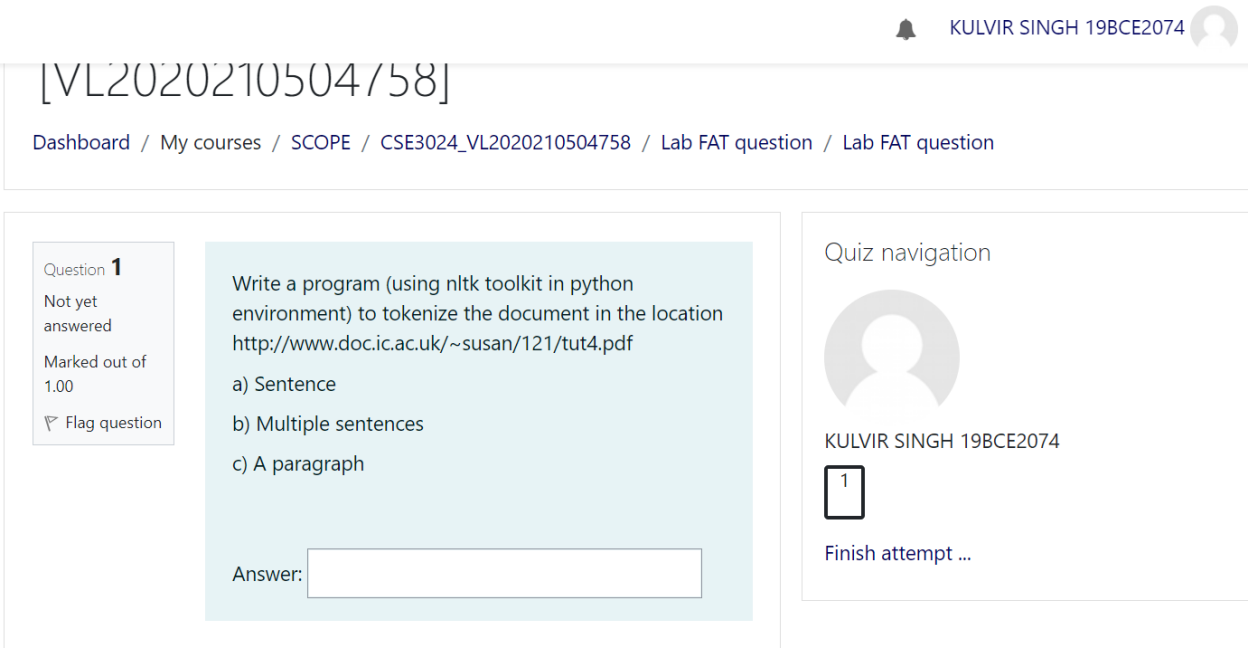
Web Mining Lab FAT

Name: Kulvir Singh
Register Number: 19BCE2074

Question :



Procedure :

Install the nltk toolkit in python environment using the pip install command. Import nltk to the code file. Download the class stopwords and punkt from nltk toolkit. Import stropwords from nltk.corpus and word_tokenize from nltk.tokenize.

Open the given file in the question. Create three variables and store a sentence, multiple sentences and a paragraph in the three varibles. Create a variable which stores the stopwords that are fetched from nltk. Use the tokenize method to create tokens of the 3 variables which contain file data. Loop through the tokens and filter out the stopwords. Display the tokens and filtered array of each of the three variables.

Code :

```
!pip install nltk
import nltk

nltk.download('stopwords')
nltk.download('punkt')

from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
print("KULVIR SINGH 19BCE2074")
sentenceFromFile = "The scorer keeps this list secret: it is called the co
de."
multipleSentencesFromFile = "The scorer keeps this list secret: it is call
ed the code. The guesser now tries to guess the code.The scorer gives a sc
ore to each guess the guesser makes."
paragraphFromFile = "Cows and Bulls, is played between two players, the sc
orer and the guesser. The scorer chooses a list of 4 numbers (repetitions
are not allowed) from the numbers 1, 2, 3, 4, 5, 6, 7, 8 and 9."

stop_words = set(stopwords.words('english'))

print("a)Sentence")
words_tokens = word_tokenize(sentenceFromFile)
filtered_paragraph = [w for w in words_tokens if not w in stop_words]
print("Tokenized Sentence = \n",words_tokens)
print("Filtered sentence = \n",filtered_paragraph)

print("b)Multiple Sentences")
words_tokens = word_tokenize(multipleSentencesFromFile)
filtered_paragraph = [w for w in words_tokens if not w in stop_words]
print("Tokenized multiple sentences = \n",words_tokens)
print("Filtered multiple sentences = \n",filtered_paragraph)

print("c)Paragraph")
words_tokens = word_tokenize(paragraphFromFile)
filtered_paragraph = [w for w in words_tokens if not w in stop_words]
print("Tokenized Paragrph = \n",words_tokens)
print("Filtered Paragraph = \n",filtered_paragraph)
```

Output Screenshots :

```
Requirement already satisfied: nltk in /usr/local/lib/python3.7/dist-packages (3.2.5)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from nltk) (1.15.0)
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
KULVIR SINGH 19BCE2074
a)Sentence
Tokenized Sentence =
 ['The', 'scorer', 'keeps', 'this', 'list', 'secret', ':', 'it', 'is', 'called', 'the', 'code', '.']
Filtered sentence =
 ['The', 'scorer', 'keeps', 'list', 'secret', ':', 'called', 'code', '.']
b)Multiple Sentences
Tokenized multiple sentences =
 ['The', 'scorer', 'keeps', 'this', 'list', 'secret', ':', 'it', 'is', 'called', 'the', 'code', '.', 'The', 'guesser', 'now', 'tries', 'to', 'guess',
Filtered multiple sentences =
 ['The', 'scorer', 'keeps', 'list', 'secret', ':', 'called', 'code', '.', 'The', 'guesser', 'tries', 'guess', 'code.The', 'scorer', 'gives', 'score',
c)Paragraph
Tokenized Paragrph =
 ['Cows', 'and', 'Bulls', ',', 'is', 'played', 'between', 'two', 'players', ',', 'the', 'scorer', 'and', 'the', 'guesser', '.', 'The', 'scorer', 'cho
Filtered Paragraph =
 ['Cows', 'Bulls', ',', 'played', 'two', 'players', ',', 'scorer', 'guesser', '.', 'The', 'scorer', 'chooses', 'list', '4', 'numbers', '(', 'repetiti
```

Output in text form :

```
Requirement already satisfied: nltk in /usr/local/lib/python3.7/dist-
packages (3.2.5)
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-
packages (from nltk) (1.15.0)
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
KULVIR SINGH 19BCE2074
a)Sentence
Tokenized Sentence =
 ['The', 'scorer', 'keeps', 'this', 'list', 'secret', ':', 'it', 'is',
'called', 'the', 'code', '.']
Filtered sentence =
 ['The', 'scorer', 'keeps', 'list', 'secret', ':', 'called', 'code', '.']
b)Multiple Sentences
Tokenized multiple sentences =
 ['The', 'scorer', 'keeps', 'this', 'list', 'secret', ':', 'it', 'is',
'called', 'the', 'code', '.', 'The', 'guesser', 'now', 'tries', 'to',
'guess', 'the', 'code.The', 'scorer', 'gives', 'a', 'score', 'to', 'each',
'guess', 'the', 'guesser', 'makes', '.']
Filtered multiple sentences =
 ['The', 'scorer', 'keeps', 'list', 'secret', ':', 'called', 'code', '.',
'The', 'guesser', 'tries', 'guess', 'code.The', 'scorer', 'gives',
'score', 'guess', 'guesser', 'makes', '.']
c)Paragraph
Tokenized Paragrph =
 ['Cows', 'and', 'Bulls', ',', 'is', 'played', 'between', 'two',
'players', ',', 'the', 'scorer', 'and', 'the', 'guesser', '.', 'The',
'scorer', 'chooses', 'a', 'list', 'of', '4', 'numbers', '(',
'repetitions', 'are', 'not', 'allowed', ')', 'from', 'the', 'numbers',
'1', ',', '2', ',', '3', ',', '4', ',', '5', ',', '6', ',', '7', ',', '8',
'and', '9', '.']
Filtered Paragraph =
```

```
 ['Cows', 'Bulls', ',', 'played', 'two', 'players', ',', 'scorer',
'guesser', '.', 'The', 'scorer', 'chooses', 'list', '4', 'numbers', '(',
'repetitions', 'allowed', ')', 'numbers', '1', ',', '2', ',', '3', ',',
'4', ',', '5', ',', '6', ',', '7', ',', '8', '9', '.']
```