



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("Artificiaial 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 worddds in uniq2:

            if words == worddds:

                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

The screenshot shows the Spyder Python IDE interface. The left pane displays a Python script with the following code:

```

72 frequency2 = {}
73 for word in lemma:
74     count = frequency2.get(word,0)
75     frequency2[word] = count + 1
76 frequency_list2 = frequency2.keys()
77 print(len(frequency_list2))
78
79 if(len(frequency_list1) <= len(frequency_list2)):
80     with open('index.txt', 'w') as f:
81         for item in stems:
82             f.write("%s\n" % item)
83 import os
84
85
86 if(len(frequency_list1) > len(frequency_list2)):
87     print("Hello")
88     with open('index.txt', 'w') as f:
89         for item in lemma:
90             f.write("%s\n" % item)
91
92 os.rename('index.txt', 'final-index.txt')
93
94 finalwords = []
95
96 # opening the text file
97
98 with open('final-index.txt','r') as file:
99
100     # reading each line
101     for line in file:
102
103         # reading each word
104         for word in line.split():
105
106             # displaying the words
107             finalwords.append(word)
108 tagged = nltk.pos_tag(finalwords)
109 print(tagged)
110
111 df = pd.DataFrame(tagged)
112 print(df)

```

The right pane shows a file explorer with the following files and their modification dates:

Name	Date Modified
Artificial 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

The bottom pane shows the console output, which includes an error message and a list of common words:

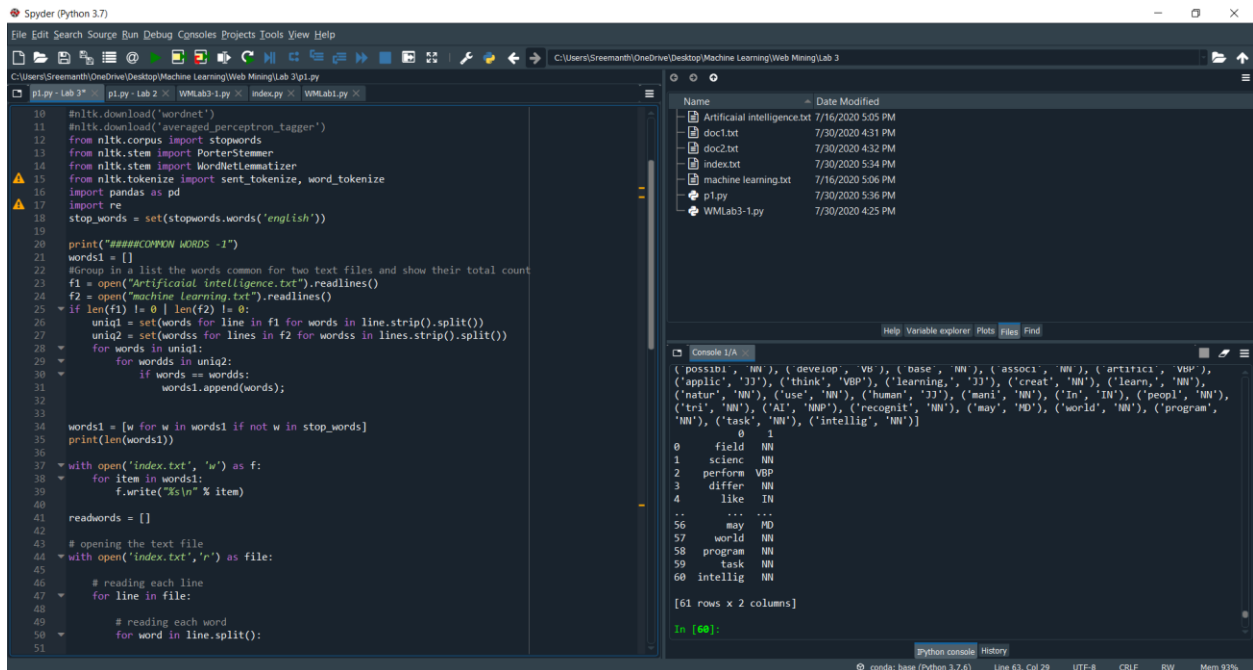
```

FileNotFoundError: [Errno 2] no such file or directory: 'index.txt'

In [42]: runfile('C:/Users/Sreemanth/OneDrive/Desktop/Machine Learning/Web Mining/Lab 3/p1.py',
wdir='C:/Users/Sreemanth/OneDrive/Desktop/Machine Learning/Web Mining/Lab 3')
####COMMON WORDS -1
61
field - field
scienc - science
perform - perform
differ - different
like - like
comput - computer
while - while
interact - interact
data - data
intellig - intelligent
chang - change
machin - machine
work - work
develop - developed
technolog - technology
made - made
the - The
paper - paper
develop - development
variou - various
applic - application
languag - language

```

The status bar at the bottom indicates the environment is conda: base (Python 3.7.4), the current line is 99, column 23, and the encoding is UTF-8.



Name	Date Modified
Artificial 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]
```