**Tokenization:-**

**Tokenization is the process of removing sensitive data and placing unique symbols of identification in its place to retain all the essential information.**

**CODE:-**

import nltk

sentence\_data = "The First sentence is about Python. The Second: about Django. You can learn Python,Django and Data Ananlysis here. "

nltk\_tokens = nltk.sent\_tokenize(sentence\_data)

print (nltk\_tokens)

**Stop-word Removal:-**

Stop — words are natural language words which have very little meaning such as ‘a’, ‘an’, ‘and’, ‘or’, ‘the’.  
⁃ These words take up space in a database and increase the processing time.  
⁃ They can be removed by storing an of stop-words.  
⁃ Stop-words are filtered out before processing of natural language data as they don’t reveal much information.

**Code:-**

import nltk

from nltk.corpus import stopwords

print(stopwords.words('english'))

**Stemming:-**

Stemming involves reducing the word “Stem” or base (root) from removing the suffix.

from nltk.stem import PorterStemmer

from nltk.tokenize import sent\_tokenize, word\_tokenize

ps = PorterStemmer( )

text\_example = “your text goes here”

words = word\_tokenize (text\_example)

for w in words :

print(ps.stem(w))

**Lemmatization:-**

**Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words.**

import nltk

from nltk.stem import WordNetLemmatizer

wordnet\_lemmatizer = WordNetLemmatizer()

text = "studies studying cries cry"

tokenization = nltk.word\_tokenize(text)

for w in tokenization:

print("Lemma for {} is {}".format(w, wordnet\_lemmatizer.lemmatize(w)))

**POS Tagging:-**

**POS tagging is a task of labelling each word in a sentence with its appropriate part of speech.**

**Code:-**

import nltk

from nltk import word\_tokenize

sentence = "I am going to school"

print (nltk.pos\_tag(word\_tokenize(sentence)))