Registration Number:- 19BCE2119 Name:- Gaurav Kumar Singh
Course:- Web Mining (L5+L6)

1) 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’]

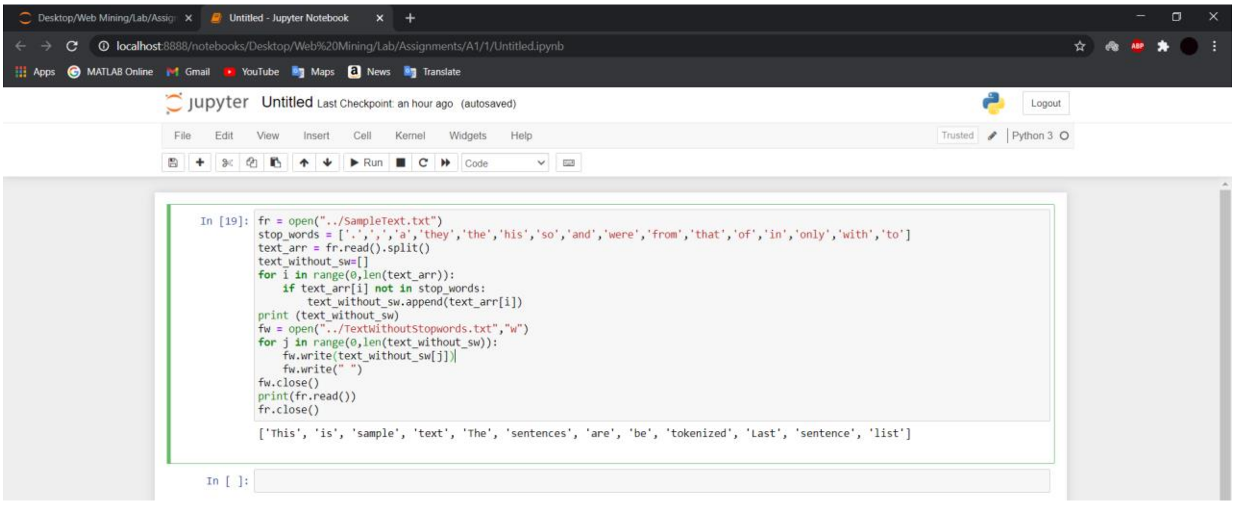
PROCEDURE

```
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()
```

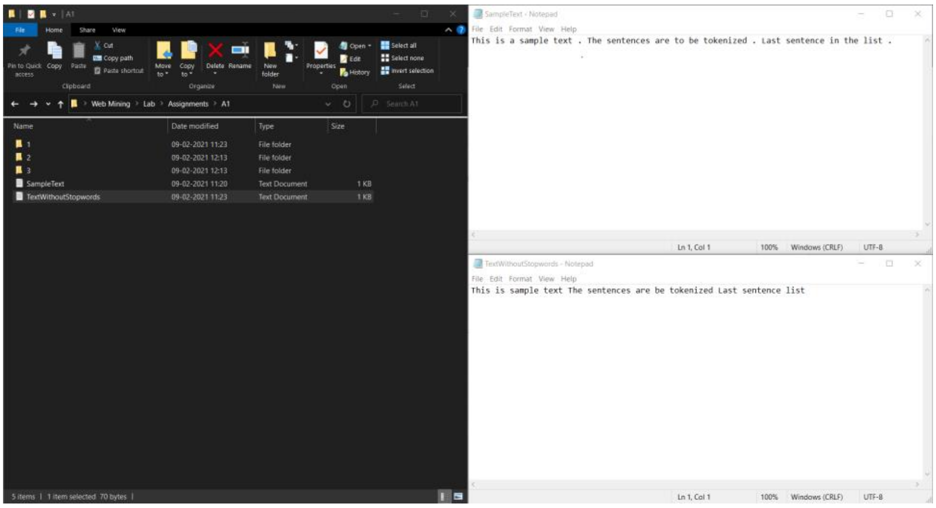
OUTPUT


Desktop/Web Mining/Lab/Assig
Untitled - Jupyter Notebook
localhost:8888/notebooks/Desktop/Web%20Mining/Lab/Assignments/A1/1/Untitled.ipynb
Apps
MATLAB Online
Gmail
YouTube
Maps
News
Translate
jupyter
Untitled Last Checkpoint: an hour ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code

```
In [19]: 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()

         ['This', 'is', 'sample', 'text', 'The', 'sentences', 'are', 'be', 'tokenized', 'Last', 'sentence', 'list']

In [ ]:
```


A1
File
Home
Share
View
Web Mining › Lab › Assignments › A1
Search A1
5 items | 1 item selected 70 bytes |

