IT 563: Data Mining

Lab 6

ID:201811024

Name- Rajat Kumar

Assignment- To make Term Document Matrix from given document file.

Code-

```
from collections import Counter
import re
import numpy as np

D=[]
A=[]
filename='D.txt'

lookup1 = '<TEXT>'
lookup2='</TEXT>'


# Taking the number of line of starting and ending Text from each Document in 1 file.

with open(filename) as myFile:
    for num, line in enumerate(myFile, 1):
        if lookup1 in line or lookup2 in line:
            A.append(num)
```

A - List (10 elements)

Index	Type	Size	Value
0	int	1	20
1	int	1	54
2	int	1	74
3	int	1	109
4	int	1	126
5	int	1	222
6	int	1	239
7	int	1	306
8	int	1	322

 # Appending each line in file to D list

```
with open(filename,'r') as f:
    for line in f:
        D.append(line)
```

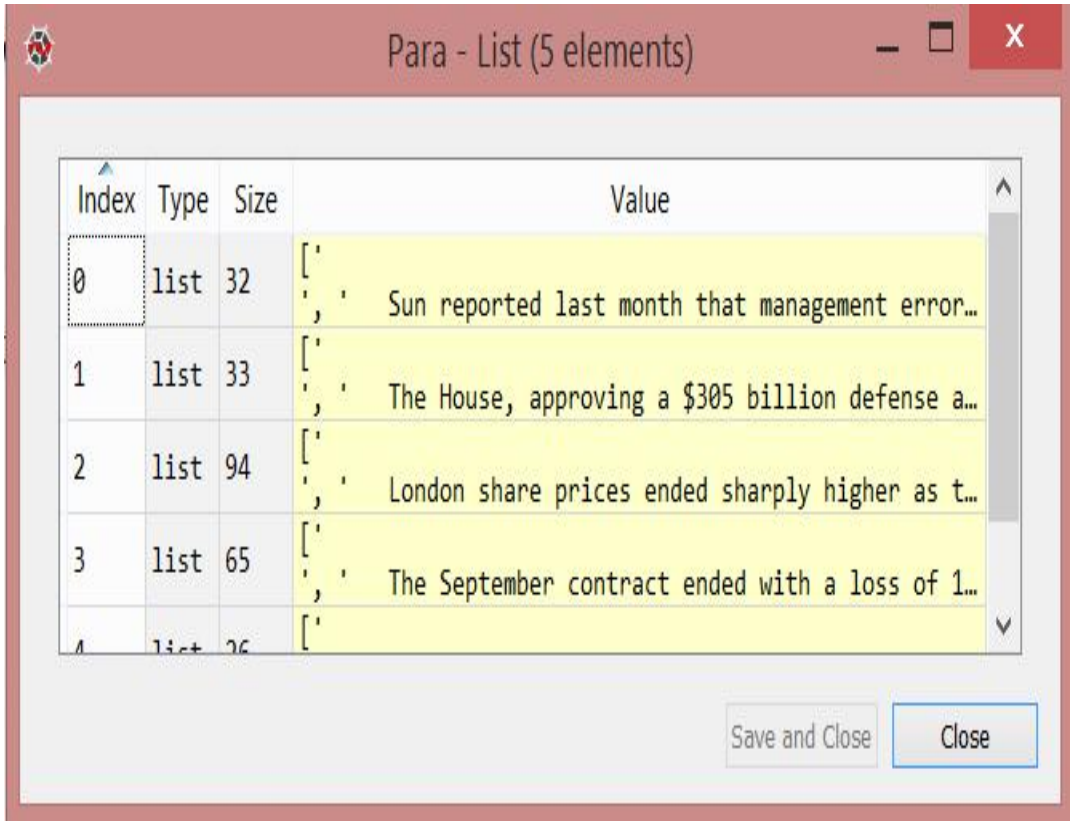
D - List (351 elements)

Index	Type	Size	Value
130	str	1	In Tokyo, the Nikkei was up 269.45 points at 34785.28, breaking the ...
131	str	1	
132	str	1	In midmorning trading Friday, the Nikkei index was up 42.91 points ...
133	str	1	
134	str	1	Positive external market factors bolstered bullishness in the marke ...
135	str	1	The "triple merits" of lower oil prices, a stronger yen and the prospe ...
136	str	1	
137	str	1	The allocation of investment trust fund money through the week will ...
138	str	1	

