Assignment 7

Text Analytics

1. Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.
2. Create representation of document by calculating Term Frequency and Inverse Document Frequency. Provide the codes with outputs and explain everything that you do in this step.

Assignment 7

Import Libraries

In [287…

```
#!pip install Document
```

In [288…

```
#!pip install spacy
#!python -m spacy download Lemmatizer


#!pip install docx
#!pip install python-docx
#!pip install -U spacy

#!python -m spacy download en_core_web_sm
```

In [289…

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
import spacy
nlp = spacy.load('en_core_web_sm')
import document
from docx import Document
import warnings
warnings.filterwarnings('ignore')
import docx2txt
```

In [ ]:

```

```

Load and review data

In [290…

```
path = 'testdoc.docx'
text = docx2txt.process(path)
#print(text)
path2 = 'testdoc2.docx'
text2 = docx2txt.process(path2)
#print(text2)
```

In [291…

```
doc = nlp(text)
# Extract tokens for the given doc
print ([token.text for token in doc])
print(len(doc))
doc2 = nlp(text2)
print(len(text2))
corpus = [doc,doc2]
```

```
['Millions', 'of', 'people', 'in', 'India', 'took', 'part', 'in', 'an', 'annual', 't
ree', 'planting', 'drive', 'Sunday', '.', 'More', 'than', '250', 'million', 'sapling
s', 'were', 'planted', 'in', 'a', 'single', 'day', 'across', 'the', 'country', "'s",
'most', '-', 'populous', 'state', '.', '\n\n', 'The', 'campaign', 'was', 'led', 'b
y', 'Uttar', 'Pradesh', 'state', 'government', 'officials', ',', 'lawmakers', ',',
'and', 'activists', ',', 'in', 'a', 'bid', 'to', 'reduce', 'carbon', 'emissions', 'a
nd', 'combat', 'climate', 'change', '.', '\n\n', 'Where', 'were', 'the', 'trees', 'p
lanted', '?', '\n\n', 'The', 'saplings', 'were', 'planted', 'by', 'volunteers', 'i
n', 'forests', ',', 'farms', ',', 'schools', ',', 'and', 'along', 'riverbanks', 'an
d', 'highways', '.', '\n\n', '"', 'We', 'are', 'committed', 'to', 'increasing', 'th
e', 'forest', 'cover', 'of', 'Uttar', 'Pradesh', 'to', 'over', '15', '%', 'of', 'th
e', 'total', 'land', 'area', 'in', 'the', 'next', 'five', 'years', ',', "'", 'sai
d', 'state', 'forest', 'official', 'Manoj', 'Singh', '.', '\n\n', 'According', 'to',
'another', 'government', 'official', ',', 'the', 'forest', 'cover', 'of', 'the', 'st
ate', 'has', 'increased', 'over', 'the', 'last', 'few', 'years', '.', '\n\n', '"',
'There', 'has', 'been', 'an', 'increase', 'of', '127', 'sqare', 'kilometers', '[',
'79', 'sqare', 'miles', ']', '\xa0', 'in', 'the', 'forest', 'cover', 'in', 'Uttar',
'Pradesh', 'as', 'compared', 'to', '2017', ',', '"', 'a', 'state', 'government', 'sp
okesperson', 'was', 'quoted', 'as', 'saying', 'in', '\xa0', 'The', 'Indian', 'Expres
s', 'newspaper', '.', '\n\n', '"', 'There', 'has', 'also', 'been', 'an', 'increase',
'in', 'trees', 'and', 'plants', '.', 'The', 'tree', 'cover', 'has', 'increased', 't
o', '3.05', '%', ',', 'as', 'compared', 'to', 'the', 'national', 'average', 'of',
'2.89', '%', ',', '"', 'the', 'official', 'said', ',', 'citing', 'the', '2019', 'For
est', 'Survey', 'of', 'India', 'report', '.', '\n\n', 'How', 'many', 'saplings', 'su
rvive', '?', '\n\n', 'Uttar', 'Pradesh', 'State', 'Forest', 'Minister', 'Dara', 'Sin
gh', 'said', 'the', 'long', '-', 'term', 'survival', 'of', 'the', 'saplings', 'remai
ns', 'a', 'concern', ',', 'adding', 'that', 'usually', 'only', '60', '%', 'of', 'th
e', 'saplings', 'survive', '.', 'The', 'rest', 'succumb', 'to', 'disease', 'or', 'la
ck', 'of', 'water', '.', '\n\n', 'However', ',', 'he', 'said', 'that', '\xa0', 'abou
t', '80', '%', 'of', 'the', 'saplings', 'planted', 'in', 'the', 'last', 'four', 'ann
ual', 'drives', 'have', 'survived', '.', '\n\n', '"', 'All', 'the', 'regions', 'wher
e', 'plantation', 'is', 'being', 'carried', 'out', 'have', 'been', 'geo', '-', 'tagg
ed', 'so', 'that', 'we', 'can', 'ascertain', 'what', 'exactly', 'happened', ',',
'"', 'Chauhan', 'told', 'The', 'Pioneer', 'newspaper', '.', '\n\n', '"', 'These', 's
aplings', 'carry', 'QR', 'codes', 'so', 'that', 'officials', 'can', 'maintain', 'a',
'record', 'and', 'verify', 'whether', 'the', 'saplings', 'survived', 'or', 'not',
'.', 'Besides', ',', 'teams', 'have', 'been', 'formed', 'to', 'monitor', 'progress',
'of', 'the', 'plantation', 'drive', ',', '"', 'he', 'said', '.', '\n\n', 'What', 'i
s', 'the', 'extent', 'of', 'India', "'s", 'tree', 'planting', 'project', '?', '\n
\n', 'India', 'has', 'vowed', 'to', 'have', 'a', 'third', 'of', 'its', 'total', 'lan
d', 'area', ',', 'or', '95', 'million', 'hectares', ',', 'under', 'forest', 'and',
'tree', 'cover', 'by', '2030', '.', '\n\n', 'The', 'government', 'has', 'allocated',
'$', '6.2', 'billion', '(', '€', '5.2', 'billion', ')', 'for', 'the', 'tree', '-',
'planting', 'across', 'the', 'country', '.', '\n\n', 'However', ',', 'industrial',
'development', 'and', 'a', 'rapidly', 'growing', 'population', 'has', 'put', 'furthe
r', 'stress', 'on', 'the', 'land', '.']
462
3790
```

In [292…

