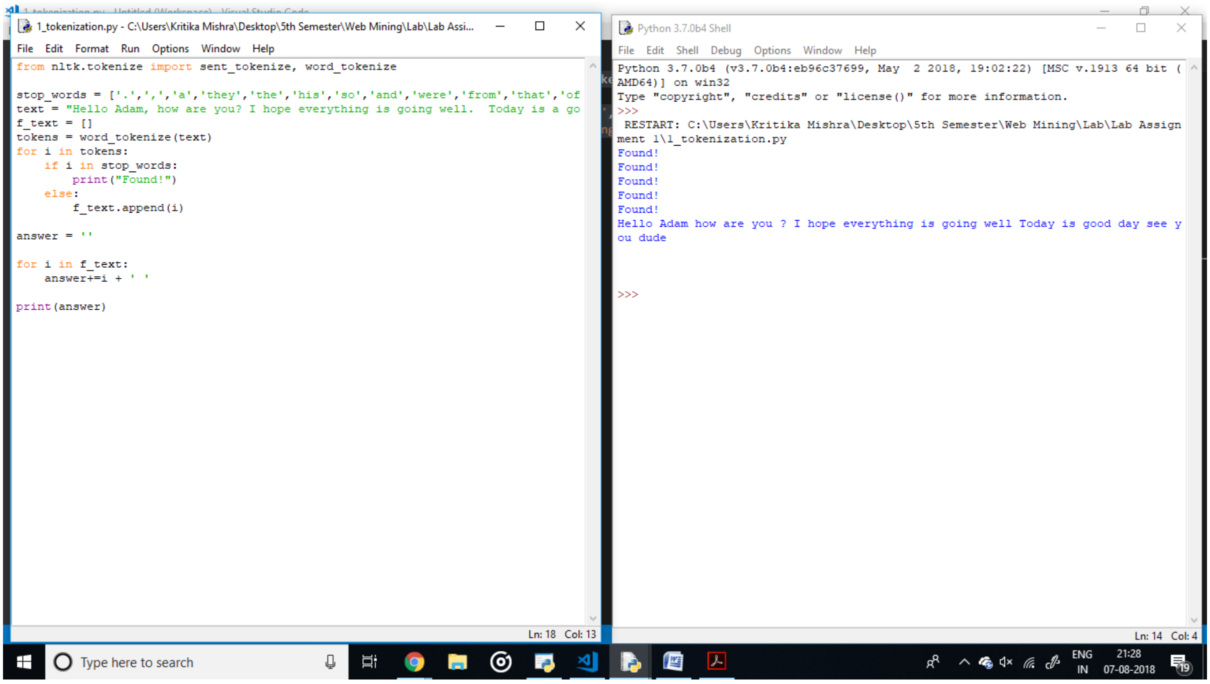
Web Mining (CSE3024)

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_tokenization.py - C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\Lab Assi...
File Edit Format Run Options Window Help

