NLP Experiment 4

Aim: To implement Advanced Text Pre-processing Techniques.

Theory:

We first remove stopwords. Stopwords are frequently occurring words that carry little to no additional information for analysing the meaning of the sentence.

Steps for removal of stopwords:

- We download a list of stopwords using nltk.download("stopwords")
- NLTK provides a list of stop words in multiple languages (we download that in the step above). We use NLTK to remove these stopwords from text data, allowing one to focus on the more meaningful words and phrases when performing text analysis, such as text classification or sentiment analysis.
- Split every sentence into words (splitting by space).
- Form a new sentence, if a word is present in the list of stopwords formed earlier we do not add that word back in the sentence.
- We form a new sentence by eliminating all stopwords using this.
-

Lemmatization is the method of reducing a word into its dictionary form (these reduced words are known as lemma) that is normalizing words to their root words. This practice makes it easier to analyse sentences.

- We perform lemmatization by using "wordnet" in nltk
- WordNet is a lexical database and resource for natural language processing and linguistic research. It's an extensive lexical database of English, developed at Princeton University. WordNet is organized around the concept of a "lexicon," which is essentially a comprehensive dictionary of English words and their relationships.
- In NLTK (Natural Language Toolkit), WordNetLemmatizer is a class that provides lemmatization functionality based on WordNet. We use this class to lemmatize words in a sentence. (As used in the function lemmatize_words(text:str))

Libraries and Tools Used:

- Pandas
- nltk