Name	Date modified	Type	Size
1	09-02-2021 11:23	File folder	
2	09-02-2021 12:13	File folder	
3	09-02-2021 12:13	File folder	
SampleText	09-02-2021 11:20	Text Document	1 KB
TextWithoutStopwords	09-02-2021 11:23	Text Document	1 KB

SampleText - Notepad

```
This is a sample text . The sentences are to be tokenized . Last sentence in the list .
```

TextWithoutStopwords - Notepad

```
This is sample text The sentences are be tokenized Last sentence list
```

2) Write a program to tokenize (without Nltk)

a) A sentence

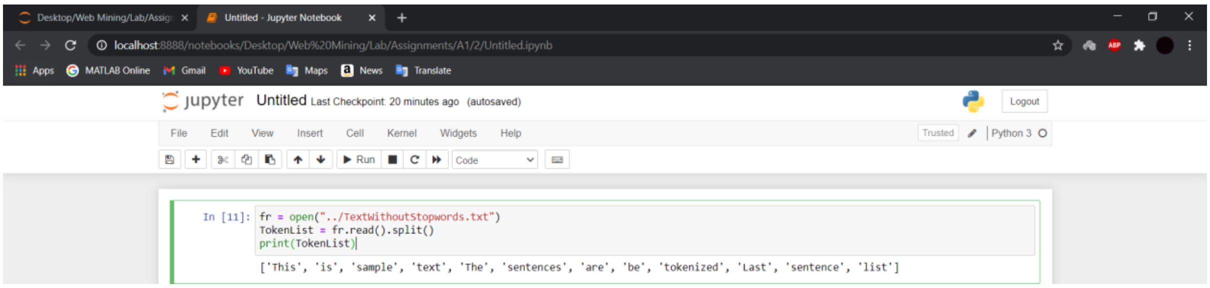
PROCEDURE

```
FileReader = open (File, read)
Tokens= FileReader.read().split()
Print(Tokens)
```

CODE

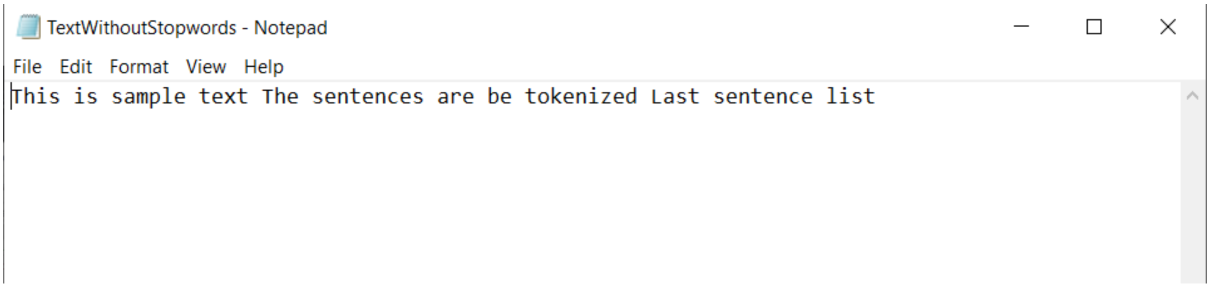
```
fr = open("../TextWithoutStopwords.txt")
TokenList = fr.read().split()
print(TokenList)
```

OUTPUT


Desktop/Web Mining/Lab/Assign
Untitled - Jupyter Notebook
localhost:8888/notebooks/Desktop/Web%20Mining/Lab/Assignments/A1/2/Untitled.ipynb
Apps
MATLAB Online
Gmail
YouTube
Maps
News
Translate
jupyter
Untitled Last Checkpoint: 20 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code

```
In [11]: fr = open("../TextWithoutStopwords.txt")
         TokenList = fr.read().split()
         print(TokenList)

         ['This', 'is', 'sample', 'text', 'The', 'sentences', 'are', 'be', 'tokenized', 'Last', 'sentence', 'list']
```



```
This is sample text The sentences are be tokenized Last sentence list
```

