VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

Register Number: 18BCE0745

Name: Gourishetty Sreemanth

Code

```
# -*- coding: utf-8 -*-
"""
Created on Thu Jul 30 16:07:35 2020

@author: Sreemanth
"""



import nltk
#nltk.download('wordnet')
#nltk.download('averaged_perceptron_tagger')
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import sent_tokenize, word_tokenize
```

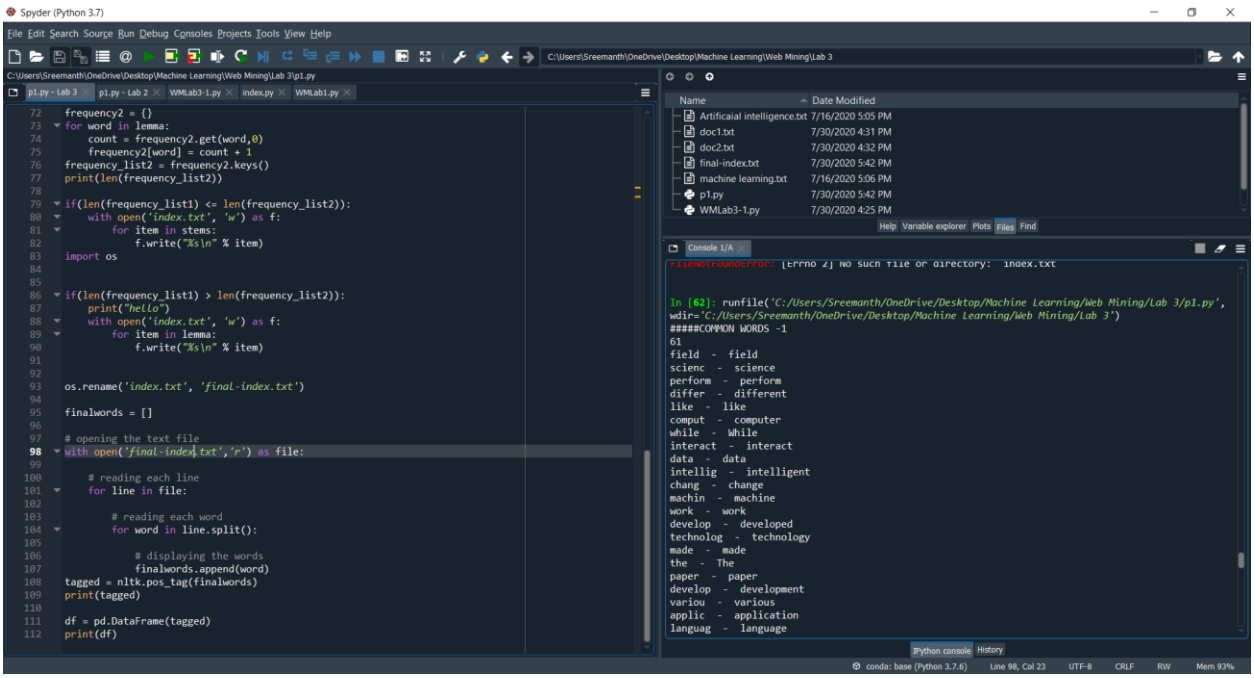
```
import pandas as pd
import re
stop_words = set(stopwords.words('english'))

print("#####COMMON WORDS -1")
words1 = []
#Group in a list the words common for two text files and show their total count
f1 = open("Artificaial intelligence.txt").readlines()
f2 = open("machine learning.txt").readlines()
if len(f1) != 0 | len(f2) != 0:
    uniq1 = set(words for line in f1 for words in line.strip().split())
    uniq2 = set(wordss for lines in f2 for wordss in lines.strip().split())
    for words in uniq1:
        for wordds in uniq2:
            if words == wordds:
                words1.append(words);



words1 = [w for w in words1 if not w in stop_words]
print(len(words1))

with open('index.txt', 'w') as f:
    for item in words1:
        f.write("%s\n" % item)

readwords = []

# opening the text file
with open('index.txt','r') as file:
```

