**Web Mining (CSE3024)**

**Lab Assignment 1**

|  |  |
| --- | --- |
| Name: | Kritika Mishra |
| Registration Number: | 16BCI0041 |
| Slot: | L15+L16 |
| Faculty: | Lokesh Kumar R |

1. **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']**

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

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

text = "Hello Adam, how are you? I hope everything is going well. Today is a good day, see you dude."

f\_text = []

tokens = word\_tokenize(text)

for i in tokens:

if i in stop\_words:

print("Found!")

else:

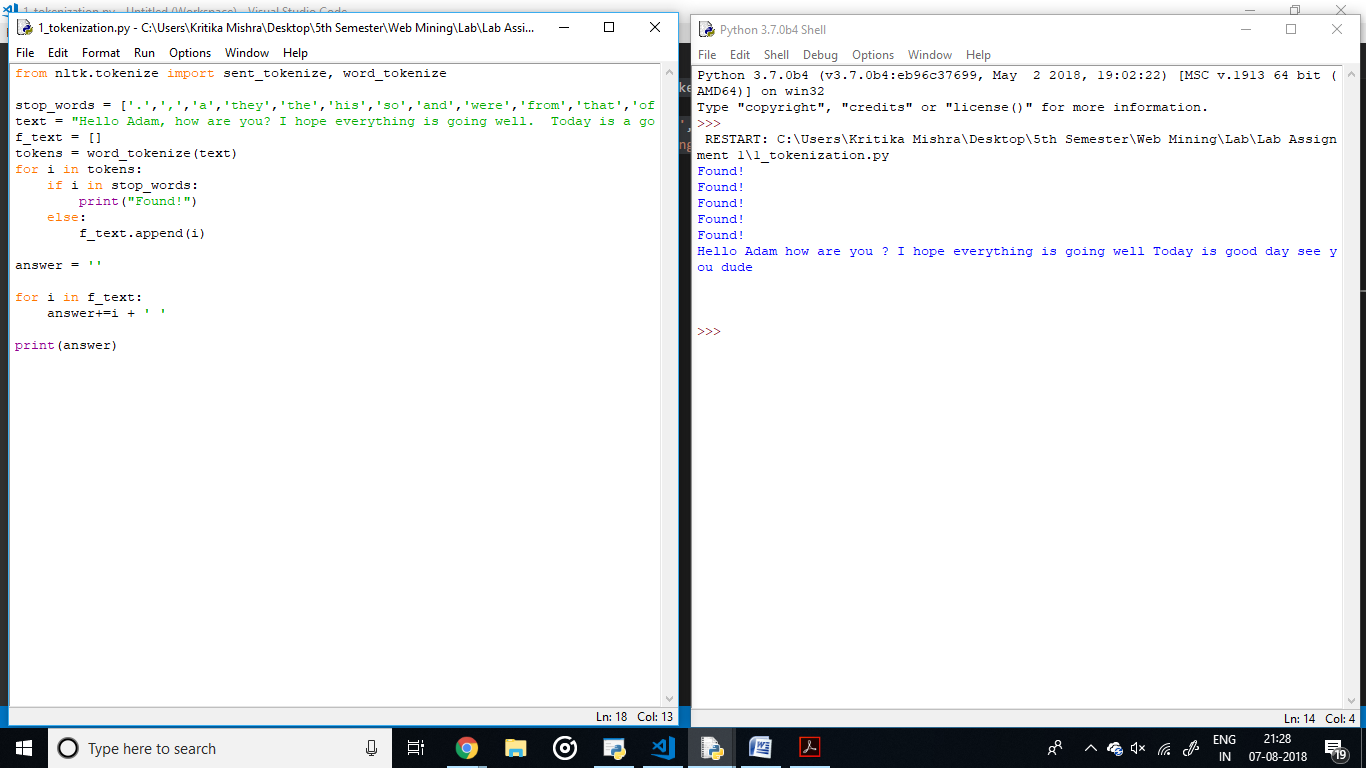
f\_text.append(i)

answer = ''

for i in f\_text:

answer+=i + ' '

print(answer)



1. **Write a program to tokenize a) A sentence b) Multiple sentences**

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

example\_sent = "This is a sample sentence, showing off the stop words filtration."

