**Name: Vishnu Mohan**

**Roll No:51**

**Batch: MCA-B**

**Date: 24-11-2022**

**DATA SCIENCE LAB**

**Experiment No.: 15**

**Aim**

Data preprocessing with NLTK

1. Counting Tags
2. Bigrams
3. Trigrams
4. Stop Words
5. Stemming

**Procedure**

!pip install -q wordcloud

import wordcloud

import nltk

nltk.download('stopwords')

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

import pandas as pd

import unicodedata

import numpy as np

import string

1. from collections import Counter

import nltk

text = "Guru99 is one of the best sites to learn WEB, SAP, Ethical Hacking and much more online."

lower\_case = text.lower()

tokens = nltk.word\_tokenize(lower\_case)

tags = nltk.pos\_tag(tokens)

counts = Counter( tag for word,  tag in tags)

print(counts)

Output

Counter({'NN': 5, ',': 2, 'VBZ': 1, 'CD': 1, 'IN': 1, 'DT': 1, 'JJS': 1, 'NNS': 1, 'TO': 1, 'VB': 1, 'JJ': 1, 'CC': 1, 'RB': 1, 'JJR': 1, '.': 1})

1. import nltk

text = "Guru99 is a totally new kind of learning experience."

Tokens = nltk.word\_tokenize(text)

output = list(nltk.bigrams(Tokens))

print(output)

Output

[('Guru99', 'is', 'a'), ('is', 'a', 'totally'), ('a', 'totally', 'new'), ('totally', 'new', 'kind'), ('new', 'kind', 'of'), ('kind', 'of', 'learning'), ('of', 'learning', 'experience'), ('learning', 'experience', '.')]

1. import nltk

text = "Guru99 is a totally new kind of learning experience."

Tokens = nltk.word\_tokenize(text)

output = list(nltk.trigrams(Tokens)) print(output)

Output

[('Guru99', 'is', 'a'), ('is', 'a', 'totally'), ('a', 'totally', 'new'), ('totally', 'new', 'kind'), ('new', 'kind', 'of'), ('kind', 'of', 'learning'), ('of', 'learning', 'experience'), ('learning', 'experience', '.')]

1. from nltk.corpus import stopwords

print(stopwords.words('english'))

en\_stopwords = stopwords.words('english')

def remove\_stopwords(text):

    result = []

    for token in text:

        if token not in en\_stopwords:

            result.append(token)

    return result

text = "this is the only solution of that question".split()  remove\_stopwords(text)

Output

['solution', 'question']

1. from nltk.stem import PorterStemmer

from nltk.tokenize import word\_tokenize

ps = PorterStemmer()

sentence = "Programmers program with programming languages"

words = word\_tokenize(sentence)

for w in words:

    print(w, " : ", ps.stem(w))

Output

Programmers : programm

program : program

with : with

programming : program

languages : languag

**Result**

The program was executed and the result was successfully obtained. Thus CO5 was obtained.