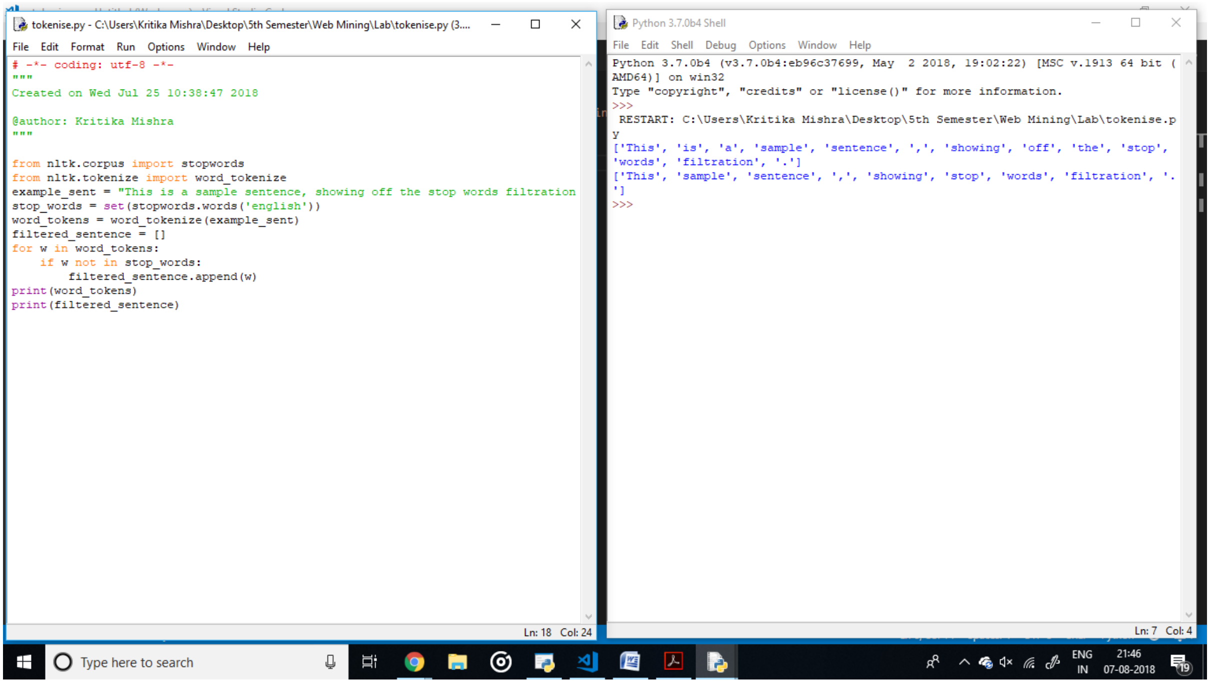
```
from nltk.tokenize import sent_tokenize, word_tokenize

stop_words = ['.',',','a','they','the','his','so','and','were','from','that','of
text = "Hello Adam, how are you? I hope everything is going well.  Today is a go
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)
```

Ln: 18 Col: 13
Python 3.7.0b4 Shell
File Edit Shell Debug Options Window Help

```
Python 3.7.0b4 (v3.7.0b4:eb96c37699, May  2 2018, 19:02:22) [MSC v.1913 64 bit (
AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
 RESTART: C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\Lab Assign
ment 1\1_tokenization.py
Found!
Found!
Found!
Found!
Found!
Hello Adam how are you ? I hope everything is going well Today is good day see y
ou dude



>>>
```

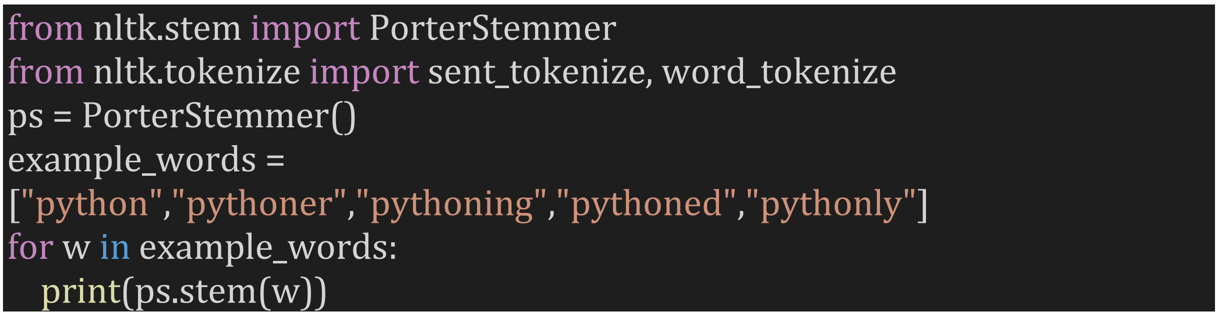
2. Write a program to tokenize a) A sentence b) Multiple sentences

```
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)
```

