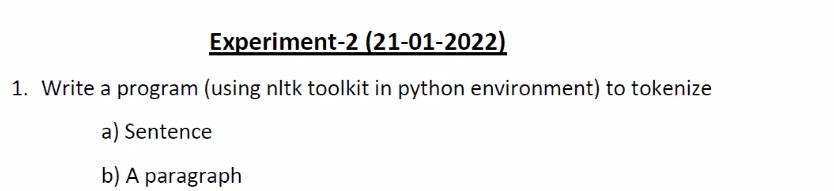
**CSE-3024 Web Mining**

**Lab Assignment 2**

**Alokam Nikhitha**

**19BCE2555**

Question 

**Problem statement:**

To write a program to tokenize using nltk toolkit in Python Environment.

**Procedure:**

* We firstly import our text file into our workspace. To do that we are able to use open method of python which will read our text file to the workspace.
* Next, we will import the necessary NLTK libraries including stopwords, sent\_tokenize and word\_tokenize.
* Using word\_tokenize we tokenize each word and store it in a variable list named tokens
* Then, we will read every word in input file as a string input. This may be carried out using nested for loop in which we split every word whenever we come across a space.
* Then , we use regex to remove the punctuations from the input string. This will render token a higher syntactic shape and break the sentence bonds.
* Then ,we split and store every token right into a list with nltk’s sentence\_tokenize method.
* Then subsequently we remove stop words in the same process as in preceding assignment.
* Finally, we print our list that contains the resultant tokens post removal of stop words as well.

**a) Sentence**

**Code:**

#Reading input (Single ) from a text file

text = ""

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

for line in file:

for word in line.split():

text = text + " " + word.lower()

#Importing libraries

import re

import nltk

from nltk.tokenize import sent\_tokenize, word\_tokenize

from nltk.corpus import stopwords

#Removing punctuations from our input .

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

text

#Printing each token

tokens = word\_tokenize(text)

print(tokens)

#Printing unique tokens

import numpy as np

tokens = np.unique(tokens)

print(tokens)

#Removing Stopwords

res = []

for x in tokens:

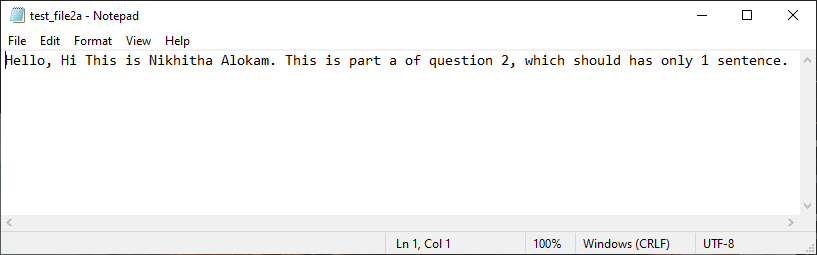
if x not in set(stopwords.words('english')):

res.append(x)

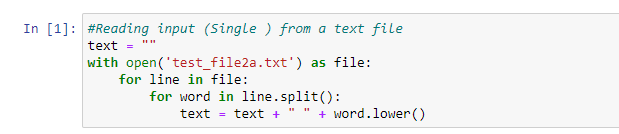
#Printing cleaned tokens in our input text file

print(res)

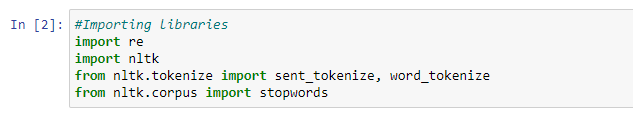
**Text File Taken as Input:**

****

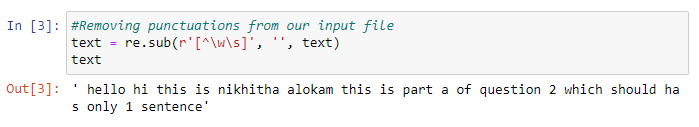
**Code Snippets and Outputs:**

****

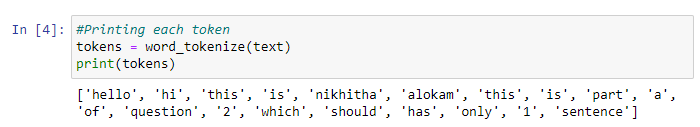
Here we're analyzing the text file by using open approach in python. Then analyzing every line we split every word and append it to a string variable with a space in between.