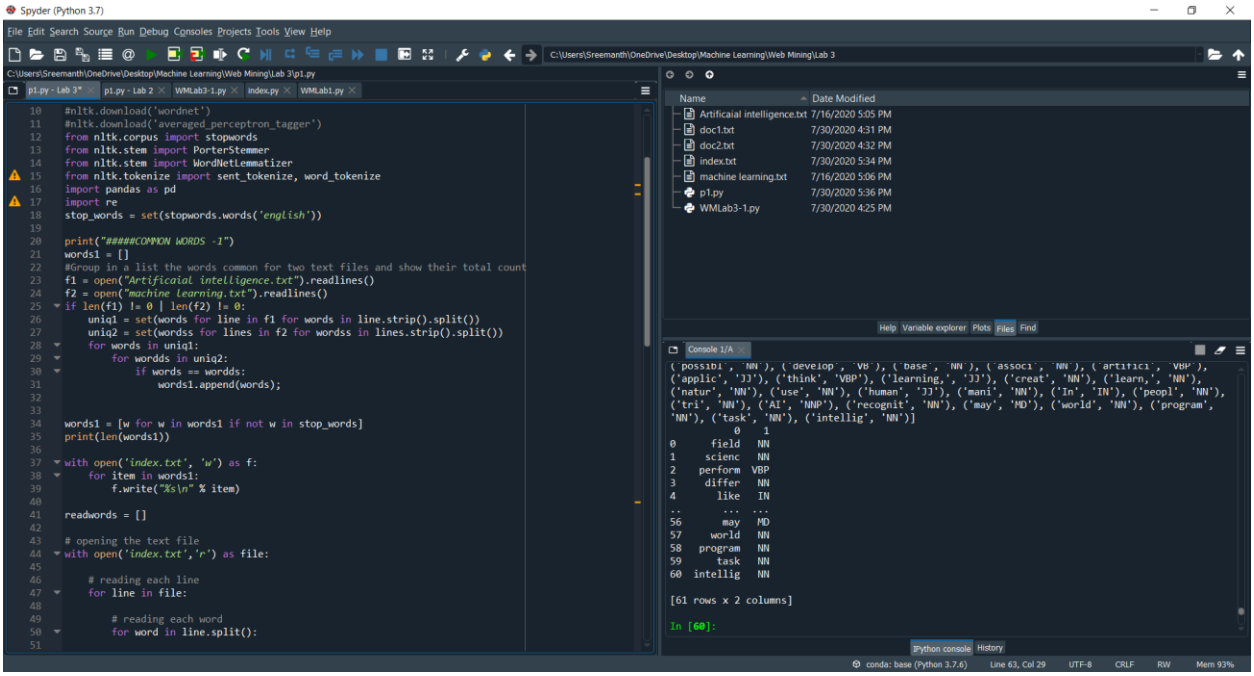
```
    # reading each line
    for line in file:

        # reading each word
        for word in line.split():

            # displaying the words
            readwords.append(word)



ps = PorterStemmer()
lemmatizer = WordNetLemmatizer()
stems = []
lemma = []
for w in readwords:
    print(ps.stem(w), " - ", lemmatizer.lemmatize(w))
    stems.append(ps.stem(w))
    lemma.append(lemmatizer.lemmatize(w))

frequency1 = {}
for word in stems:
    count = frequency1.get(word,0)
    frequency1[word] = count + 1
frequency_list1 = frequency1.keys()
print(len(frequency_list1))

frequency2 = {}
for word in lemma:
```

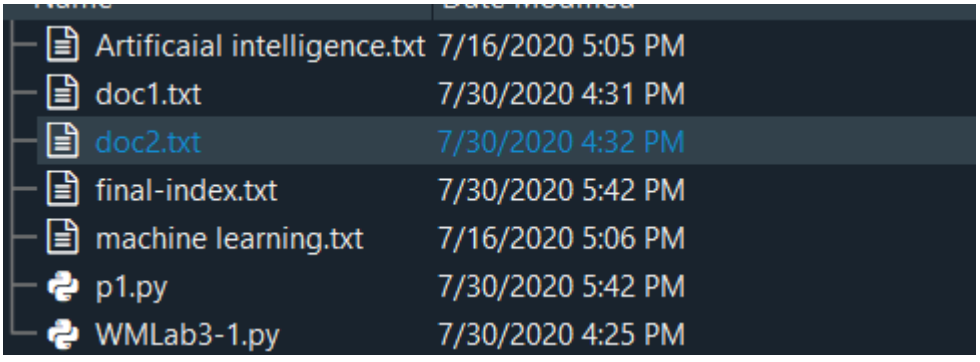
```
    count = frequency2.get(word,0)
    frequency2[word] = count + 1
frequency_list2 = frequency2.keys()
print(len(frequency_list2))

if(len(frequency_list1) <= len(frequency_list2)):
    with open('index.txt', 'w') as f:
        for item in stems:
            f.write("%s\n" % item)
import os



if(len(frequency_list1) > len(frequency_list2)):
    print("hello")
    with open('index.txt', 'w') as f:
        for item in lemma:
            f.write("%s\n" % item)



os.rename('index.txt', 'final-index.txt')

finalwords = []

# opening the text file
with open('index.txt','r') as file:

    # reading each line
    for line in file:
```

```
        # reading each word
        for word in line.split():


            # displaying the words
            finalwords.append(word)
tagged = nltk.pos_tag(finalwords)
print(tagged)


df = pd.DataFrame(tagged)
print(df)
```

