Experiment No.3
Apply various other text preprocessing techniques for any given text: Stop Word Removal, Lemmatization / Stemming.
Date of Performance:
Date of Submission:

Aim: Apply various other text preprocessing techniques for any given text: Stop Word Removal, Lemmatization / Stemming.

Objective: To write a program for Stop word removal from a sentence given in English and any Indian Language.

Theory:

The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

Stopwords are the most common words in any natural language. For the purpose of analyzing text data and building NLP models, these stopwords might not add much value to the meaning of the document.

Stop Words: A stop word is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We need to perform tokenization before removing any stopwords.

Why do we need to Remove Stopwords?

Removing stopwords is not a hard and fast rule in NLP. It depends upon the task that we are working on. For tasks like text classification, where the text is to be classified into different categories, stopwords are removed or excluded from the given text so that more focus can be given to those words which define the meaning of the text.

Here are a few key benefits of removing stopwords:

- On removing stopwords, dataset size decreases and the time to train the model also decreases
- Removing stopwords can potentially help improve the performance as there are fewer and only meaningful tokens left. Thus, it could increase classification accuracy
- Even search engines like Google remove stopwords for fast and relevant retrieval of data from the database

We can remove stopwords while performing the following tasks:

- Text Classification
    - Spam Filtering

  - Language Classification
  - Genre Classification
- Caption Generation
- Auto-Tag Generation

Avoid Stopword Removal

- Machine Translation
- Language Modeling
- Text Summarization
- Question-Answering problems

Different Methods to Remove Stopwords

1. **Stopword Removal using NLTK**

   NLTK, or the Natural Language Toolkit, is a treasure trove of a library for text preprocessing. It's one of my favorite Python libraries. NLTK has a list of stopwords stored in 16 different languages.

   You can use the below code to see the list of stopwords in NLTK:

   ```
   import nltk
   from nltk.corpus import stopwords
   set(stopwords.words('english')
   ```

2. **Stopword Removal using spaCy:**

   spaCy is one of the most versatile and widely used libraries in NLP. We can quickly and efficiently remove stopwords from the given text using SpaCy.

   It has a list of its own stopwords that can be imported as **STOP_WORDS** from the **spacy.lang.en.stop_words** class.

3. **Stopword Removal using Gensim**

   Gensim is a pretty handy library to work with on NLP tasks. While pre-processing, gensim provides methods to remove stopwords as well. We can easily import the remove_stopwords method from the class gensim.parsing.preprocessing.

Code:

```
### Importing Required Libraries
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer, SnowballStemmer, LancasterStemmer, WordNetLemmatizer
# Download necessary NLTK resources (uncomment if you haven't downloaded them yet)
# nltk.download('punkt')
# nltk.download('stopwords')
# nltk.download('wordnet')
### Text
text = 'TON 618 is a hyperluminous, broad-absorption-line, radio-loud quasar and Lyman-alpha blob located near the border of the constellations Canes Venatici and Coma Berenices, with the projected comoving distance of approximately 18.2 billion light-years from Earth.'
print("Original Text: ", text)
### Stopwords
stop_words = set(stopwords.words('english'))  # Convert to set for faster lookups
words = word_tokenize(text)
#### Applying stop words
holder = []
for w in words:
    if w not in stop_words:  # Check if the word is not a stop word
        holder.append(w)
print("Filtered Words (Stopwords Removed): ", holder)
#### List Comprehension for stop words
holder_comprehension = [w for w in words if w not in stop_words]
print("Filtered Words using List Comprehension: ", holder_comprehension)
### Stemming
porter = PorterStemmer()
snow = SnowballStemmer(language='english')
lancaster = LancasterStemmer()
words_list = ['play', 'plays', 'played', 'playing', 'player']
#### Porter Stemmer
porter_stemmed = []
for w in words_list:
    stemmed_word = porter.stem(w)
    porter_stemmed.append(stemmed_word)
print("Porter Stemmed Words: ", porter_stemmed)
#### Porter Stemmer List Comprehension
porter_stemmed_comprehension = [porter.stem(x) for x in words_list]
print("Porter Stemmed Words using List Comprehension: ", porter_stemmed_comprehension)
#### Snowball Stemmer
snow_stemmed = []
for w in words_list:
    stemmed_word = snow.stem(w)
    snow_stemmed.append(stemmed_word)
```

```
print("Snowball Stemmed Words: ", snow_stemmed)
#### Snowball Stemmer List Comprehension
snow_stemmed_comprehension = [snow.stem(x) for x in words_list]
print("Snowball Stemmed Words using List Comprehension: ", snow_stemmed_comprehension)
#### Lancaster Stemmer
lancaster_stemmed = []
for w in words_list:
    stemmed_word = lancaster.stem(w)
    lancaster_stemmed.append(stemmed_word)
print("Lancaster Stemmed Words: ", lancaster_stemmed)
#### Lancaster Stemmer List Comprehension
lancaster_stemmed_comprehension = [lancaster.stem(x) for x in words_list]
print("Lancaster Stemmed Words using List Comprehension: ", lancaster_stemmed_comprehension)
### Lemmatization: This has a more expansive vocabulary than Stemming
wordnet = WordNetLemmatizer()
lemmatized = [wordnet.lemmatize(x) for x in words_list]
print("Lemmatized Words: ", lemmatized)
```