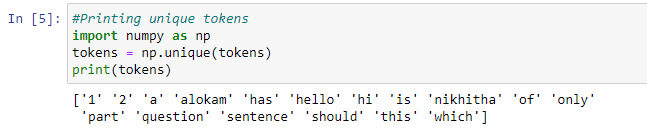
Here we're importing the necessary libraries which incorporates our library of challenge this is nltk. We additionally import stopwords and word\_tokenize, sentence\_tokenize from nltk. To eliminate punctuations we import the regex library.



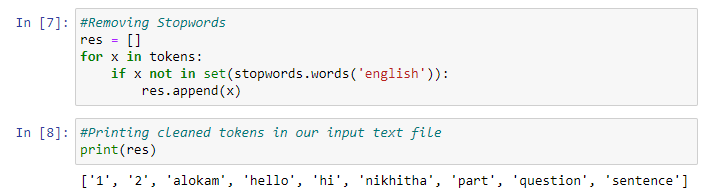
Here we're removing punctuations from the input taken from input file. This is carried out by using regex, wherein we keep only alphanumeric inputs in our text string. We can see all of the intervals and commas from original input file are eliminated here.



Since we have to tokenize every word, we have used the word\_tokenize and as we can see each token in identified here and we print them.



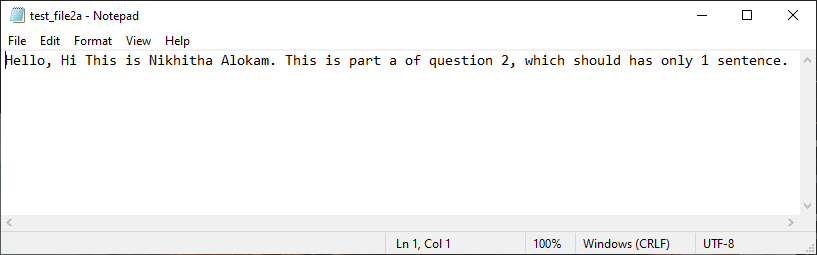
Here we print all the unique tokens in after tokenizing. And print them.



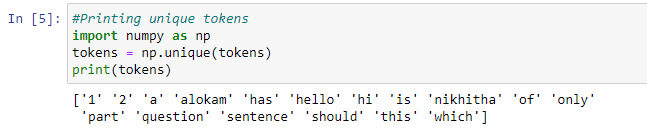
Here we remove all the stop words from a self-defined list of stop words. We use nested loop to check if given token belongs to both tokens list and stopwords list. If it does, we don’t add it to our result else we add it to our results.

**Results and Output**

1. Input Sentence

****

1. Tokens with out removing Stopwords.



1. Tokens after removing Stopwords



**b) A Paragraph**

**Code:**

#Reading input (Single ) from a text file

text = ""

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

for line in file:

for word in line.split():

text = text + " " + word.lower()

#Importing libraries

import re

import nltk

from nltk.tokenize import sent\_tokenize, word\_tokenize

from nltk.corpus import stopwords

#Removing punctuations from our input .

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

text

#Printing each token

tokens = word\_tokenize(text)

print(tokens)

#Printing unique tokens

import numpy as np

tokens = np.unique(tokens)

print(tokens)

#Removing Stopwords

res = []

for x in tokens:

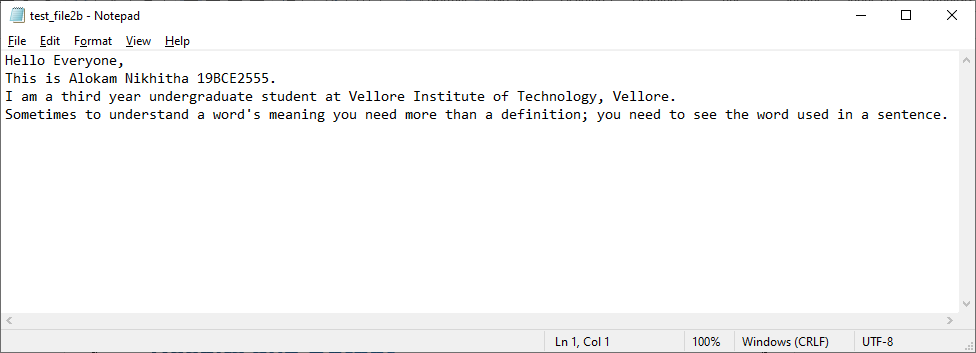
if x not in set(stopwords.words('english')):

res.append(x)

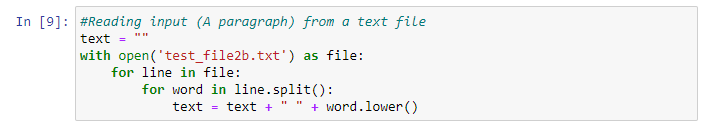
#Printing cleaned tokens in our input text file

print(res)

