import nltk

import re

import numpy as np

myfile = open(r"C:\Users\Aniket\Desktop\CCNLP\1\test.txt")

text = myfile.read()

dataset = nltk.sent\_tokenize(text)

for i in range(len(dataset)):

dataset[i] = dataset[i].lower()

dataset[i] = re.sub(r'\W', ' ', dataset[i])

dataset[i] = re.sub(r'\s+', ' ', dataset[i])

print(dataset)

print('\n')

word2count = {}

for data in dataset:

words = nltk.word\_tokenize(data)

for word in words:

if word not in word2count.keys():

word2count[word] = 1

else:

word2count[word] += 1

print(word2count)

print('\n')

import heapq

freq\_words = heapq.nlargest(100, word2count, key=word2count.get)

print(freq\_words)

print('\n')

X = []

for data in dataset:

vector = []

for word in freq\_words:

if word in nltk.word\_tokenize(data):

vector.append(1)

else:

vector.append(0)

X.append(vector)

X = np.asarray(X)

print(X)

print('\n')

//////////////////////////////////////////////////////////////////////////////////////////////////////////

Output:-

runfile('C:/Users/Aniket/.spyder-py3/temp.py', wdir='C:/Users/Aniket/.spyder-py3')

['birsa munda about this soundpronunciation was an indian tribal freedom fighter religious leader and folk hero who belonged to the munda tribe ', 'he spearheaded a tribal religious millenarian movement that arose in the bengal presidency now jharkhand in the late 19th century during the british raj thereby making him an important figure in the history of the indian independence movement ', 'the revolt mainly concentrated in the munda belt of khunti tamar sarwada and bandgaon ', 'his portrait hangs in the indian parliament museum he is the only tribal leader to have been so honored ']

{'birsa': 1, 'munda': 3, 'about': 1, 'this': 1, 'soundpronunciation': 1, 'was': 1, 'an': 2, 'indian': 3, 'tribal': 3, 'freedom': 1, 'fighter': 1, 'religious': 2, 'leader': 2, 'and': 2, 'folk': 1, 'hero': 1, 'who': 1, 'belonged': 1, 'to': 2, 'the': 10, 'tribe': 1, 'he': 2, 'spearheaded': 1, 'a': 1, 'millenarian': 1, 'movement': 2, 'that': 1, 'arose': 1, 'in': 5, 'bengal': 1, 'presidency': 1, 'now': 1, 'jharkhand': 1, 'late': 1, '19th': 1, 'century': 1, 'during': 1, 'british': 1, 'raj': 1, 'thereby': 1, 'making': 1, 'him': 1, 'important': 1, 'figure': 1, 'history': 1, 'of': 2, 'independence': 1, 'revolt': 1, 'mainly': 1, 'concentrated': 1, 'belt': 1, 'khunti': 1, 'tamar': 1, 'sarwada': 1, 'bandgaon': 1, 'his': 1, 'portrait': 1, 'hangs': 1, 'parliament': 1, 'museum': 1, 'is': 1, 'only': 1, 'have': 1, 'been': 1, 'so': 1, 'honored': 1}

['the', 'in', 'munda', 'indian', 'tribal', 'an', 'religious', 'leader', 'and', 'to', 'he', 'movement', 'of', 'birsa', 'about', 'this', 'soundpronunciation', 'was', 'freedom', 'fighter', 'folk', 'hero', 'who', 'belonged', 'tribe', 'spearheaded', 'a', 'millenarian', 'that', 'arose', 'bengal', 'presidency', 'now', 'jharkhand', 'late', '19th', 'century', 'during', 'british', 'raj', 'thereby', 'making', 'him', 'important', 'figure', 'history', 'independence', 'revolt', 'mainly', 'concentrated', 'belt', 'khunti', 'tamar', 'sarwada', 'bandgaon', 'his', 'portrait', 'hangs', 'parliament', 'museum', 'is', 'only', 'have', 'been', 'so', 'honored']

[[1 0 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

[1 1 0 1 1 1 1 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

[1 1 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0]

[1 1 0 1 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1]]