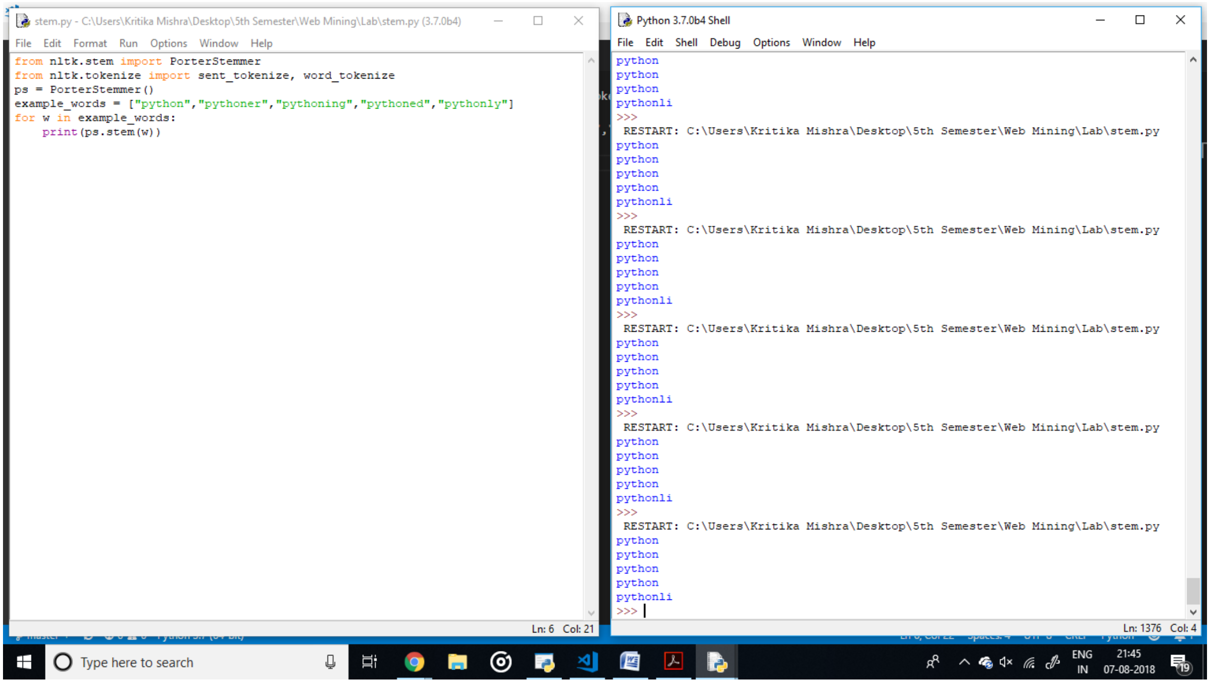

tokenise.py - C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\tokenise.py (3....
File Edit Format Run Options Window Help
Ln: 18 Col: 24
Python 3.7.0b4 Shell
File Edit Shell Debug Options Window Help
Ln: 7 Col: 4
Type here to search
ENG IN 21:46 07-08-2018

```
# -*- coding: utf-8 -*-
"""
Created on Wed Jul 25 10:38:47 2018

@author: Kritika Mishra
"""

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)
```

```
Python 3.7.0b4 (v3.7.0b4:eb96c37699, May  2 2018, 19:02:22) [MSC v.1913 64 bit (
AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
 RESTART: C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\tokenise.p
y
['This', 'is', 'a', 'sample', 'sentence', ',', 'showing', 'off', 'the', 'stop',
'words', 'filtration', '.']
['This', 'sample', 'sentence', ',', 'showing', 'stop', 'words', 'filtration', '.
']
>>>
```