```
sentences = list(doc.sents)
print(len(sentences))
```

```
21
```

In [293…

```
for token in doc:
     print ("\n",token, token.idx, token.text_with_ws,
            token.is_alpha, token.is_punct, token.is_space,
            token.shape_, token.is_stop)
```

```
Millions 0 Millions  True False False Xxxxx False

of 9 of  True False False xx True

people 12 people  True False False xxxx False

in 19 in  True False False xx True

India 22 India  True False False Xxxxx False

took 28 took  True False False xxxx False

part 33 part  True False False xxxx True

in 38 in  True False False xx True

an 41 an  True False False xx True

annual 44 annual  True False False xxxx False

tree 51 tree  True False False xxxx False

planting 56 planting  True False False xxxx False

drive 65 drive  True False False xxxx False

Sunday 71 Sunday True False False Xxxxx False

. 77 .  False True False . False

More 79 More  True False False Xxxx True

than 84 than  True False False xxxx True

250 89 250  False False False ddd False

million 93 million  True False False xxxx False

saplings 101 saplings  True False False xxxx False

were 110 were  True False False xxxx True

planted 115 planted  True False False xxxx False

in 123 in  True False False xx True

a 126 a  True False False x True

single 128 single  True False False xxxx False

day 135 day  True False False xxx False

across 139 across  True False False xxxx True

the 146 the  True False False xxx True

country 150 country True False False xxxx False
```

```
's 157 's  False False False 'x True

most 160 most True False False xxxx True

- 164 - False True False - False

populous 165 populous  True False False xxxx False

state 174 state True False False xxxx False

. 179 . False True False . False


180

False False True

False

The 182 The  True False False Xxx True

campaign 186 campaign  True False False xxxx False

was 195 was  True False False xxx True

led 199 led  True False False xxx False

by 203 by  True False False xx True

Uttar 206 Uttar  True False False Xxxxx False

Pradesh 212 Pradesh  True False False Xxxxx False

state 220 state  True False False xxxx False

government 226 government  True False False xxxx False

officials 237 officials True False False xxxx False

, 246 ,  False True False , False

lawmakers 248 lawmakers True False False xxxx False

, 257 ,  False True False , False

and 259 and  True False False xxx True

activists 263 activists True False False xxxx False

, 272 ,  False True False , False

in 274 in  True False False xx True

a 277 a  True False False x True

bid 279 bid  True False False xxx False

to 283 to  True False False xx True

reduce 286 reduce  True False False xxxx False

carbon 293 carbon  True False False xxxx False
```

```
emissions 300 emissions  True False False xxxx False
and 310 and  True False False xxx True
combat 314 combat  True False False xxxx False
climate 321 climate  True False False xxxx False
change 329 change True False False xxxx False
. 335 . False True False . False

