**Experiment 1**

*Question 1*

***Problem Statement :***

Write a program to remove the stopwords for any given paragraph. 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']

***Procedure :***

Store the given stop words in as a string. Use the inbuilt split() method to get the stop\_words array. Create a paragraph variable to store the paragraph to be tokenized. Use split() method to get the paragraph as tokens. Loop through the tokens and remove the stop words from it. Display the result.

***Code :***

stop\_words\_string = '. , a they the his so and were from that of in only with to'

stop\_words = stop\_words\_string.split()

print(stop\_words)

paragraph = "Ronaldo made his senior international debut for Portugal in 2003 at age 18 , and has since earned 170 caps , including appearing and scoring in ten major tournaments , becoming Portugal's most capped player and his country's all-time top goalscorer. He scored his first international goal at Euro 2004 where he helped Portugal reach the final and assumed full captaincy of the national team in July 2008. In 2015 Ronaldo was named the best Portuguese player of all time by the Portuguese Football Federation. The following year he led Portugal to their first triumph in a major tournament by winning Euro 2016, and received the Silver Boot as the second-highest goalscorer of the tournament."

result = ""

paragraph = paragraph.split()

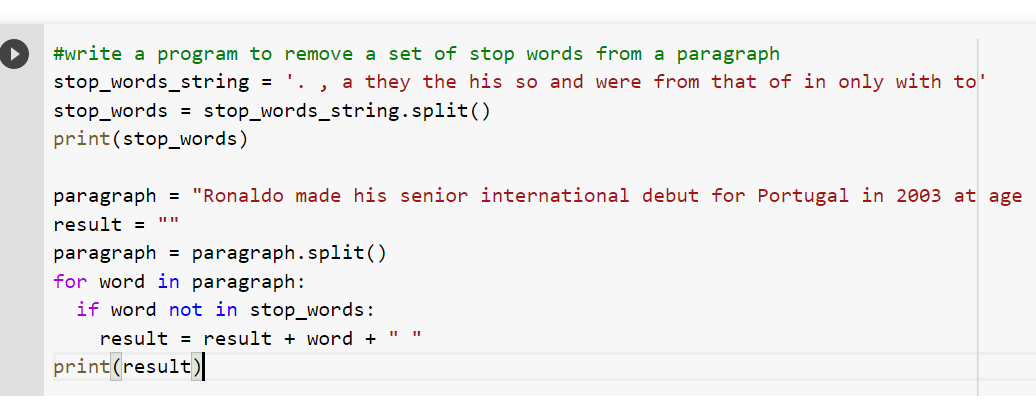
for word in paragraph:

if word not in stop\_words:

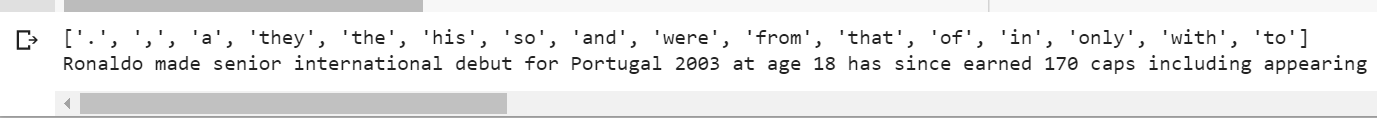
result = result + word + " "

print(result)

***Code Screenshot:***



***Output Screenshots :***



*Question 2*

***Problem Statement :***

2. Write a program to tokenize

a) A sentence

b) Multiple sentences (Without Nltk)

***Procedure :***

Store a sentence in a variable. Split the sentence using split() method and store the result in a variable. Display the array. Store a set of sentences, ie, paragraph in an array. Initialize an array variable. Loop through all the sentences in the array and using split() method split each sentence. Display the result.