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

word\_tokens = word\_tokenize(example\_sent)

filtered\_sentence = []

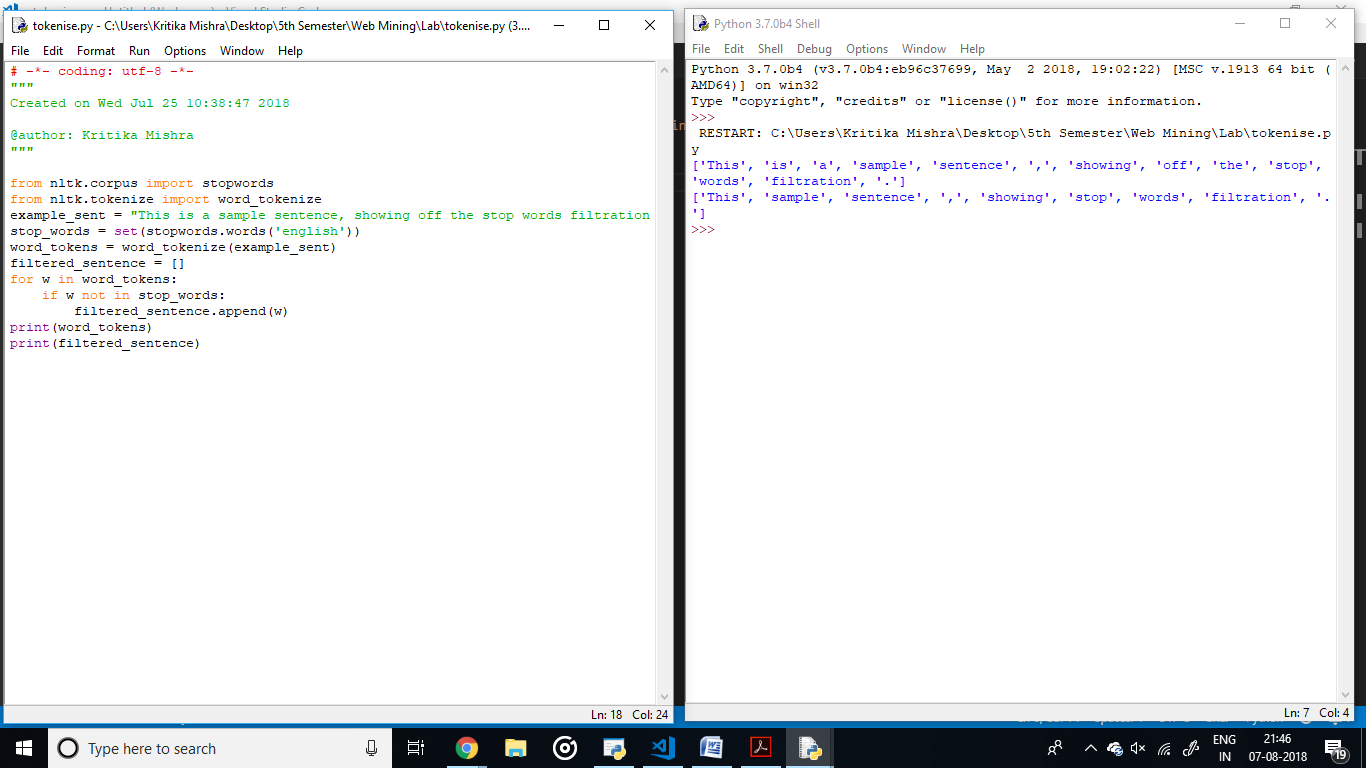
for w in word\_tokens:

if w not in stop\_words:

filtered\_sentence.append(w)

print(word\_tokens)

print(filtered\_sentence)