b) Multiple Sentences

PROCEDURE

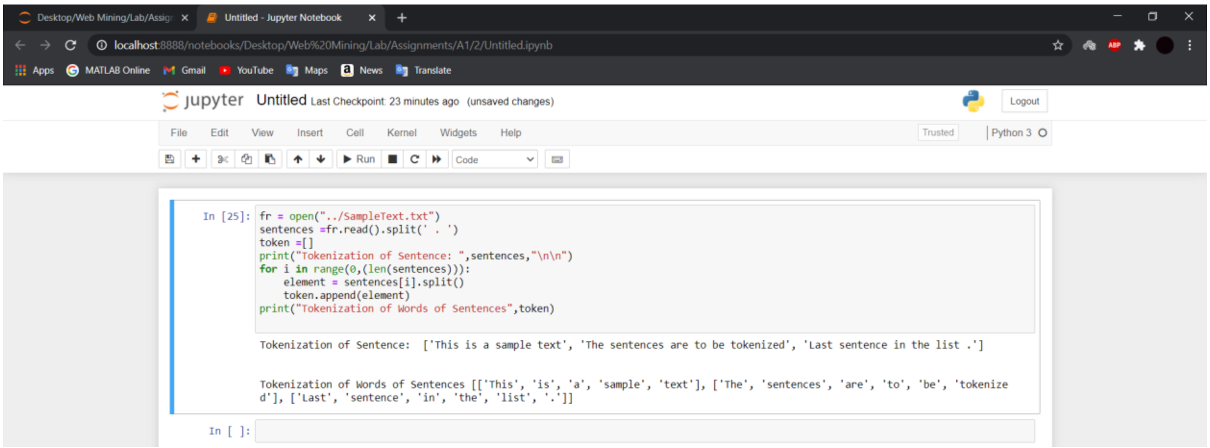
```
FileReader = open (File, read)
Sentences= FileReader.read().split(‘ . ’)
Print(Sentences)
For i in (0 to lengthof(Sentences))
    Arr[] = Sentences[i].split()
    Text.append(Arr)

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



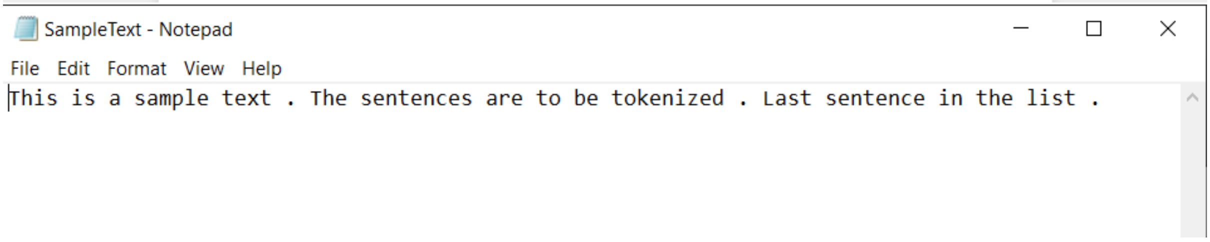
jupyter Untitled Last Checkpoint: 23 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code

```
In [25]: 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)

         Tokenization of Sentence:  ['This is a sample text', 'The sentences are to be tokenized', 'Last sentence in the list .']


         Tokenization of Words of Sentences [['This', 'is', 'a', 'sample', 'text'], ['The', 'sentences', 'are', 'to', 'be', 'tokenize
         d'], ['Last', 'sentence', 'in', 'the', 'list', '.']]

In [ ]:
```



```
This is a sample text . The sentences are to be tokenized . Last sentence in the list .
```

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

a) A sentence

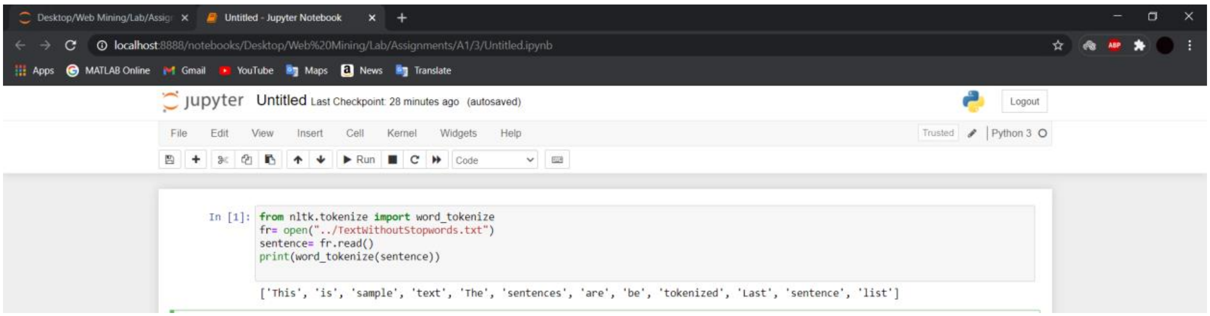
PROCEDURE