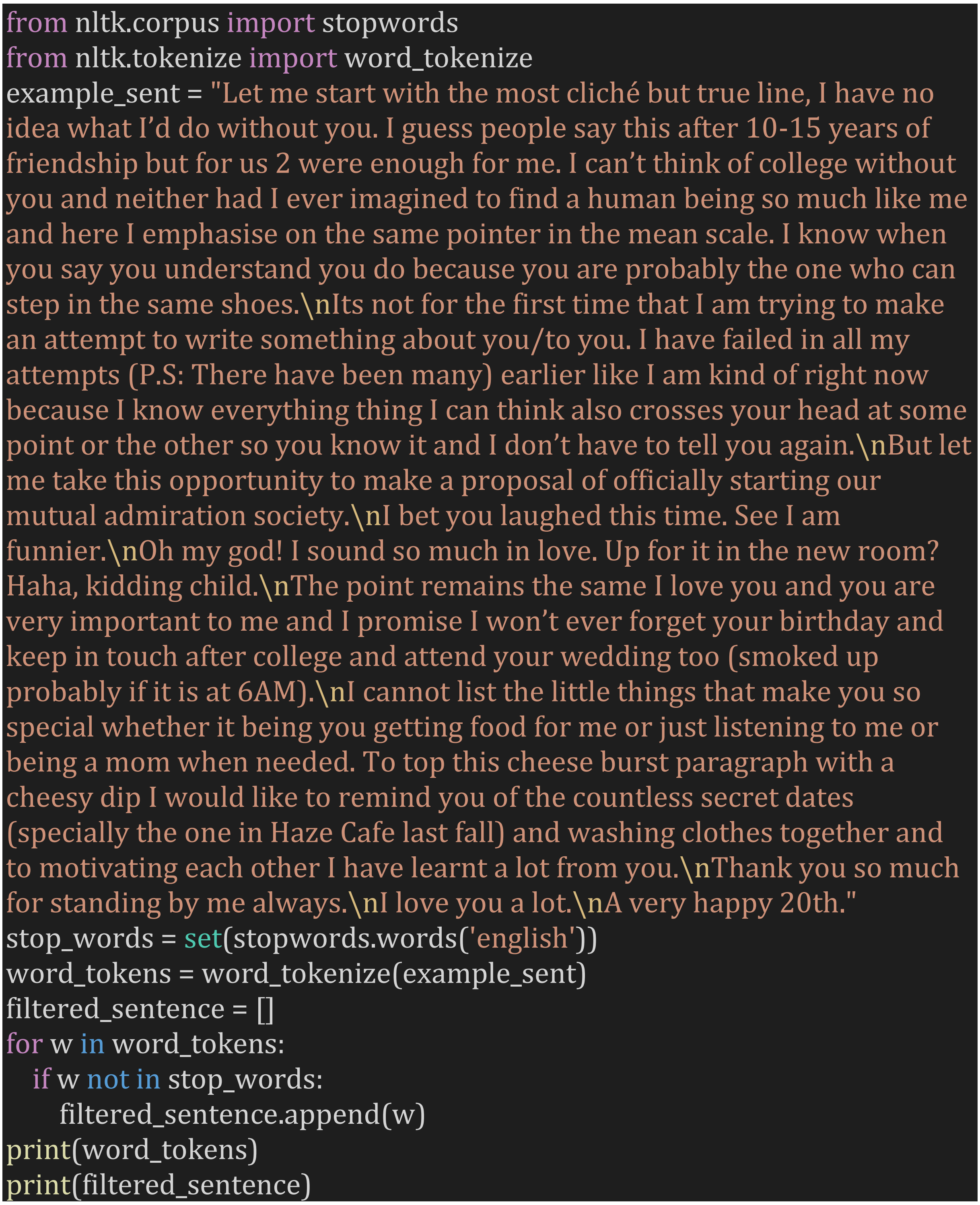


```
from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize
ps = PorterStemmer()
example_words =
["python","pythoner","pythoning","pythoned","pythonly"]
for w in example_words:
   print(ps.stem(w))
```