from nltk.stem import PorterStemmer

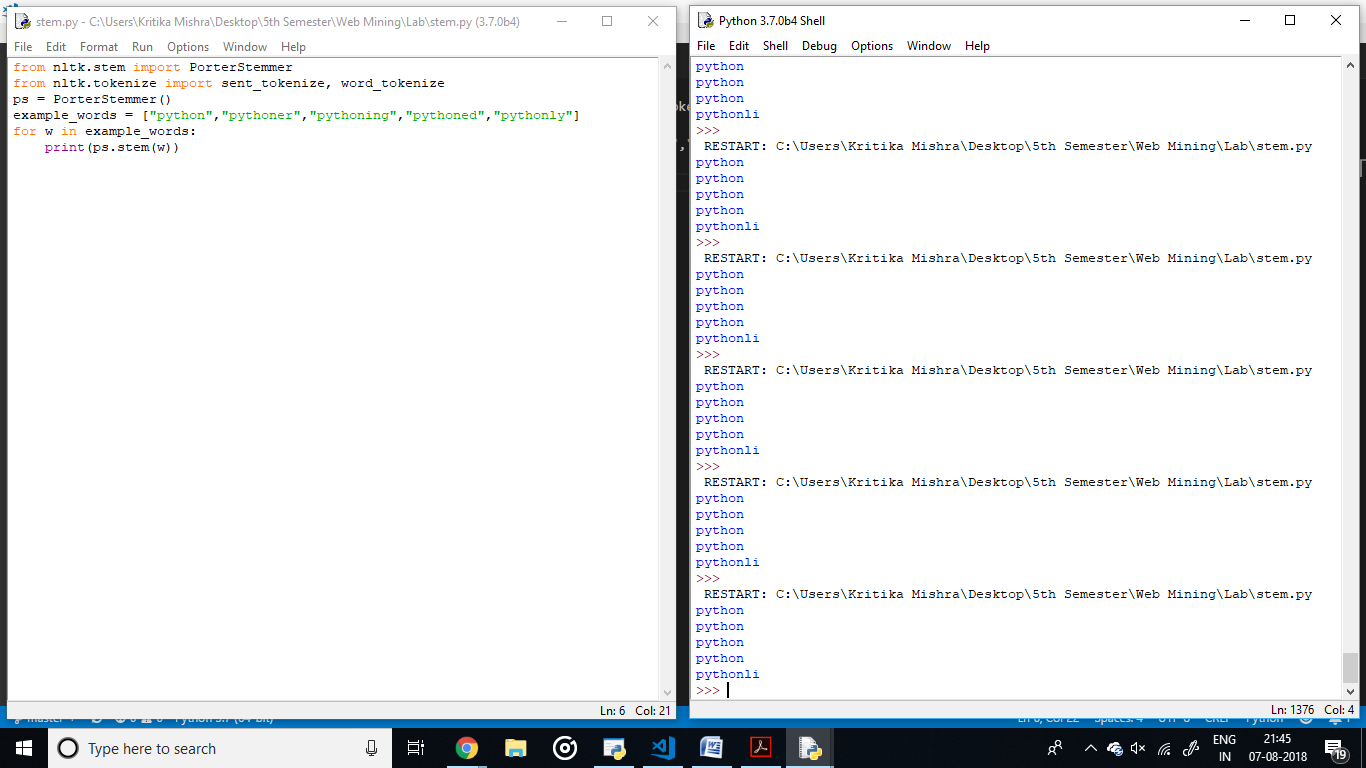
from nltk.tokenize import sent\_tokenize, word\_tokenize

ps = PorterStemmer()

example\_words = ["python","pythoner","pythoning","pythoned","pythonly"]

for w in example\_words:

print(ps.stem(w))



from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

example\_sent = "Let me start with the most cliché but true line, I have no idea what I’d do without you. I guess people say this after 10-15 years of friendship but for us 2 were enough for me. I can’t think of college without you and neither had I ever imagined to find a human being so much like me and here I emphasise on the same pointer in the mean scale. I know when you say you understand you do because you are probably the one who can step in the same shoes.\nIts not for the first time that I am trying to make an attempt to write something about you/to you. I have failed in all my attempts (P.S: There have been many) earlier like I am kind of right now because I know everything thing I can think also crosses your head at some point or the other so you know it and I don’t have to tell you again.\nBut let me take this opportunity to make a proposal of officially starting our mutual admiration society.\nI bet you laughed this time. See I am funnier.\nOh my god! I sound so much in love. Up for it in the new room? Haha, kidding child.\nThe point remains the same I love you and you are very important to me and I promise I won’t ever forget your birthday and keep in touch after college and attend your wedding too (smoked up probably if it is at 6AM).\nI cannot list the little things that make you so special whether it being you getting food for me or just listening to me or being a mom when needed. To top this cheese burst paragraph with a cheesy dip I would like to remind you of the countless secret dates (specially the one in Haze Cafe last fall) and washing clothes together and to motivating each other I have learnt a lot from you.\nThank you so much for standing by me always.\nI love you a lot.\nA very happy 20th."

