
print("Sentence Tokenization Output:")
print(tokenized_text)
```

```
Sentence Tokenization Output:
['Tokenization is the first step in text analytics.', 'The\nprocess of breaking
down a text paragraph into smaller\nchunks such as words or sentences is called
Tokenization.']
```

[4]:

```
from nltk.tokenize import word_tokenize
tokenized_word = word_tokenize(text)
print("\nWord Tokenization Output:")
print(tokenized_word)
```

```

Word Tokenization Output:
['Tokenization', 'is', 'the', 'first', 'step', 'in', 'text', 'analytics', '.',
'The', 'process', 'of', 'breaking', 'down', 'a', 'text', 'paragraph', 'into',
'smaller', 'chunks', 'such', 'as', 'words', 'or', 'sentences', 'is', 'called',
'Tokenization', '.']
```

[5]:

```
import string
tokens_without_punctuations = [word for word in tokenized_word if word not in␣
 ↪string.punctuation]
print("\nTokens after Removing Punctuation:")
print(tokens_without_punctuations)
```

```

Tokens after Removing Punctuation:
['Tokenization', 'is', 'the', 'first', 'step', 'in', 'text', 'analytics', 'The',
'process', 'of', 'breaking', 'down', 'a', 'text', 'paragraph', 'into',
'smaller', 'chunks', 'such', 'as', 'words', 'or', 'sentences', 'is', 'called',
'Tokenization']
```

[6]:

```
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
filtered_tokens = [word for word in tokens_without_punctuations if word.lower()␣
 ↪not in stop_words]
print("\nTokens after Removing Stop Words:")
print(filtered_tokens)
```

```

Tokens after Removing Stop Words:
['Tokenization', 'first', 'step', 'text', 'analytics', 'process', 'breaking',
'text', 'paragraph', 'smaller', 'chunks', 'words', 'sentences', 'called',
'Tokenization']
```

[7]:

```
pos_tags = nltk.pos_tag(filtered_tokens)
print("\nPOS Tagging Output:")
print(pos_tags)
```

```
POS Tagging Output:
[('Tokenization', 'NN'), ('first', 'RB'), ('step', 'VB'), ('text', 'JJ'),
('analytics', 'NNS'), ('process', 'NN'), ('breaking', 'VBG'), ('text', 'NN'),
('paragraph', 'NN'), ('smaller', 'JJR'), ('chunks', 'NNS'), ('words', 'NNS'),
('sentences', 'NNS'), ('called', 'VBD'), ('Tokenization', 'NN')]
```

[8]:

```
from nltk.stem import PorterStemmer
stemmer = PorterStemmer()
stemmed_words = [stemmer.stem(word) for word in filtered_tokens]
print("\nStemmed Words:")
print(stemmed_words)
```

```
Stemmed Words:
['token', 'first', 'step', 'text', 'analyt', 'process', 'break', 'text',
'paragraph', 'smaller', 'chunk', 'word', 'sentenc', 'call', 'token']
```

[9]:

```
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
lemmatized_words = [lemmatizer.lemmatize(word) for word in filtered_tokens]
print("\nLemmatized Words:")
print(lemmatized_words)
```

```
Lemmatized Words:
['Tokenization', 'first', 'step', 'text', 'analytics', 'process', 'breaking',
'text', 'paragraph', 'smaller', 'chunk', 'word', 'sentence', 'called',
'Tokenization']
```

[ ]:

```
Name: Mansi Nirbhavane
        Roll No.:13251
```