**Practical No : 12**

**Problem Statement:**

Implement the following tasks for NLP on the data mentioned in the document below. Also try the same steps for your own data. Use the NLTK library

**Link 1:** <https://www.turing.com/kb/natural-language-processing-function-in-ai>

**1. Segmenting and Tokenizing: Tokenize**

import nltk

nltk.download('punkt')

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

data="""Hello my name is Harsh Gharsandiya. My Enrollment number is ET22BTCO037.

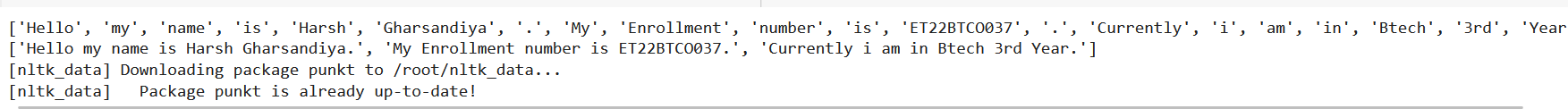
Currently i am in Btech 3rd Year."""

wt = nltk.word\_tokenize(data)

print(wt)

st = nltk.sent\_tokenize(data)

print(st)



**2. Stop Words**

import nltk

nltk.download('stopwords')

from nltk.corpus import stopwords

len(stopwords.words('english'))

stopwords.words('english')

sw = set(stopwords.words('english'))

data1 = wt

[word for word in data1 if word not in sw]

**Link2:** <https://realpython.com/nltk-nlp-python/#tokenizing>

**3. Stemming**

**//**Example 1

from nltk.stem import PorterStemmer

pst = PorterStemmer()

print(pst.stem("Harsh Gharsandiya Achievement"))



//Example 2

from nltk.stem import LancasterStemmer

lan = LancasterStemmer()

print(lan.stem("Harsh Gharsandiya Achivement"))



**4. Lemmatization**

import nltk

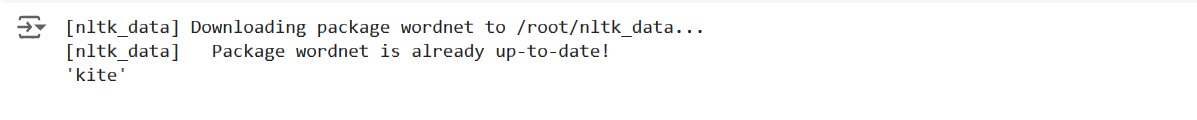
nltk.download('wordnet')

from nltk.stem import wordnet

from nltk.stem import WordNetLemmatizer

word\_lem=WordNetLemmatizer()

word\_lem.lemmatize("kites")

****

**5. POS (Parts of Speech Tagging) / Speech Tagging**

import nltk

nltk.download('punkt')

sent = "I am a Harsh doing nothing"

token=nltk.word\_tokenize(sent)

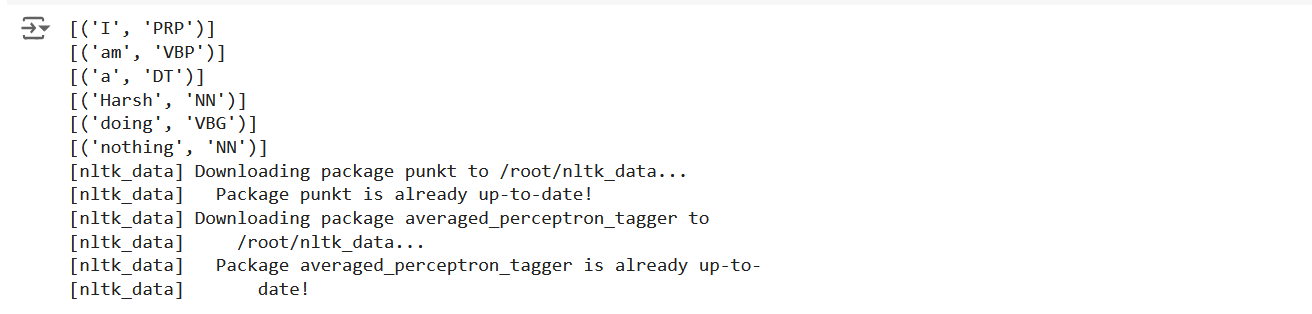
token

import nltk

nltk.download('averaged\_perceptron\_tagger')

for tokens in token:

print(nltk.pos\_tag([tokens]))

****

**Own data:**

**Code:**

import nltk

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

from nltk.corpus import stopwords

from nltk.stem import PorterStemmer, WordNetLemmatizer

from nltk import pos\_tag

# Download necessary resources

nltk.download('punkt')

nltk.download('stopwords')

nltk.download('averaged\_perceptron\_tagger')

nltk.download('wordnet')

# Your custom data (you can replace this with any text)

your\_data = "Hi I am coder. I do nothing. Since i become coder i do nothing. btw my name is harsh gharsandiya"

# Step 1: Tokenization

# Word tokenization

words = word\_tokenize(your\_data)

print("Word Tokens:", words)

# Sentence tokenization

sentences = sent\_tokenize(your\_data)

print("Sentence Tokens:", sentences)

# Step 2: Stopword Removal

# Load stop words

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

# Remove stop words with a check for string values

filtered\_words = [word for word in words if isinstance(word, str) and word.lower() not in stop\_words]

print("Filtered Words (without stopwords):", filtered\_words)

# Step 3: Stemming

# Initialize stemmer

stemmer = PorterStemmer()

# Stem each word

stemmed\_words = [stemmer.stem(word) for word in filtered\_words]

print("Stemmed Words:", stemmed\_words)

# Step 4: Lemmatization

# Initialize lemmatizer

lemmatizer = WordNetLemmatizer()

# Lemmatize each word

lemmatized\_words = [lemmatizer.lemmatize(word) for word in filtered\_words]

print("Lemmatized Words:", lemmatized\_words)

# Step 5: Part-of-Speech Tagging (POS)

# Perform POS tagging

pos\_tags = pos\_tag(filtered\_words)

print("POS Tags:", pos\_tags)

**Output:**