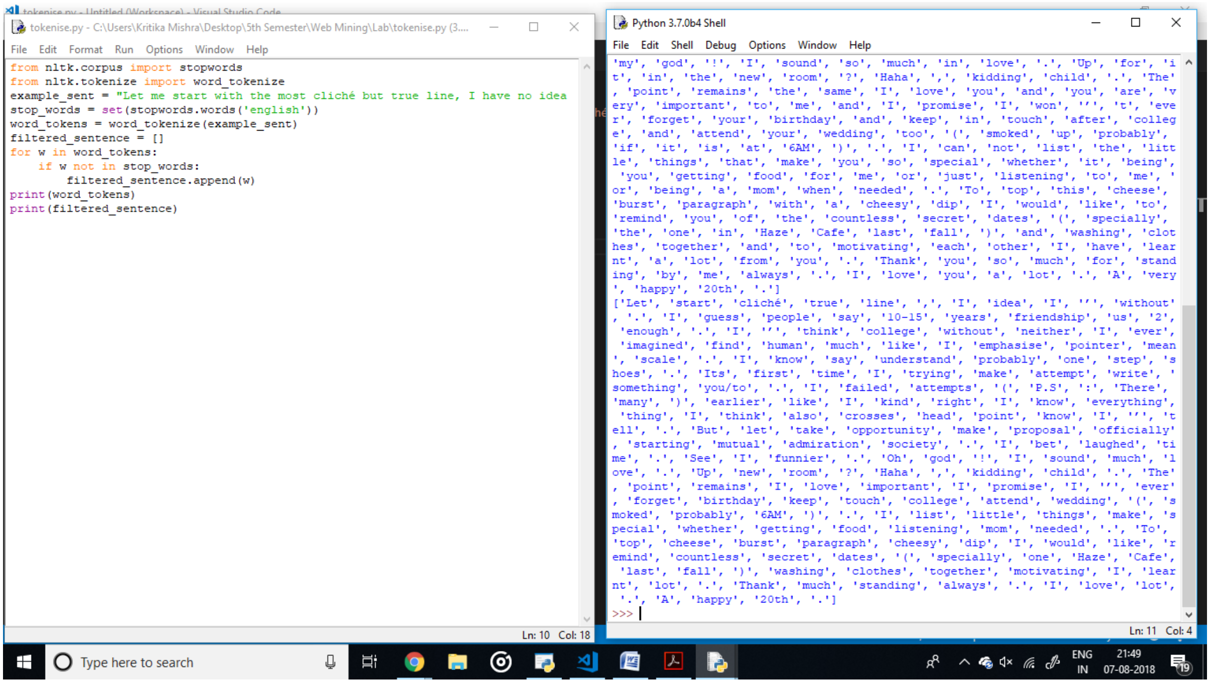

stem.py - C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\stem.py (3.7.0b4)
File Edit Format Run Options Window Help
Ln: 6 Col: 21
Python 3.7.0b4 Shell
File Edit Shell Debug Options Window Help
Ln: 1376 Col: 4
Type here to search
ENG IN 21:45 07-08-2018

```
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))
```

```
python
python
python
pythonli
>>>
 RESTART: C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\stem.py
python
python
python
python
pythonli
>>>
 RESTART: C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\stem.py
python
python
python
python
pythonli
>>>
 RESTART: C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\stem.py
python
python
python
python
pythonli
>>>
 RESTART: C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\stem.py
python
python
python
python
pythonli
>>>
 RESTART: C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\stem.py
python
python
python
python
pythonli
>>>
```



```
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)
```



tokenise.py - Untitled (Workspace) - Visual Studio Code
tokenise.py - C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\tokenise.py (3....
File Edit Format Run Options Window Help