***Code :***

sentence = 'My name is Kulvir'

sentence = sentence.split()

print('tokenized sentence = ',sentence)

paragraph = {'My first name is Kulvir','My surname is Singh','I am 19 years old'}

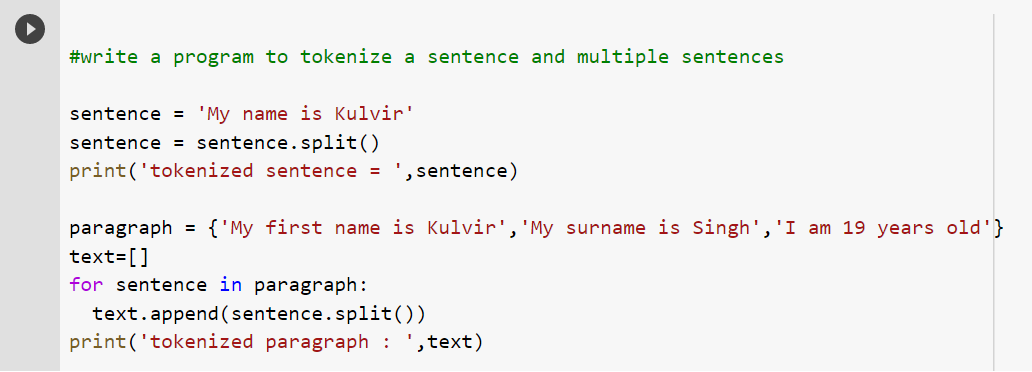
text=[]

for sentence in paragraph:

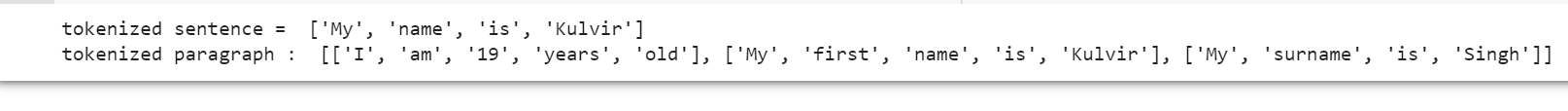
text.append(sentence.split())

print('tokenized paragraph : ',text)

***Code Screenshot:***



***Output Screenshots :***



**Experiment 2**

*Question 3*

***Problem Statement :***

Write a program (using nltk toolkit in python environment) to tokenize

a) Sentence

b) A paragraph

***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 a file using open method and store its contents in a variable using read method. Create a variable which stores the stopwords that are fetched from nltk. Use the tokenize method to create tokens of the document/file read. Loop through the tokens and filter out the stopwords. Display the tokens and filtered paragraph.

***Code :***

!pip install nltk

import nltk

nltk.download('stopwords')

nltk.download('punkt')

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

f = open('demo.txt','r')

paragraph = f.read()

stop\_words = set(stopwords.words('english'))

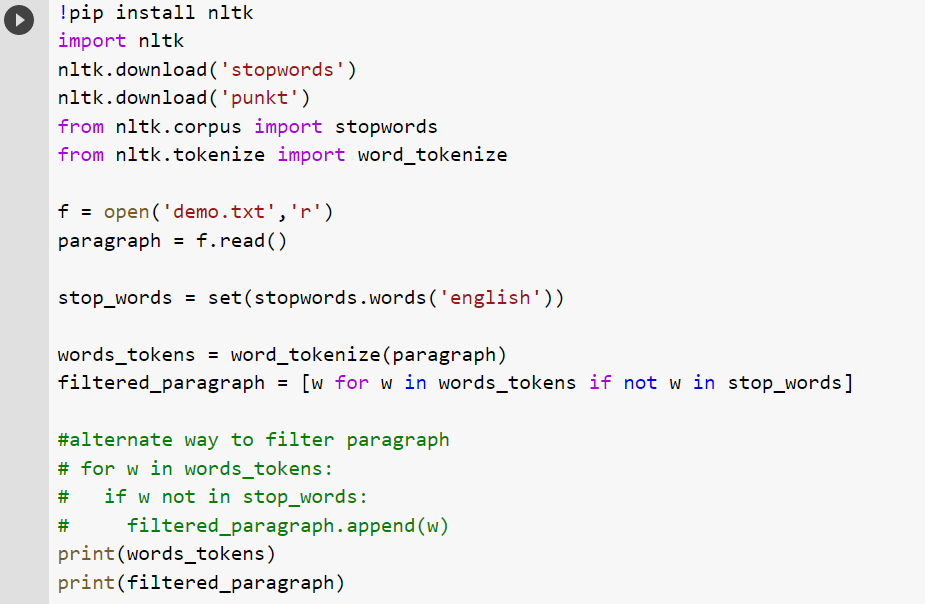
words\_tokens = word\_tokenize(paragraph)

filtered\_paragraph = [w for w in words\_tokens if not w in stop\_words]

print(words\_tokens)

print(filtered\_paragraph)

***Code Screenshot:***



***Output Screenshots :***

