

WEB MINING

by Sharadindu Adhikari, 19BCE2105

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

2. Write a program to tokenize
a) A sentence b) Multiple sentences (Without Nltk)

Experiment 2

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

a) Sentence
b) A paragraph
c) Information of a complete web page

The screenshot also shows the Zoom meeting interface with a toolbar at the top and a participant list at the bottom.

Experiment-2

1(a,b).Tokenize multiple sentences using nltk.

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
word_data = "It originated from the idea that there are readers who prefer learning
new skills from the comforts of their drawing rooms"
nltk_tokens = word_tokenize(word_data)
print (nltk_tokens)
```

Output:

```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.19042.804]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>python moj.txt
['the quick brown fox jumps over a lazy dog', 'the five boxing wizards jump quickly', '']

C:\WINDOWS\system32>python lan.txt
['It', 'originated', 'from', 'the', 'idea', 'that', 'there', 'are', 'readers', 'who', 'prefer', 'learning', 'new', 'skills', 'from', 'the', 'comforts', 'of', 'their', 'drawing', 'rooms']

C:\WINDOWS\system32>
```

1(c).Tokenize a webpage using nltk.

Solution:

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from urllib import request
url = "http://www.gutenberg.org/files/2554/2554-0.txt"
response = request.urlopen(url)
raw = response.read().decode('utf8')
tokens = word_tokenize(raw)
print(tokens[:10])
```

Output:

```
IDLE Shell 3.9.1
File Edit Shell Debug Options Window Help
Python 3.9.1 (tags/v3.9.1:1e5d33e, Dec 7 2020, 17:08:21) [MSC v.1927 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: C:\Users\rishi\OneDrive\Desktop\p.py =====
['\uffeffThe', 'Project', 'Gutenberg', 'EBook', 'of', 'Crime', 'and', 'Punishment', ',', 'by']
>>> |
```