[Day-7 2211cs020196]Write a Python script that: 1. Use Genism to preprocess data from a sample text file, follow basic procedures like tokenization, stemming, lemmatization etc. ¶

In [3]:

```python
!pip install gensim nltk spacy
import re
import gensim
from nltk.stem.porter import PorterStemmer
from nltk.corpus import stopwords
import spacy
import nltk
nltk.download('stopwords')
nlp = spacy.load("en_core_web_sm")
porter_stemmer = PorterStemmer()
stop_words = set(stopwords.words('english'))
def preprocess_text(text):
    text = re.sub(r'[^\w\s]', '', text.lower())
    tokens = [word for word in gensim.utils.simple_preprocess(text) if wor
    stemmed_tokens = [porter_stemmer.stem(token) for token in tokens]
    doc = nlp(' '.join(stemmed_tokens))
    lemmatized_tokens = [token.lemma_ for token in doc]
    return lemmatized_tokens
text_content = """
Write a Python script that uses Gensim to preprocess data from a sample te
file. Follow basic procedures like tokenization, stemming, and lemmatizati
Print the final output to verify the preprocessing steps.
"""
processed_text = preprocess_text(text_content)
print(processed_text)

```

```
[nltk_data] Downloading package stopwords to C:\Users\Sai Krishna
[nltk_data]     Hari\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
['write', 'python', 'script', 'use', 'gensim', 'preprocess', 'data', 'sampl',
'text', 'file', 'follow', 'basic', 'procedur', 'like', 'token', 'stem', 'lemm
at', 'print', 'final', 'output', 'verifi', 'preprocess', 'step']
```

In [ ]:

```
```