```
//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))
```

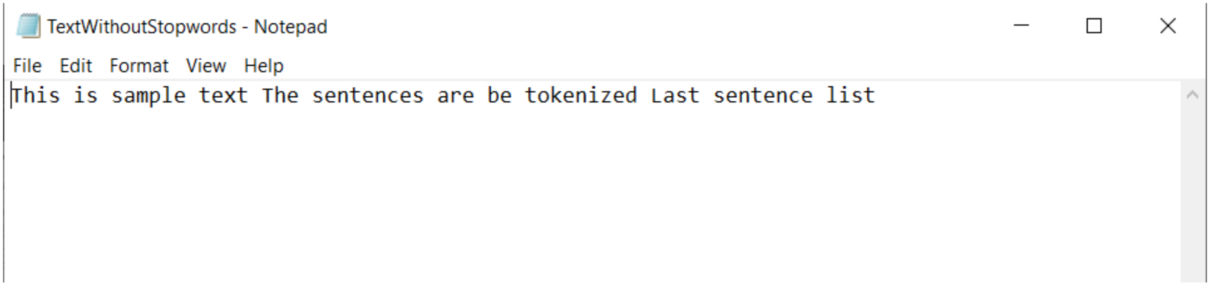
OUTPUT



Desktop/Web Mining/Lab/Assig
Untitled - Jupyter Notebook
localhost:8888/notebooks/Desktop/Web%20Mining/Lab/Assignments/A1/3/Untitled.ipynb
Apps
MATLAB Online
Gmail
YouTube
Maps
News
Translate
jupyter Untitled Last Checkpoint: 28 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code

```
In [1]: from nltk.tokenize import word_tokenize
        fr= open("../TextWithoutStopwords.txt")
        sentence= fr.read()
        print(word_tokenize(sentence))

        ['This', 'is', 'sample', 'text', 'The', 'sentences', 'are', 'be', 'tokenized', 'Last', 'sentence', 'list']
```



```
This is sample text The sentences are be tokenized Last sentence list
```

b) A paragraph

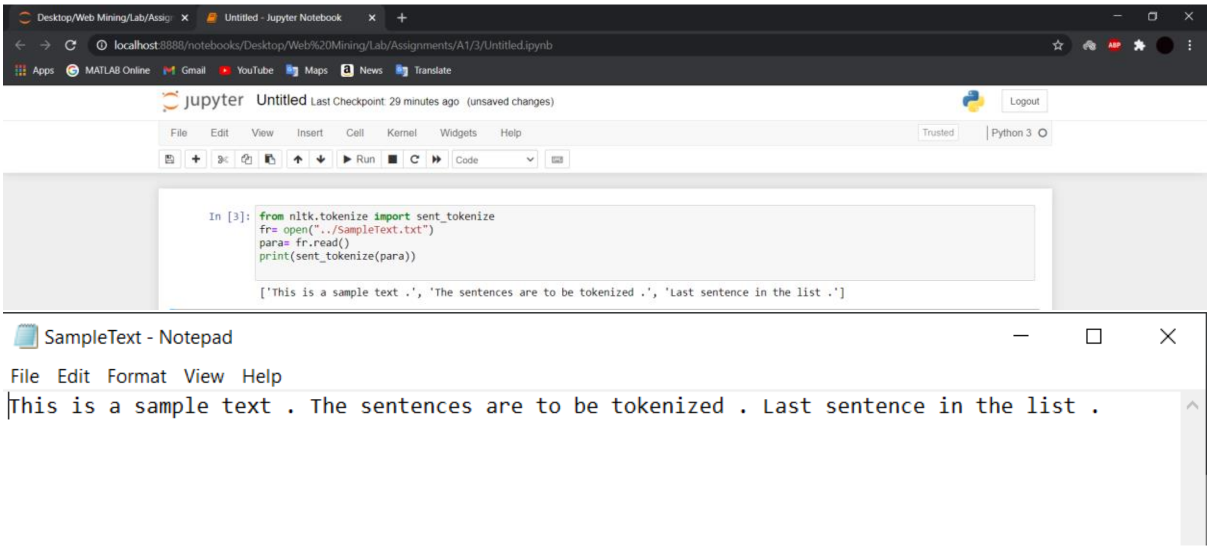
PROCEDURE

```
//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



Desktop/Web Mining/Lab/Assigr
Untitled - Jupyter Notebook
localhost:8888/notebooks/Desktop/Web%20Mining/Lab/Assignments/A1/3/Untitled.ipynb
Apps
MATLAB Online
Gmail
YouTube
Maps
News
Translate
jupyter Untitled Last Checkpoint: 29 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
SampleText - Notepad
File
Edit
Format
View
Help

```
In [3]: from nltk.tokenize import sent_tokenize
        fr= open("../SampleText.txt")
        para= fr.read()
        print(sent_tokenize(para))

        ['This is a sample text .', 'The sentences are to be tokenized .', 'Last sentence in the list .']
```

```
This is a sample text . The sentences are to be tokenized . Last sentence in the list .
```