Save and Close Close

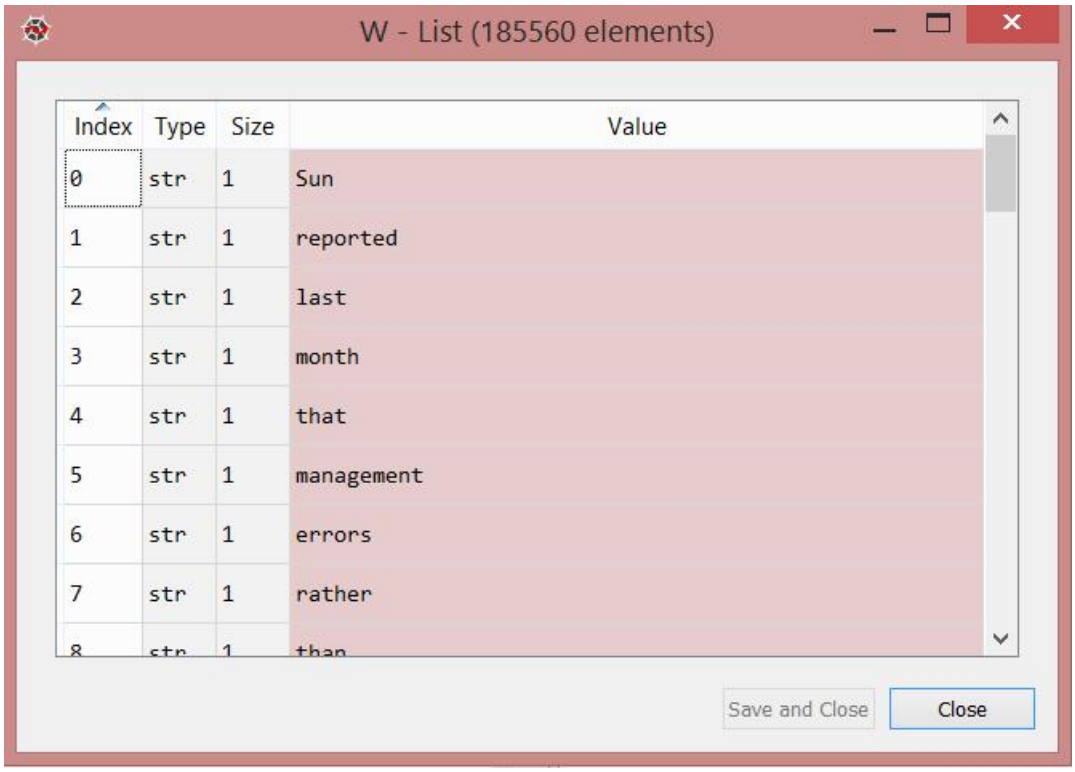
 # Extracting Meaningful text from each document and appending it to Para from the indexes of lines

```
itr=len(A)
Para=[]
i=0
while i<itr:
      Para.append(D[A[i]+1:A[i+1]-1])
      i=i+2
      if i>itr-1:
        Break
```



 # Extracting Meaningful words from Para by using regular expression

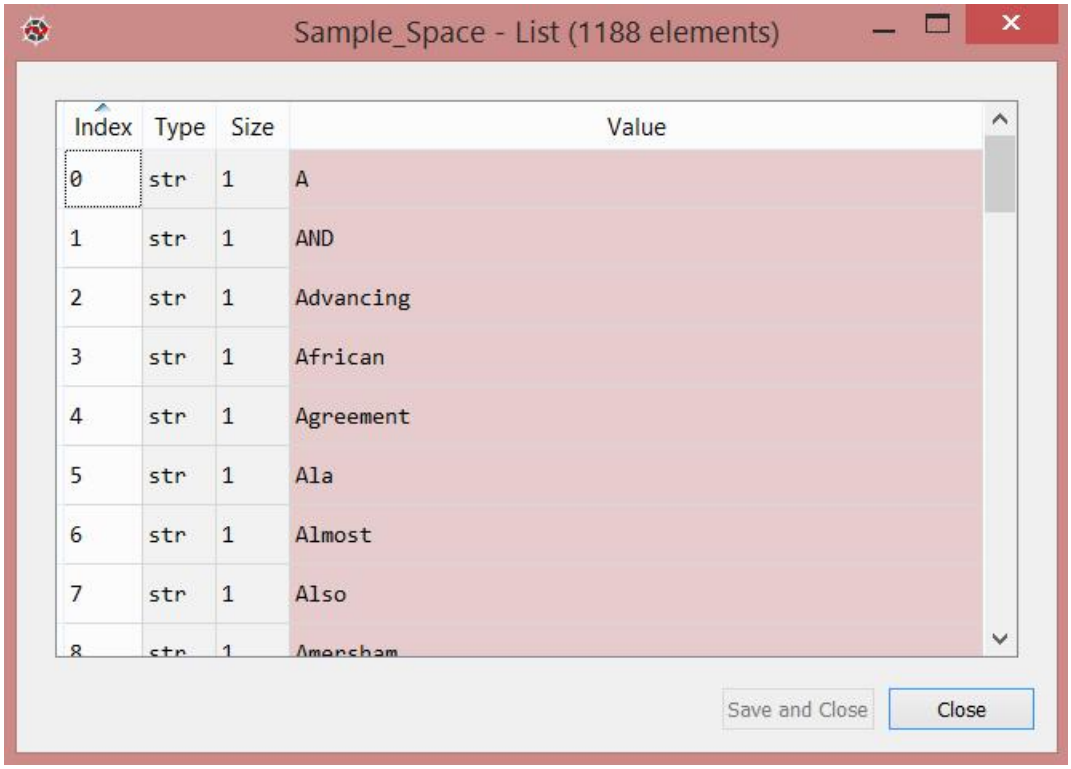
```
W=[]
for i in range(len(Para)):
    for line in Para[i]:
        for j in range(len(Para[i])):
            #print (i,j)
            W=W+(re.findall(r'[A-Za-z]+',Para[i][j]))
```