Output:

```
(venv) PS D:\Vartak college\sem 7\NLP\EXP\New folder> python .\exp3.py
Original Text:  TON 618 is a hyperluminous, broad-absorption-line, radio-loud quasar and Lyman-alpha blob located near the border of the constellations Canes Venatici and Coma Berenices, with the projected comoving distance of approximately 18.2 billion light-years from Earth.
Filtered Words (Stopwords Removed):  ['TON', '618', 'hyperluminous', ',', 'broad-absorption-line', ',', 'radio-loud', 'quasar', 'Lyman-alpha', 'blob', 'located', 'near', 'border', 'constellations', 'Canes', 'Venatici', 'Coma', 'Berenices', ',', 'projected', 'comoving', 'distance', 'approximately', '18.2', 'billion', 'light-years', 'Earth', '.']
Filtered Words using List Comprehension:  ['TON', '618', 'hyperluminous', ',', 'broad-absorption-line', ',', 'radio-loud', 'quasar', 'Lyman-alpha', 'blob', 'located', 'near', 'border', 'constellations', 'Canes', 'Venatici', 'Coma', 'Berenices', ',', 'projected', 'comoving', 'distance', 'approximately', '18.2', 'billion', 'light-years', 'Earth', '.']
Porter Stemmed Words:  ['play', 'play', 'play', 'play', 'player']
Porter Stemmed Words using List Comprehension:  ['play', 'play', 'play', 'play', 'player']
Snowball Stemmed Words:  ['play', 'play', 'play', 'play', 'player']
Snowball Stemmed Words using List Comprehension:  ['play', 'play', 'play', 'play', 'player']
Lancaster Stemmed Words:  ['play', 'play', 'play', 'play', 'play']
Lancaster Stemmed Words using List Comprehension:  ['play', 'play', 'play', 'play', 'play']
Lemmatized Words:  ['play', 'play', 'played', 'playing', 'player']
```

Conclusion:

Stop word removal is an essential preprocessing step in natural language processing (NLP) that involves eliminating common words (e.g., "the," "is," "and") that carry little meaning in a given context. For Indian languages, the process can be more challenging due to the diversity and complexity of languages. Here are some commonly used tools and methods for stop word removal in Indian languages, along with the steps involved:

Tools for Stop Word Removal in Indian Languages:

1. NLTK: While primarily focused on English, NLTK offers stop word lists for some Indian languages and can be customized with user-defined stop word lists.
2. spaCy: This library supports multiple languages and can be extended with custom stop word lists for Indian languages.
4. Indic NLP Library: A dedicated library for Indian languages that includes functionalities for tokenization, stop word removal, and more, specifically tailored for languages like Hindi, Bengali, Tamil, etc.
5. Gensim: A Python library for topic modeling that includes options for removing stop words. It can be customized with language-specific stop word lists.
6. Custom Lists: Creating tailored stop word lists for specific Indian languages based on domain requirements can significantly improve the effectiveness of stop word removal.

Steps Involved in Stop Word Removal:

1. Text Input: Start with the raw text data in the target Indian language.
2. Tokenization: Break down the text into individual tokens (words) using a suitable tokenizer that can handle the complexities of the language.
3. Stop Word List: Use an existing stop word list specific to the target language, or create a custom list based on the context of your application.
4. Filtering Tokens: Iterate through the tokenized words and remove those that are present in the stop word list. This can be done using set operations for efficiency.
5. Output Processed Text: Collect the filtered tokens back into a single string or a list, which can then be used for further NLP tasks, such as text classification or sentiment analysis.