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

word\_tokens = word\_tokenize(example\_sent)

filtered\_sentence = []

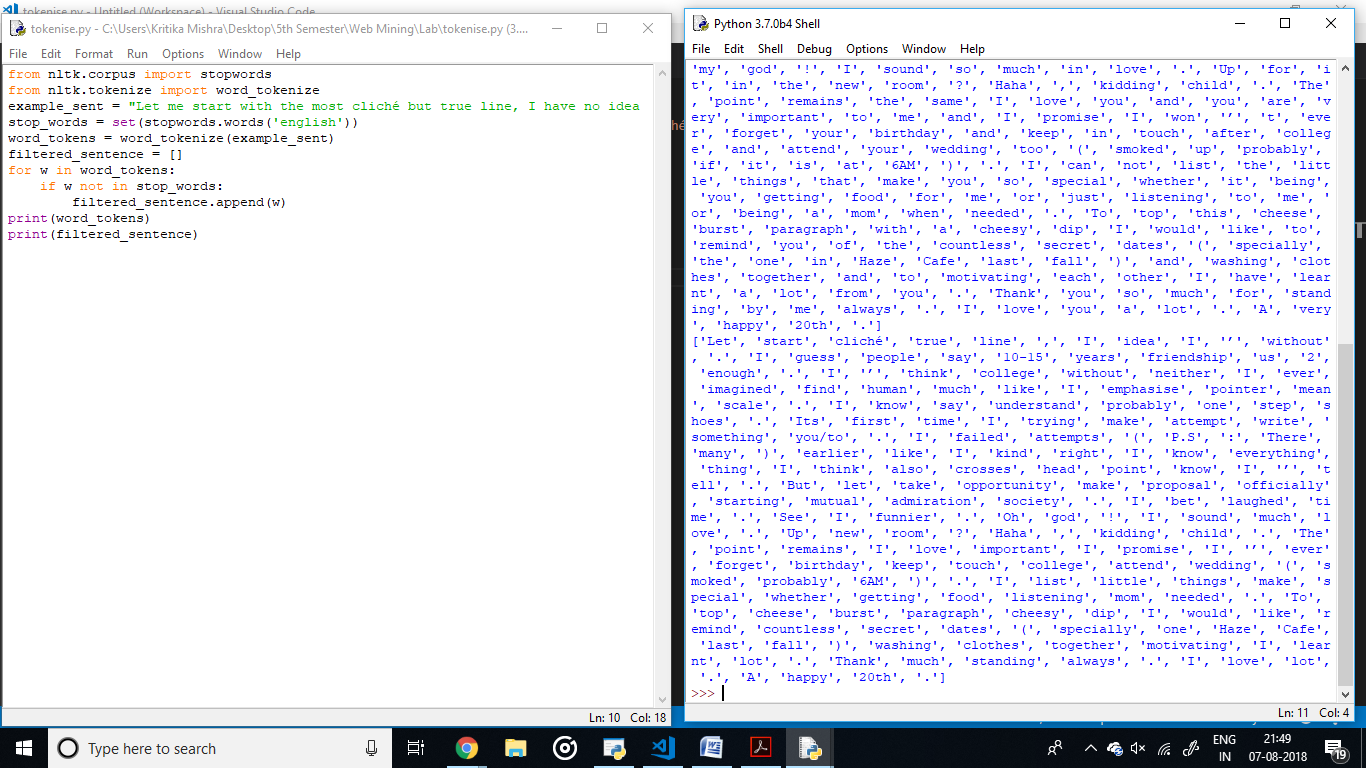
for w in word\_tokens:

if w not in stop\_words:

filtered\_sentence.append(w)

print(word\_tokens)

print(filtered\_sentence)



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

**a) Sentence**

**b) Multiple sentences**

**c) A paragraph**

**d) Information of a complete web page**

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

import urllib.request

from bs4 import BeautifulSoup

url = "https://python.org"

html = urllib.request.urlopen(url)

soup = BeautifulSoup(html,"lxml")

for script in soup(["script", "style"]):

script.extract() # rip it out

text = soup.get\_text()

tokens = word\_tokenize(text)

text\_2 = []

stopwords\_total = 0

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

for i in tokens:

if i in stop\_words:

stopwords\_total+=1

else:

text\_2.append(i)

print(text\_2)