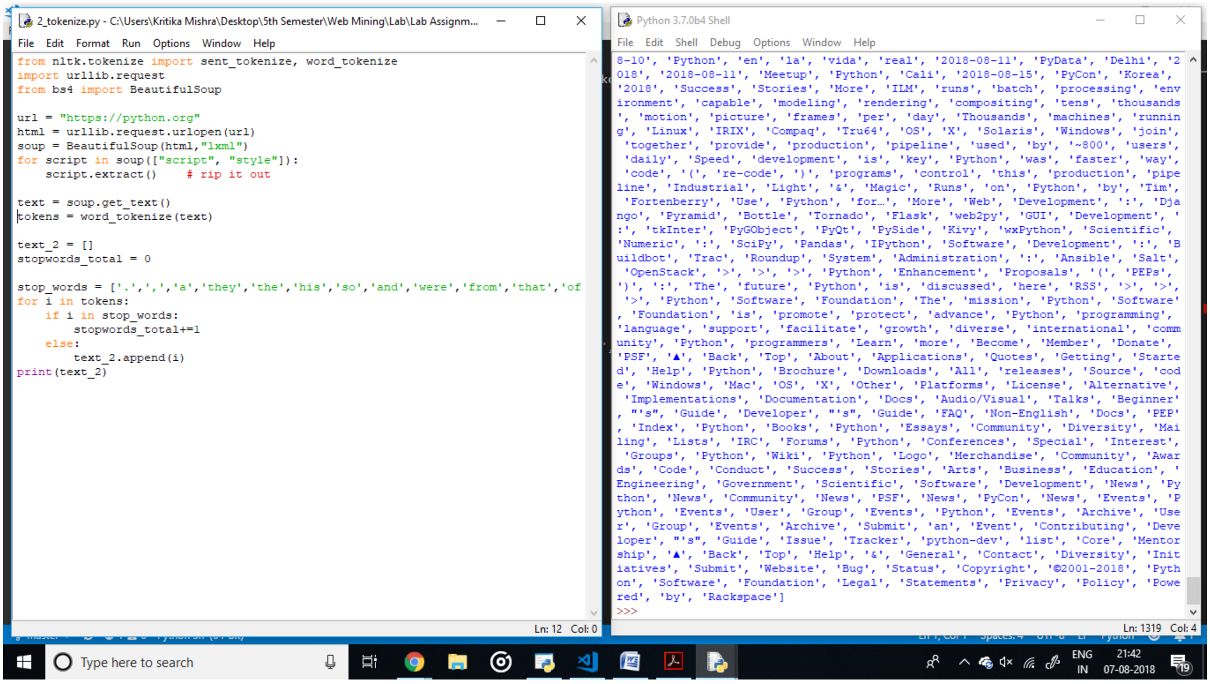
```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
example_sent = "Let me start with the most cliché but true line, I have no idea
stop_words = set(stopwords.words('english'))
word_tokens = word_tokenize(example_sent)
filtered_sentence = []
for w in word_tokens:
    if w not in stop_words:
        filtered_sentence.append(w)
print(word_tokens)
print(filtered_sentence)
```