Index	Type	Size	Value
0	str	1	Sun
1	str	1	reported
2	str	1	last
3	str	1	month
4	str	1	that
5	str	1	management
6	str	1	errors
7	str	1	rather
8	str	1	than

 # Filling unique words in Sample_Space of Words and sorting it

```
counts = Counter(W)
Sample_Space=list(counts.keys())
Sample_Space.sort()
```

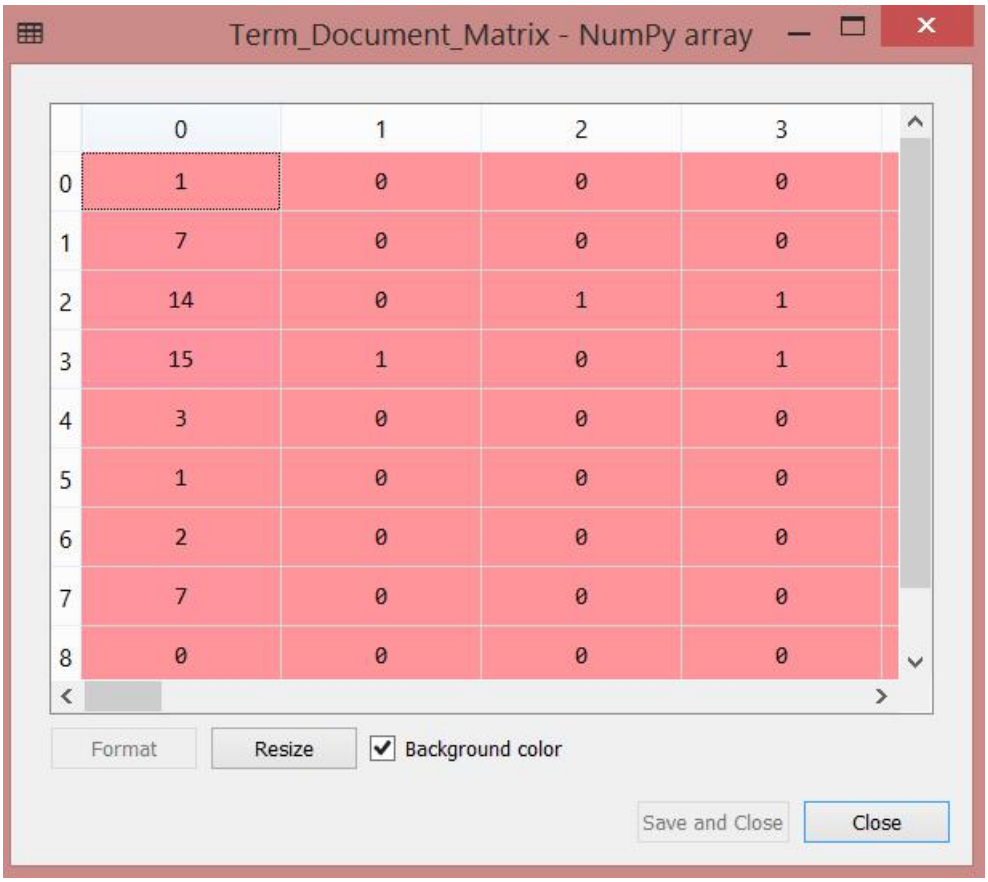


Index	Type	Size	Value
0	str	1	A
1	str	1	AND
2	str	1	Advancing
3	str	1	African
4	str	1	Agreement
5	str	1	Ala
6	str	1	Almost
7	str	1	Also
8	str	1	Amersham

 # Creating Term Document Matrix by counting freuencies of each word in each document

```
M=np.empty([len(Para),len(Sample_Space)])
M=M.astype(int)
str=[]
Final=[]
for i in range(len(Para)):
  str=" ".join(Para[i])
  Final.append(str)
```

```
for i in range(len(Para)):
    for j in range(len(Sample_Space)):
        M[i][j]=Final[i].count(Sample_Space[j])
```



	0	1	2	3
0	1	0	0	0
1	7	0	0	0
2	14	0	1	1
3	15	1	0	1
4	3	0	0	0
5	1	0	0	0
6	2	0	0	0
7	7	0	0	0
8	0	0	0	0

Thanks