1. Write a Python code using NLP to Pre-Process the text data and convert Text-Numeric vectors.

I. Use Tokenization, Stopword removal, Stemming/Lemmatization , text preprocess logic using NLTK

In [8]:
```
import os
os.chdir(r'D:')
```

In [6]:
```
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer

# Download the NLTK resources
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')

# Load the text data
with open('novel.txt', 'r', encoding='utf8') as file:
    text_data = file.read()

# Tokenize the text
tokens = word_tokenize(text_data)

# Remove stopwords
stop_words = set(stopwords.words('english'))
filtered_tokens = [token for token in tokens if token.lower() not in stop_words]

# Lemmatize the words
lemmatizer = WordNetLemmatizer()
lemmatized_tokens = [lemmatizer.lemmatize(token) for token in filtered_tokens]

# Join the tokens back into a single string
preprocessed_text = ' '.join(lemmatized_tokens)

# Convert the preprocessed text into text-numeric vectors
vectorizer = TfidfVectorizer()
text_numeric_vectors = vectorizer.fit_transform([preprocessed_text])

# Print the text-numeric vectors
print(text_numeric_vectors.toarray())
```

```
[nltk_data] Downloading package punkt to C:\Users\G BHAVANI
[nltk_data]     SHANKAR\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to C:\Users\G BHAVANI
[nltk_data]     SHANKAR\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to C:\Users\G BHAVANI
[nltk_data]     SHANKAR\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[[0.00162125 0.00162125 0.00162125 ... 0.00162125 0.00486374 0.00162125]]
```

In [5]:
```
import nltk
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package omw-1.4 to C:\Users\G BHAVANI
[nltk_data]     SHANKAR\AppData\Roaming\nltk_data...
```

Out[5]:
```
True
```

Use SKLearn for converting Text-Numeric vectors using TF-IDF model

In [10]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

# Sample input text
text = [
    "This is the first document.",
    "This document is the second document.",
    "And this is the third one.",
    "Is this the first document?",
]

# Create the TF-IDF vectorizer object
vectorizer = TfidfVectorizer()

# Fit the vectorizer to the input text and transform the text into numerical vecto
vectorized_text = vectorizer.fit_transform(text)

# Print the vectorized text
print(vectorized_text.toarray())
```

```
[[0.         0.46979139 0.58028582 0.38408524 0.         0.
  0.38408524 0.         0.38408524]
 [0.         0.6876236  0.         0.28108867 0.         0.53864762
  0.28108867 0.         0.28108867]
 [0.51184851 0.         0.         0.26710379 0.51184851 0.
  0.26710379 0.51184851 0.26710379]
 [0.         0.46979139 0.58028582 0.38408524 0.         0.
  0.38408524 0.         0.38408524]]
```

In [ ]:
```
```