Write a program to remove the stopwords for any given paragraph. Create a set of stopwords given below and print the output.

stop\_words = [‘.’,’,’,’a’,’they’,’the’,’his’,’so’,’and’,’were’,’from’,that’,’of’,’in’,’only’,’with’,’to’]

PROCEDURE

Filereader = open(File, read)

Text[]= []

Stop\_words[]=['.',',','a','they','the','his','so','and','were','from','that','of','in','only','with','to']

Textarray[] = Filereader.read().split()

For i in (0 to lengthof(Textarray))

If TextArray[i] is not in Stop\_words

Text.append(TextArray[i])

Print (Text)

CODE

fr = open("../SampleText.txt")

stop\_words = ['.',',','a','they','the','his','so','and','were','from','that','of','in','only','with','to']

text\_arr = fr.read().split()

text\_without\_sw=[]

for i in range(0,len(text\_arr)):

if text\_arr[i] not in stop\_words:

text\_without\_sw.append(text\_arr[i])

print (text\_without\_sw)

fw = open("../TextWithoutStopwords.txt","w")

for j in range(0,len(text\_without\_sw)):

fw.write(text\_without\_sw[j])

fw.write(" ")

fw.close()

print(fr.read())

fr.close()

OUTPUT

Write a program to tokenize (without Nltk)

A sentence

PROCEDURE

FileReader = open (File, read)

Tokens= FileReader.read().split()

Print(Tokens)

CODE

fr = open("../TextWithoutStopwords.txt")

TokenList = fr.read().split()

print(TokenList)

OUTPUT

Multiple Sentences

PROCEDURE

FileReader = open (File, read)

Sentences= FileReader.read().split(‘ . ’)

Print(Sentences)

For i in (0 to lengthof(Sentences))

Arr[] = Sentences[i].split()

Text.append(Arr)

Print (Text)

CODE

fr = open("../SampleText.txt")

sentences =fr.read().split(' . ')

token =[]

print("Tokenization of Sentence: ",sentences,"\n\n")

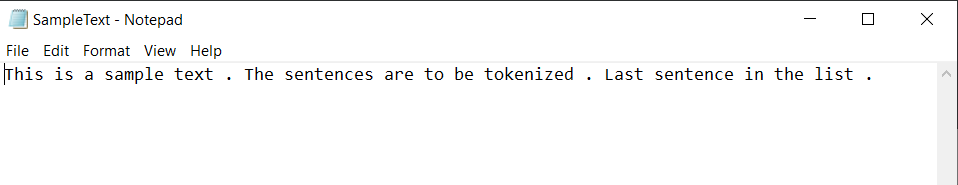
for i in range(0,(len(sentences))):

element = sentences[i].split()

token.append(element)

print("Tokenization of Words of Sentences",token)

OUTPUT



Write a program (using nltk toolkit in python environment) to tokenize

A sentence

PROCEDURE

//Importing NLTK library to use the word\_tokenize function

Import nltk.tokenize.word\_tokenize

FileReader=open(File, read)

Sentence=FileReader.read()

Print(word\_tokenize(Sentence))

CODE

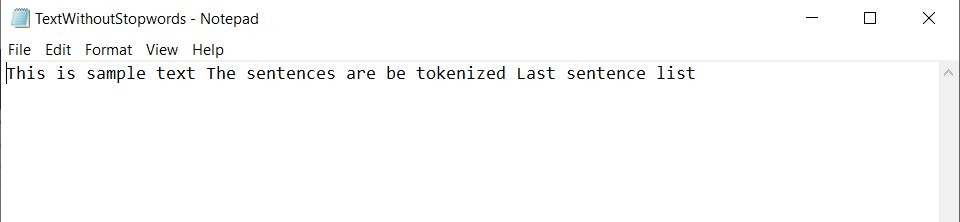
from nltk.tokenize import word\_tokenize

fr= open("../TextWithoutStopwords.txt")

sentence= fr.read()

print(word\_tokenize(sentence))

OUTPUT



A paragraph

PROCEDURE

//Importing NLTK library to use the sent\_tokenize function

Import nltk.tokenize.sent\_tokenize

FileReader=open(File, read)

Paragraph=FileReader.read()

Print(sent\_tokenize(Paragraph))

CODE

from nltk.tokenize import sent\_tokenize

fr= open("../SampleText.txt")

para= fr.read()

print(sent\_tokenize(para))

OUTPUT