336
False False True
False
Where 338 Where  True False False Xxxxx True
were 344 were  True False False xxxx True
the 349 the  True False False xxx True
trees 353 trees  True False False xxxx False
planted 359 planted True False False xxxx False
? 366 ? False True False ? False

367
False False True
False
The 369 The  True False False Xxx True
saplings 373 saplings  True False False xxxx False
were 382 were  True False False xxxx True
planted 387 planted  True False False xxxx False
by 395 by  True False False xx True
volunteers 398 volunteers  True False False xxxx False
in 409 in  True False False xx True
forests 412 forests True False False xxxx False
, 419 ,  False True False , False
farms 421 farms True False False xxxx False
, 426 ,  False True False , False
schools 428 schools True False False xxxx False
```

```
, 435 ,  False True False , False

and 437 and  True False False xxx True

along 441 along  True False False xxxx True

riverbanks 447 riverbanks  True False False xxxx False

and 458 and  True False False xxx True

highways 462 highways True False False xxxx False

. 470 . False True False . False


471

False False True

False

" 473 " False True False " False

We 474 We  True False False Xx True

are 477 are  True False False xxx True

committed 481 committed  True False False xxxx False

to 491 to  True False False xx True

increasing 494 increasing  True False False xxxx False

the 505 the  True False False xxx True

forest 509 forest  True False False xxxx False

cover 516 cover  True False False xxxx False

of 522 of  True False False xx True

Uttar 525 Uttar  True False False Xxxxx False

Pradesh 531 Pradesh  True False False Xxxxx False

to 539 to  True False False xx True

over 542 over  True False False xxxx True

15 547 15 False False False dd False

% 549 %  False True False % False

of 551 of  True False False xx True

the 554 the  True False False xxx True

total 558 total  True False False xxxx False

land 564 land  True False False xxxx False

area 569 area  True False False xxxx False
```

```
in 574 in  True False False xx True

the 577 the  True False False xxx True

next 581 next  True False False xxxx True

five 586 five  True False False xxxx True

years 591 years True False False xxxx False

, 596 , False True False , False

'' 597 ''  False True False '' False

said 600 said  True False False xxxx False

state 605 state  True False False xxxx False

forest 611 forest  True False False xxxx False

official 618 official  True False False xxxx False

Manoj 627 Manoj  True False False Xxxxx False

Singh 633 Singh True False False Xxxxx False

. 638 . False True False . False


639

False False True

False

According 641 According  True False False Xxxxx False

to 651 to  True False False xx True

another 654 another  True False False xxxx True

government 662 government  True False False xxxx False

official 673 official True False False xxxx False

, 681 ,  False True False , False

the 683 the  True False False xxx True

forest 687 forest  True False False xxxx False

cover 694 cover  True False False xxxx False

of 700 of  True False False xx True

the 703 the  True False False xxx True

state 707 state  True False False xxxx False

has 713 has  True False False xxx True

increased 717 increased  True False False xxxx False
```

```
over 727 over  True False False xxxx True

the 732 the  True False False xxx True

last 736 last  True False False xxxx True

few 741 few  True False False xxx True

years 745 years True False False xxxx False

. 750 . False True False . False


751

False False True

False

" 753 " False True False " False

There 754 There  True False False Xxxxx True

has 760 has  True False False xxx True

been 764 been  True False False xxxx True

an 769 an  True False False xx True

increase 772 increase  True False False xxxx False

of 781 of  True False False xx True

127 784 127  False False False ddd False

sqare 788 sqare  True False False xxxx False

kilometers 794 kilometers  True False False xxxx False

[ 805 [ False True False [ False

79 806 79  False False False dd False

sqare 809 sqare  True False False xxxx False

miles 815 miles True False False xxxx False

] 820 ] False True False ] False

  821   False False True   False

in 822 in  True False False xx True

the 825 the  True False False xxx True

forest 829 forest  True False False xxxx False

cover 836 cover  True False False xxxx False

in 842 in  True False False xx True

Uttar 845 Uttar  True False False Xxxxx False
```

```
Pradesh 851 Pradesh  True False False Xxxxx False

as 859 as  True False False xx True

compared 862 compared  True False False xxxx False

to 871 to  True False False xx True

2017 874 2017 False False False dddd False

