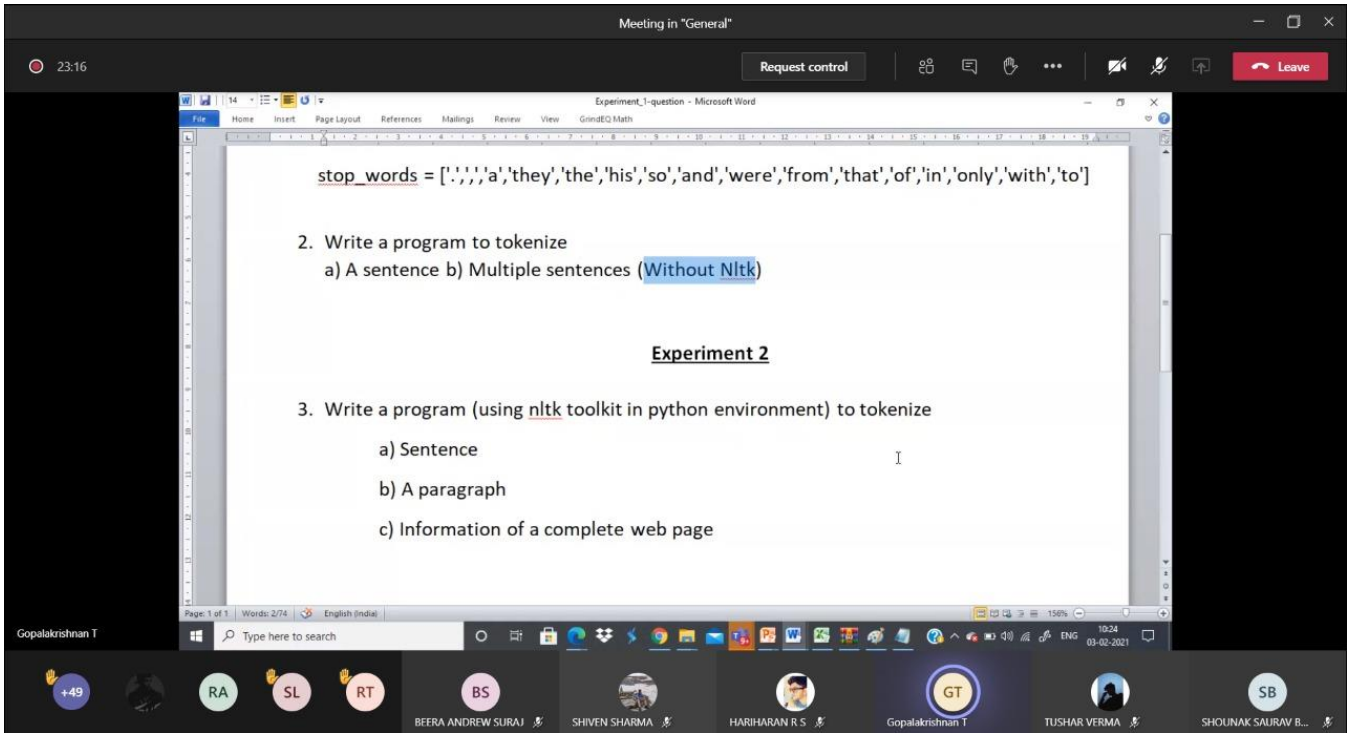
Lab DA-2, Winter 2020-21, L15+L16

WEB MINING

by Sharadindu Adhikari, 19BCE2105



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:

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: