**CSE-3024 Web Mining**

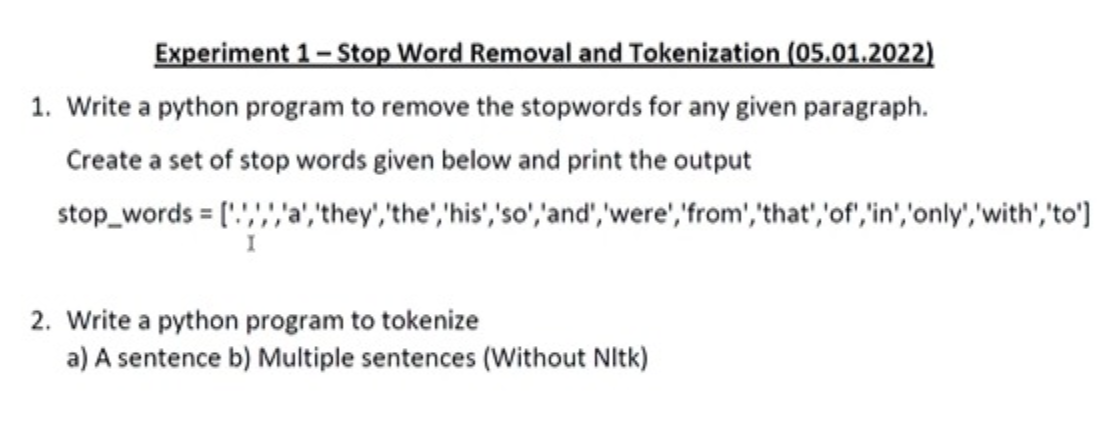
**Lab Assignment 1**

**Alokam Nikhitha**

**19BCE2555**

Question 1

**Problem statement:**



**Procedure:**

* At First, we import the text file in our work space. To do this we can use open method of python which reads the file into our workspace.
* Next, we read each word into a variable as string using a nested for loop wherein we split each word whenever we encounter a space.
* Next, using regex in python we remove the punctuations from our string input. This will make sure that tokens are free from sentence structure.
* We tokenize each token in our text using split() function of NumPy lists and save it in a list.
* Then we use NumPy’s unique method to only include unique tokens from our identified set of tokens.
* A tentative list of stop words and using a nested for loop we check if the given token belongs to that list or not, to remove stopwords. If it doesn’t then we save it else we discard it.
* Finally, we print our list that contains the resultant tokens after removal of stop words.

**Code:**

#Reading input from a text file and saving it as a string

text = ""

with open('test\_file.txt') as file:

for line in file:

for word in line.split():

text= text + " " + word

#Removing punctuations from our input file

import re

text = re.sub(r'[^\w\s]', '', text)

text

#Printing each token

print(text.split())

#Printing unique tokens

import numpy as np

print(np.unique(text.split()))

#Removing StopWords

stopwords = ["i", "a", "am", "and", "at", "for", "in", "is", "my", "of", "this"]

res = []

for x in tokens:

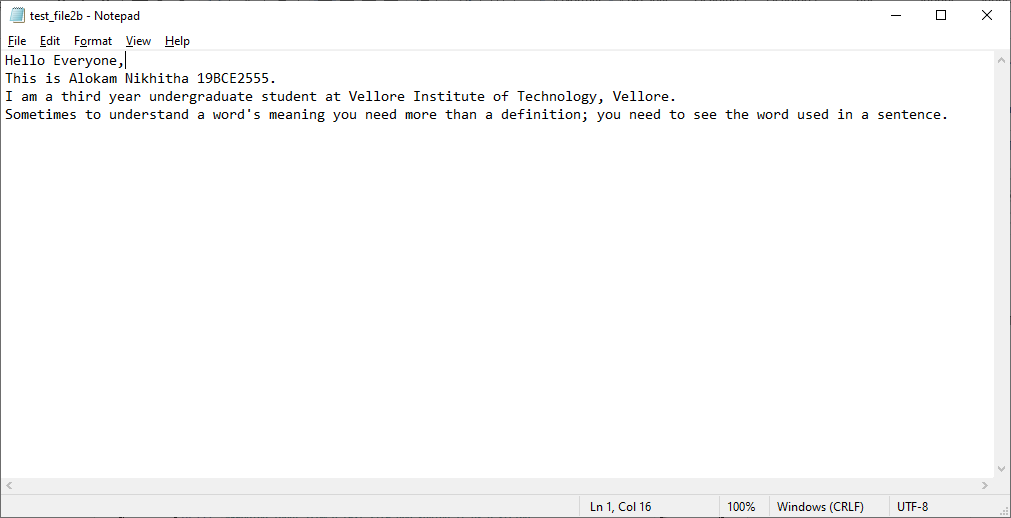
if x not in stopwords:

res.append(x)

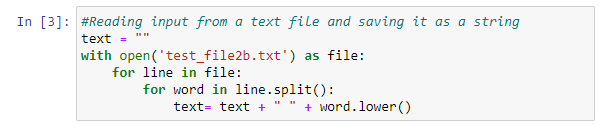
#Printing cleaned tokens in our input text file

print(res)

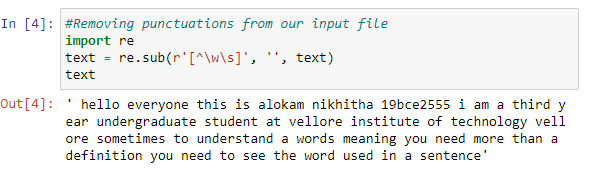
**Text File Taken as Input:**

****

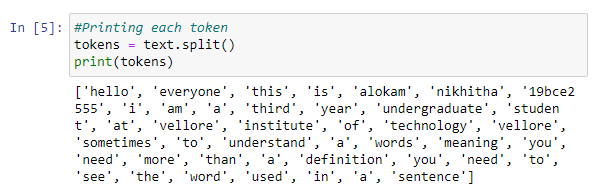
**Code Snippets and Outputs:**



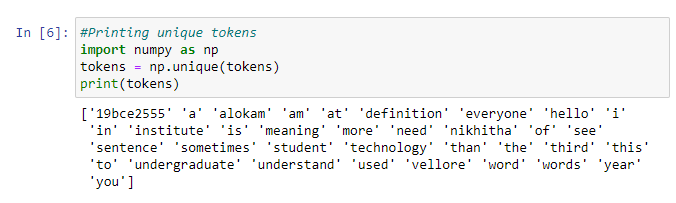
Here we are reading the text file using open method in python. Then reading each line we split each word and append it to a string variable with a space in between.



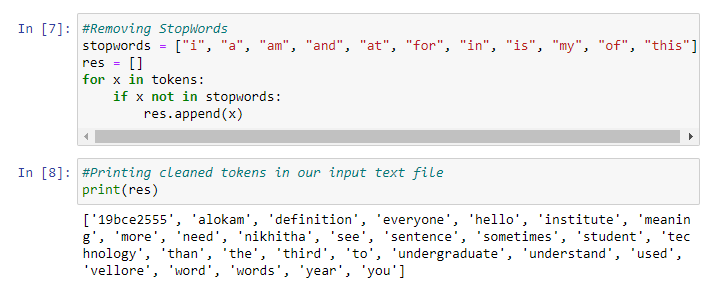
Here we are removing punctuations from our input file. This is done using regex, where we keep only alphanumeric inputs in our text string. We can see all the periods and commas from original input files are removed here.



Next, we are splitting each word in our string using space character. Clearly, they form a token and hence we print each token.



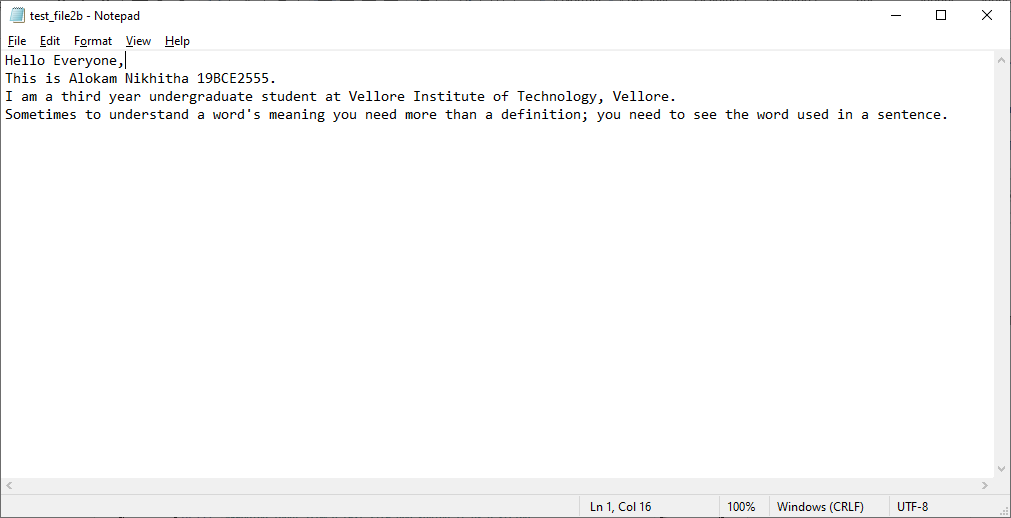
Here we are only printing unique tokens from all the generated tokens using split method. This is done using NumPy’s unique function, which identifies all the unique elements from a list.



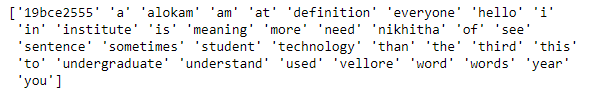
Here we remove all the stop words from a self-defined list of stop words. We use nested loop to check if given token belongs to both tokens list and stopwords list. If it does, we don’t add it to our result else we add it to our results.

**Results and Output**

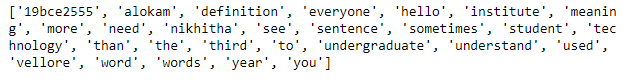
**- Input text:**

****

**- Tokens of input text:**

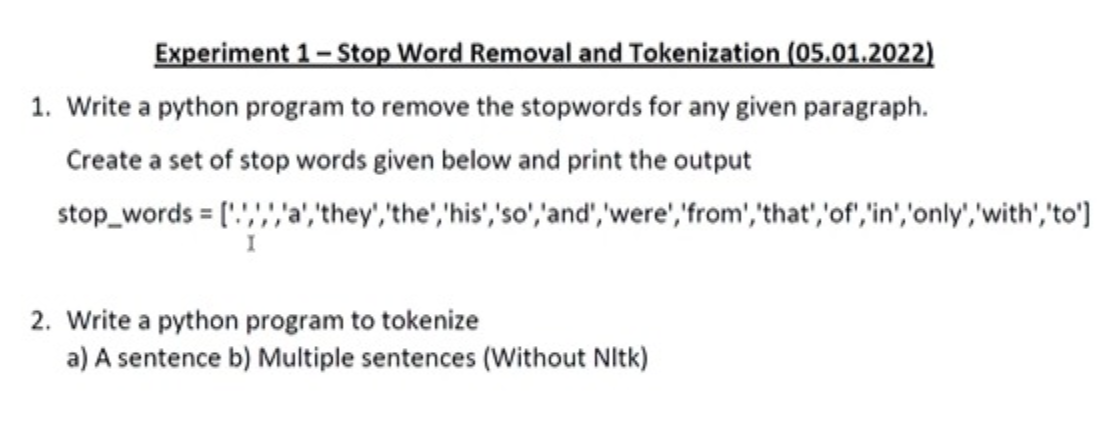
****

**-Tokens after removal of stop words:**

****

Question 2

**Problem statement:**



**Procedure:**

* Firstly, we import the text file in our work space. To do this we can use open method of python which reads the file into our workspace.
* Next, we read each word into a variable as string. This can be done using a nested for loop wherein we split each word whenever we encounter a space.
* Next, using regex in python we remove the punctuations from our string input. This will make sure that tokens are free from sentence structure.
* Finally, we will print each token and then unique tokens.

**a)**

**Code:**

#Reading input from a text file and saving it as a string

text = ""

with open('test\_file2a.txt') as file:

for line in file:

for word in line.split():

text= text + " " + word

#Removing punctuations from our input file

import re

text = re.sub(r'[^\w\s]', '', text)

text

#Printing each token

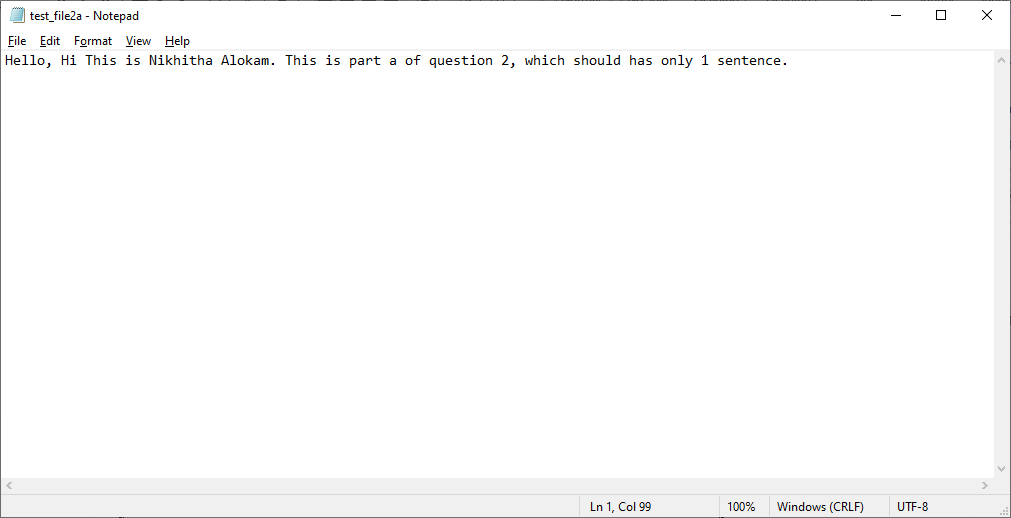
print(text.split())

#Printing unique tokens

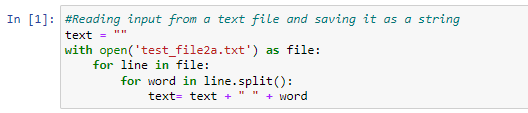
import numpy as np

print(np.unique(text.split()))

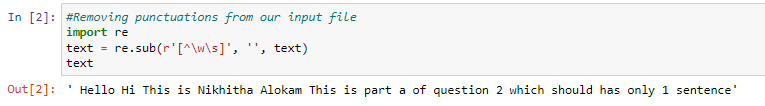
**Text File Taken as Input:**

****

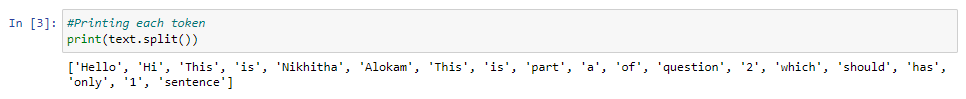
**Code Snippets and Outputs:**



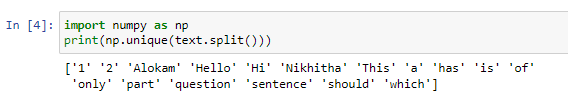
Here we are reading the text file using open method in python. Later we are reading each line we split each word and append it to a string variable with a space in between.



Here we are removing punctuations from our input file using regex, where we keep only alphanumeric inputs in our text string. We can see all the periods and commas from original input files are removed here.



Next, we are splitting each word in our string using space character. Clearly, they form a token and hence we print each token.



Here we are only printing unique tokens from all the generated tokens using split method. This is done using NumPy’s unique function, which identifies all the unique elements from a list.

**Results and Output**

**- Input text:**

****

**- Tokens of input text:**

****

**b)**

**Code:**

#Reading input from a text file and saving it as a string

text = ""

with open('test\_file2b.txt') as file:

for line in file:

for word in line.split():

text= text + " " + word

#Removing punctuations from our input file

import re

text = re.sub(r'[^\w\s]', '', text)

text

#Printing each token

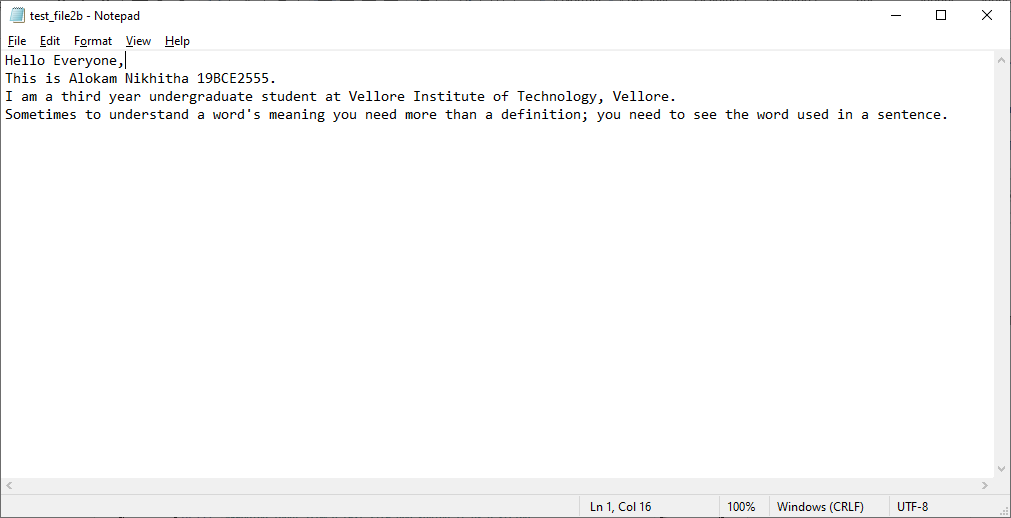
print(text.split())

#Printing unique tokens

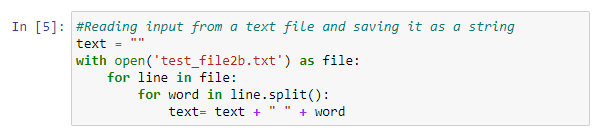
import numpy as np

print(np.unique(text.split()))

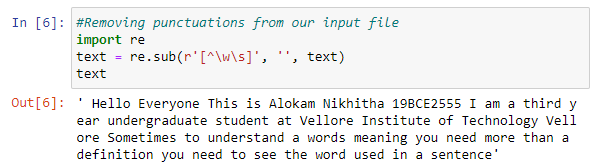
**Text File Taken as Input:**

****

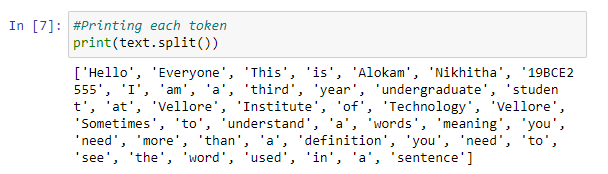
**Code Snippets and Outputs:**



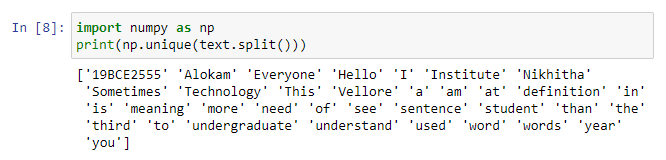
Here we are reading the text file using open method in python. Then reading each line we split each word and append it to a string variable with a space in between.



Here we are removing punctuations from our input file. This is done using regex, where we keep only alphanumeric inputs in our text string. We can see all the periods and commas from original input files are removed here.



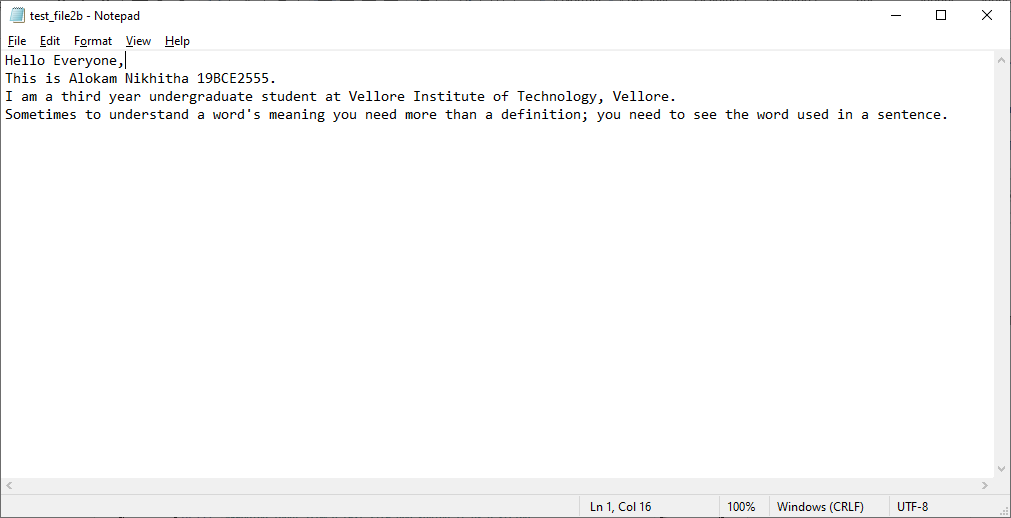
Next, we are splitting each word in our string using space character. Clearly, they form a token and hence we print each token.



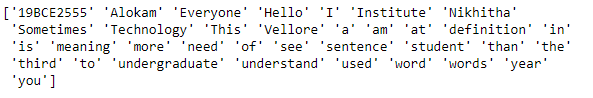
Here we are only printing unique tokens from all the generated tokens using split method. This is done using NumPy’s unique function, which identifies all the unique elements from a list.

**Results and Output:**

**- Input text:**

****

**- Tokens of input text:**

****