Ln: 10 Col: 18
Python 3.7.0b4 Shell
File Edit Shell Debug Options Window Help

```
'my', 'god', '!', 'I', 'sound', 'so', 'much', 'in', 'love', '.', 'Up', 'for', 'i
t', 'in', 'the', 'new', 'room', '?', 'Haha', ',', 'kidding', 'child', '.', 'The'
, 'point', 'remains', 'the', 'same', 'I', 'love', 'you', 'and', 'you', 'are', 'v
ery', 'important', 'to', 'me', 'and', 'I', 'promise', 'I', 'won', ''', 't', 'eve
r', 'forget', 'your', 'birthday', 'and', 'keep', 'in', 'touch', 'after', 'colleg
e', 'and', 'attend', 'your', 'wedding', 'too', '(', 'smoked', 'up', 'probably',
'if', 'it', 'is', 'at', '6AM', ')', '.', 'I', 'can', 'not', 'list', 'the', 'litt
le', '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', 'clot
hes', 'together', 'and', 'to', 'motivating', 'each', 'other', 'I', 'have', 'lear
nt', 'a', 'lot', 'from', 'you', '.', 'Thank', 'you', 'so', 'much', 'for', 'stand
ing', 'by', 'me', 'always', '.', 'I', 'love', 'you', 'a', 'lot', '.', 'A', 'very
', 'happy', '20th', '.']
['Let', 'start', 'cliché', 'true', 'line', ',', 'I', 'idea', 'I', ''', 'without'
, '.', 'I', 'guess', 'people', 'say', '10-15', 'years', 'friendship', 'us', '2',
 'enough', '.', 'I', ''', 'think', 'college', 'without', 'neither', 'I', 'ever',
 'imagined', 'find', 'human', 'much', 'like', 'I', 'emphasise', 'pointer', 'mean
', 'scale', '.', 'I', 'know', 'say', 'understand', 'probably', 'one', 'step', 's
hoes', '.', 'Its', 'first', 'time', 'I', 'trying', 'make', 'attempt', 'write', '
something', 'you/to', '.', 'I', 'failed', 'attempts', '(', 'P.S', ':', 'There',
'many', ')', 'earlier', 'like', 'I', 'kind', 'right', 'I', 'know', 'everything',
 'thing', 'I', 'think', 'also', 'crosses', 'head', 'point', 'know', 'I', ''', 't
ell', '.', 'But', 'let', 'take', 'opportunity', 'make', 'proposal', 'officially'
, 'starting', 'mutual', 'admiration', 'society', '.', 'I', 'bet', 'laughed', 'ti
me', '.', 'See', 'I', 'funnier', '.', 'Oh', 'god', '!', 'I', 'sound', 'much', 'l
ove', '.', 'Up', 'new', 'room', '?', 'Haha', ',', 'kidding', 'child', '.', 'The'
, 'point', 'remains', 'I', 'love', 'important', 'I', 'promise', 'I', ''', 'ever'
, 'forget', 'birthday', 'keep', 'touch', 'college', 'attend', 'wedding', '(', 's
moked', 'probably', '6AM', ')', '.', 'I', 'list', 'little', 'things', 'make', 's
pecial', 'whether', 'getting', 'food', 'listening', 'mom', 'needed', '.', 'To',
'top', 'cheese', 'burst', 'paragraph', 'cheesy', 'dip', 'I', 'would', 'like', 'r
emind', 'countless', 'secret', 'dates', '(', 'specially', 'one', 'Haze', 'Cafe',
 'last', 'fall', ')', 'washing', 'clothes', 'together', 'motivating', 'I', 'lear
nt', 'lot', '.', 'Thank', 'much', 'standing', 'always', '.', 'I', 'love', 'lot',
 '.', 'A', 'happy', '20th', '.']
>>> 
```

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)
```



2_tokenize.py - C:\Users\Kritika Mishra\Desktop\5th Semester\Web Mining\Lab\Lab Assignm...
File Edit Format Run Options Window Help
Ln: 12 Col: 0
Python 3.7.0b4 Shell
File Edit Shell Debug Options Window Help
Ln: 1319 Col: 4
Type here to search
ENG IN 21:42 07-08-2018

```
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
for i in tokens:
    if i in stop_words:
        stopwords_total+=1
    else:
        text_2.append(i)
print(text_2)
```

