PRACTICAL NO 6

**AIM:** Illustrate utility of NLTK: Write a program to remove stop words using NLTK.

**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.

What are Stop words?

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 would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK(Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages. You can find them in the nltk\_data directory. home/pratima/nltk\_data/corpora/stopwords is the directory address.(Do not forget to change your home directory name).

**CODE:**

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

example\_sent = """ 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. """

stop\_words = set(stopwords.words('english'))

word\_tokens = word\_tokenize(example\_sent)

# converts the words in word\_tokens to lower case and then checks whether

#they are present in stop\_words or not

filtered\_sentence = [w for w in word\_tokens if not w.lower() in stop\_words]

#with no lower case conversion

filtered\_sentence = []

for w in word\_tokens:

if w not in stop\_words:

filtered\_sentence.append(w)

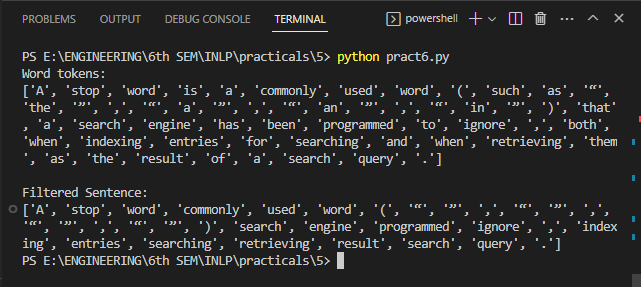
print("Word tokens: ")

print(word\_tokens)

print("\nFiltered Sentence:")

print(filtered\_sentence)

**OUTPUT:**



**CONCLUSION:** Hence we successfully executed the practical on the removal of stop words using NLTK.