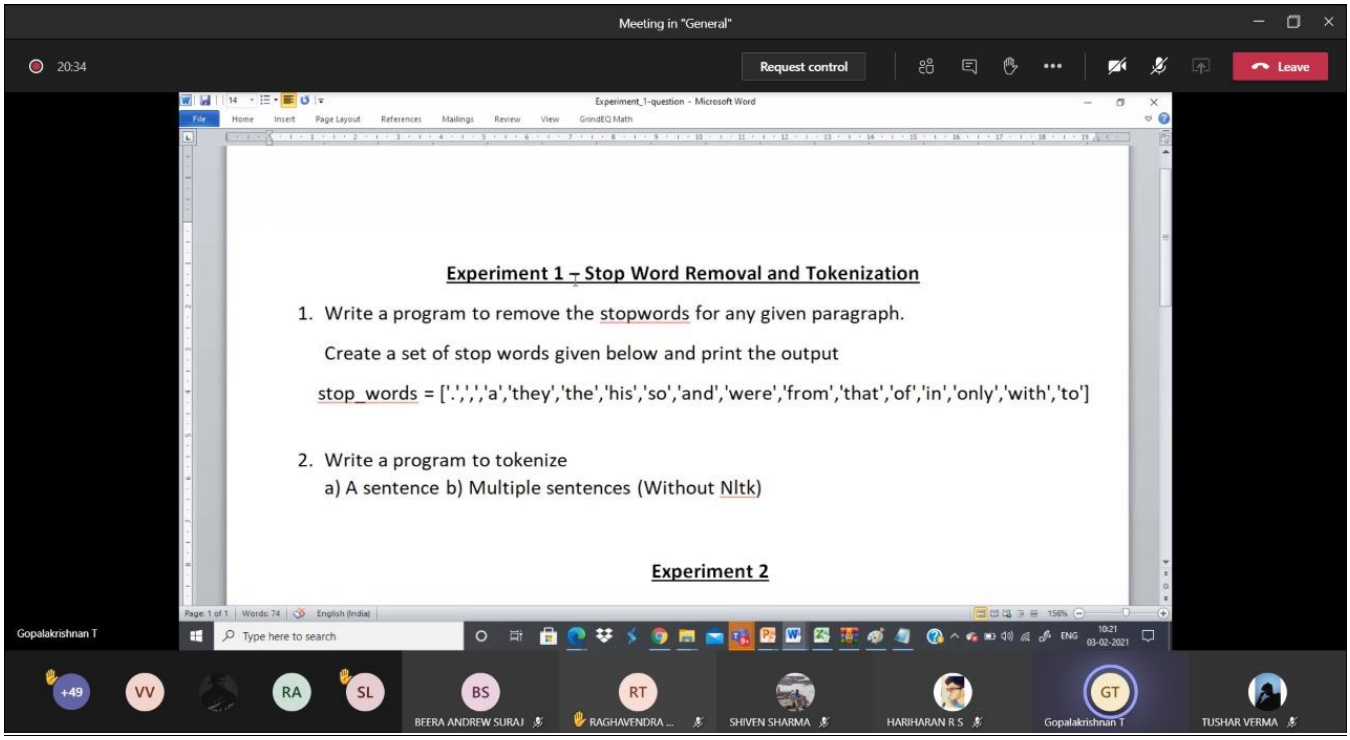
Lab DA-1, Winter 2020-21, L15+L16

WEB MINING

by Sharadindu Adhikari, 19BCE2105



Experiment 1:

1(a).Stop Words removal from a paragraph.

Solution:-

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
ex = """Web Mining is the process of Data Mining techniques to automatically
discover and extract information from Web documents and services. The main purpose
of web mining is discovering useful information from the World-Wide Web and its
usage patterns."""
  stop_words = set(stopwords.words('english'))
  token = word_tokenize(ex)
  fs = [w for w in word_tokens if not w in stop_words]
  fs = []
  for w in word_tokens:
    if w not in stop_words:
        fs.append(w)
print("Before Removal=>")
print(token)
print("After Removal=>")
print(fs)
```

Output:-

1(b). Create a set of stop words given below and print the output:-

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

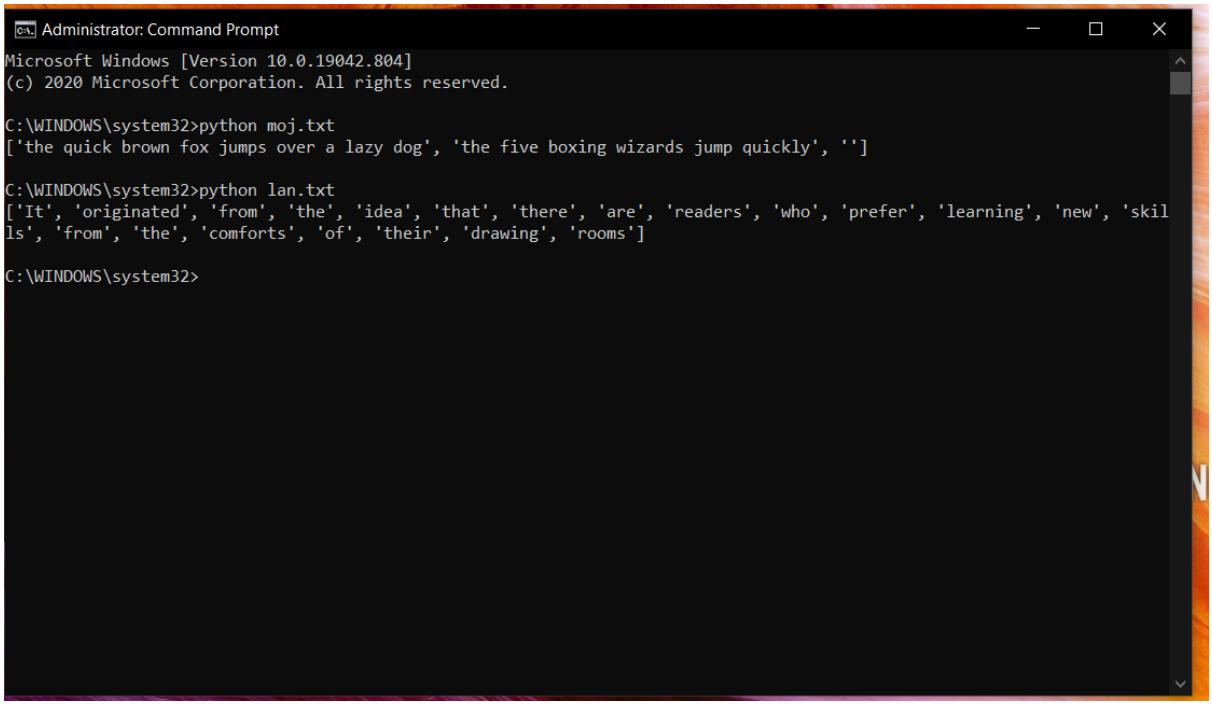
Solution: -

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

ex = """. , a they the his so and were from that of in only with to"""
stop_words = set(stopwords.words('english'))
token = word_tokenize(example_sent)
fs = [w for w in word_tokens if not w in stop_words]
fs = []
for w in word_tokens:
    if w not in stop_words:
        fs.append(w)

print("Before Removal=>")
print(token)
print("After Removal=>")
print(fs)
```

Output:

2(a,b). Write a program to tokenise a sentence(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)
```



2(c). Write a program to tokenize multiple sentences without using nltk.

```
text = """the quick brown fox jumps over a lazy dog.the five boxing wizards jump quickly"""
# Splits at '.'
print(text.split('. '))
```