```
8-10', 'Python', 'en', 'la', 'vida', 'real', '2018-08-11', 'PyData', 'Delhi', '2
018', '2018-08-11', 'Meetup', 'Python', 'Cali', '2018-08-15', 'PyCon', 'Korea',
'2018', 'Success', 'Stories', 'More', 'ILM', 'runs', 'batch', 'processing', 'env
ironment', 'capable', 'modeling', 'rendering', 'compositing', 'tens', 'thousands
', 'motion', 'picture', 'frames', 'per', 'day', 'Thousands', 'machines', 'runnin
g', 'Linux', 'IRIX', 'Compaq', 'Tru64', 'OS', 'X', 'Solaris', 'Windows', 'join',
 'together', 'provide', 'production', 'pipeline', 'used', 'by', '~800', 'users',
 'daily', 'Speed', 'development', 'is', 'key', 'Python', 'was', 'faster', 'way',
 'code', '(', 're-code', ')', 'programs', 'control', 'this', 'production', 'pipe
line', 'Industrial', 'Light', '&', 'Magic', 'Runs', 'on', 'Python', 'by', 'Tim',
 'Fortenberry', 'Use', 'Python', 'for…', 'More', 'Web', 'Development', ':', 'Dja
ngo', 'Pyramid', 'Bottle', 'Tornado', 'Flask', 'web2py', 'GUI', 'Development', '
:', 'tkInter', 'PyGObject', 'PyQt', 'PySide', 'Kivy', 'wxPython', 'Scientific',
'Numeric', ':', 'SciPy', 'Pandas', 'IPython', 'Software', 'Development', ':', 'B
uildbot', 'Trac', 'Roundup', 'System', 'Administration', ':', 'Ansible', 'Salt',
 'OpenStack', '>', '>', '>', 'Python', 'Enhancement', 'Proposals', '(', 'PEPs',
')', ':', 'The', 'future', 'Python', 'is', 'discussed', 'here', 'RSS', '>', '>',
 '>', 'Python', 'Software', 'Foundation', 'The', 'mission', 'Python', 'Software'
, 'Foundation', 'is', 'promote', 'protect', 'advance', 'Python', 'programming',
'language', 'support', 'facilitate', 'growth', 'diverse', 'international', 'comm
unity', 'Python', 'programmers', 'Learn', 'more', 'Become', 'Member', 'Donate',
'PSF', '▲', 'Back', 'Top', 'About', 'Applications', 'Quotes', 'Getting', 'Starte
d', 'Help', 'Python', 'Brochure', 'Downloads', 'All', 'releases', 'Source', 'cod
e', 'Windows', 'Mac', 'OS', 'X', 'Other', 'Platforms', 'License', 'Alternative',
 'Implementations', 'Documentation', 'Docs', 'Audio/Visual', 'Talks', 'Beginner'
, "'s", 'Guide', 'Developer', "'s", 'Guide', 'FAQ', 'Non-English', 'Docs', 'PEP'
, 'Index', 'Python', 'Books', 'Python', 'Essays', 'Community', 'Diversity', 'Mai
ling', 'Lists', 'IRC', 'Forums', 'Python', 'Conferences', 'Special', 'Interest',
 'Groups', 'Python', 'Wiki', 'Python', 'Logo', 'Merchandise', 'Community', 'Awar
ds', 'Code', 'Conduct', 'Success', 'Stories', 'Arts', 'Business', 'Education', '
Engineering', 'Government', 'Scientific', 'Software', 'Development', 'News', 'Py
thon', 'News', 'Community', 'News', 'PSF', 'News', 'PyCon', 'News', 'Events', 'P
ython', 'Events', 'User', 'Group', 'Events', 'Python', 'Events', 'Archive', 'Use
r', 'Group', 'Events', 'Archive', 'Submit', 'an', 'Event', 'Contributing', 'Deve
loper', "'s", 'Guide', 'Issue', 'Tracker', 'python-dev', 'list', 'Core', 'Mentor
ship', '▲', 'Back', 'Top', 'Help', '&', 'General', 'Contact', 'Diversity', 'Init
iatives', 'Submit', 'Website', 'Bug', 'Status', 'Copyright', '©2001-2018', 'Pyth
on', 'Software', 'Foundation', 'Legal', 'Statements', 'Privacy', 'Policy', 'Powe
red', 'by', 'Rackspace']
>>>
```