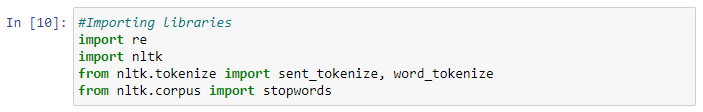
**Text File Taken as Input:**

****

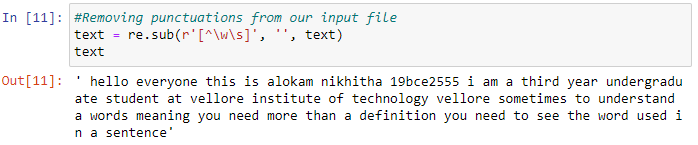
**Code Snippets and Outputs:**

****

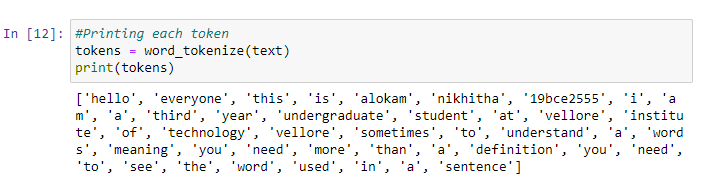
Here we're analyzing the text file by using open approach in python. Then analyzing every line we split every word and append it to a string variable with a space in between.



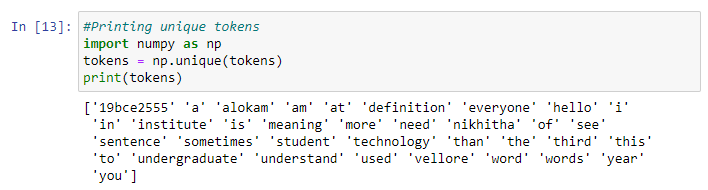
Here we're importing the necessary libraries which incorporates our library of challenge this is nltk. We additionally import stopwords and word\_tokenize, sentence\_tokenize from nltk. To eliminate punctuations we import the regex library



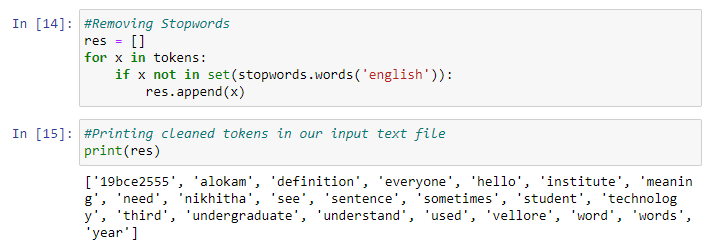
Here we're removing punctuations from the input taken from input file. This is carried out by using regex, wherein we keep only alphanumeric inputs in our text string. We can see all of the intervals and commas from original input file are eliminated here.



Since we have to tokenize every word, we have used the word\_tokenize and as we can see each token in identified here and we print them.



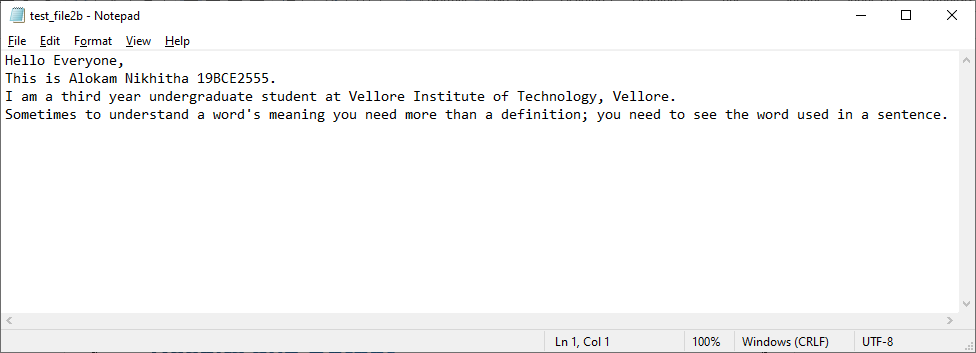
Here we print all the unique tokens in after tokenizing. And print them.

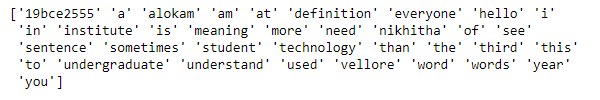


Here we remove all the stop words from a self-defined list of stop words. We use nested loop to check if given token belongs to both tokens list and stopwords list. If it does, we don’t add it to our result else we add it to our results.

**Results and Output**

1.Input Sentence

****2.Tokens with out removing Stopwords.

****

3.Tokens after removing Stopwords