Output




Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\Sreemanth\OneDrive\Desktop\Machine Learning\Web Mining\Lab 3
C:\Users\Sreemanth\OneDrive\Desktop\Machine Learning\Web Mining\Lab 3\p1.py
p1.py - Lab 3*
p1.py - Lab 2
WMLab3-1.py
index.py
WMLab1.py
#nltk.download('wordnet')
#nltk.download('averaged_perceptron_tagger')
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import sent_tokenize, word_tokenize
import pandas as pd
import re
stop_words = set(stopwords.words('english'))
print("#####COMMON WORDS -1")
words1 = []
#Group in a list the words common for two text files and show their total count
f1 = open("Artificaial intelligence.txt").readlines()
f2 = open("machine learning.txt").readlines()
if len(f1) != 0 | len(f2) != 0:
uniq1 = set(words for line in f1 for words in line.strip().split())
uniq2 = set(wordss for lines in f2 for wordss in lines.strip().split())
for words in uniq1:
for wordds in uniq2:
if words == wordds:
words1.append(words);
words1 = [w for w in words1 if not w in stop_words]
print(len(words1))
with open('index.txt', 'w') as f:
for item in words1:
f.write("%s\n" % item)
readwords = []
# opening the text file
with open('index.txt','r') as file:
# reading each line
for line in file:
# reading each word
for word in line.split():
Name
Date Modified
Artificaial intelligence.txt 7/16/2020 5:05 PM
doc1.txt 7/30/2020 4:31 PM
doc2.txt 7/30/2020 4:32 PM
index.txt 7/30/2020 5:34 PM
machine learning.txt 7/16/2020 5:06 PM
p1.py 7/30/2020 5:36 PM
WMLab3-1.py 7/30/2020 4:25 PM
Help
Variable explorer
Plots
Files
Find
Console 1/A
('possibl', 'NN'), ('develop', 'VB'), ('base', 'NN'), ('associ', 'NN'), ('artifici', 'VBP'),
('applic', 'JJ'), ('think', 'VBP'), ('learning,', 'JJ'), ('creat', 'NN'), ('learn,', 'NN'),
('natur', 'NN'), ('use', 'NN'), ('human', 'JJ'), ('mani', 'NN'), ('In', 'IN'), ('peopl', 'NN'),
('tri', 'NN'), ('AI', 'NNP'), ('recognit', 'NN'), ('may', 'MD'), ('world', 'NN'), ('program',
'NN'), ('task', 'NN'), ('intellig', 'NN')]
0 1
0 field NN
1 scienc NN
2 perform VBP
3 differ NN
4 like IN
.. ..
56 may MD
57 world NN
58 program NN
59 task NN
60 intellig NN
[61 rows x 2 columns]
In [60]:
IPython console
History
conda: base (Python 3.7.6)
Line 63, Col 29
UTF-8
CRLF
RW
Mem 93%


Artificaial intelligence.txt 7/16/2020 5:05 PM
doc1.txt 7/30/2020 4:31 PM
doc2.txt 7/30/2020 4:32 PM
final-index.txt 7/30/2020 5:42 PM
machine learning.txt 7/16/2020 5:06 PM
p1.py 7/30/2020 5:42 PM
WMLab3-1.py 7/30/2020 4:25 PM

```
 NN ), ( task ,  NN ), ( in
           0    1
0      field   NN
1     scienc   NN
2    perform  VBP
3     differ   NN
4       like   IN
..       ...  ...
56       may   MD
57     world   NN
58   program   NN
59      task   NN
60  intellig   NN

[61 rows x 2 columns]
```