, 878 , False True False , False

" 879 "  False True False " False

a 881 a  True False False x True

state 883 state  True False False xxxx False

government 889 government  True False False xxxx False

spokesperson 900 spokesperson  True False False xxxx False

was 913 was  True False False xxx True

quoted 917 quoted  True False False xxxx False

as 924 as  True False False xx True

saying 927 saying  True False False xxxx False

in 934 in True False False xx True

  936   False False True   False

The 937 The  True False False Xxx True

Indian 941 Indian  True False False Xxxxx False

Express 948 Express  True False False Xxxxx False

newspaper 956 newspaper True False False xxxx False

. 965 . False True False . False


966

False False True

False

" 968 " False True False " False

There 969 There  True False False Xxxxx True

has 975 has  True False False xxx True

also 979 also  True False False xxxx True

been 984 been  True False False xxxx True

an 989 an  True False False xx True
```

```
increase 992 increase  True False False xxxx False
in 1001 in  True False False xx True
trees 1004 trees  True False False xxxx False
and 1010 and  True False False xxx True
plants 1014 plants True False False xxxx False
. 1020 .  False True False . False
The 1022 The  True False False Xxx True
tree 1026 tree  True False False xxxx False
cover 1031 cover  True False False xxxx False
has 1037 has  True False False xxx True
increased 1041 increased  True False False xxxx False
to 1051 to  True False False xx True
3.05 1054 3.05 False False False d.dd False
% 1058 % False True False % False
, 1059 ,  False True False , False
as 1061 as  True False False xx True
compared 1064 compared  True False False xxxx False
to 1073 to  True False False xx True
the 1076 the  True False False xxx True
national 1080 national  True False False xxxx False
average 1089 average  True False False xxxx False
of 1097 of  True False False xx True
2.89 1100 2.89 False False False d.dd False
% 1104 % False True False % False
, 1105 , False True False , False
" 1106 "  False True False " False
the 1108 the  True False False xxx True
official 1112 official  True False False xxxx False
said 1121 said True False False xxxx False
, 1125 ,  False True False , False
citing 1127 citing  True False False xxxx False
the 1134 the  True False False xxx True
```

```
2019 1138 2019  False False False dddd False

Forest 1143 Forest  True False False Xxxxx False

Survey 1150 Survey  True False False Xxxxx False

of 1157 of  True False False xx True

India 1160 India  True False False Xxxxx False

report 1166 report True False False xxxx False

. 1172 . False True False . False


1173

False False True

False

How 1175 How  True False False Xxx True

many 1179 many  True False False xxxx True

saplings 1184 saplings  True False False xxxx False

survive 1193 survive True False False xxxx False

? 1200 ? False True False ? False


1201

False False True

False

Uttar 1203 Uttar  True False False Xxxxx False

Pradesh 1209 Pradesh  True False False Xxxxx False

State 1217 State  True False False Xxxxx False

Forest 1223 Forest  True False False Xxxxx False

Minister 1230 Minister  True False False Xxxxx False

Dara 1239 Dara  True False False Xxxx False

Singh 1244 Singh  True False False Xxxxx False

said 1250 said  True False False xxxx False

the 1255 the  True False False xxx True

long 1259 long True False False xxxx False

- 1263 - False True False - False

term 1264 term  True False False xxxx False
```

```
survival 1269 survival  True False False xxxx False

of 1278 of  True False False xx True

the 1281 the  True False False xxx True

saplings 1285 saplings  True False False xxxx False

remains 1294 remains  True False False xxxx False

a 1302 a  True False False x True

concern 1304 concern True False False xxxx False

, 1311 ,  False True False , False

adding 1313 adding  True False False xxxx False

that 1320 that  True False False xxxx True

usually 1325 usually  True False False xxxx False

only 1333 only  True False False xxxx True

60 1338 60 False False False dd False

% 1340 %  False True False % False

of 1342 of  True False False xx True

the 1345 the  True False False xxx True

saplings 1349 saplings  True False False xxxx False

survive 1358 survive True False False xxxx False

. 1365 .  False True False . False

The 1367 The  True False False Xxx True

rest 1371 rest  True False False xxxx False

succumb 1376 succumb  True False False xxxx False

to 1384 to  True False False xx True

disease 1387 disease  True False False xxxx False

or 1395 or  True False False xx True

lack 1398 lack  True False False xxxx False

of 1403 of  True False False xx True

water 1406 water True False False xxxx False

. 1411 . False True False . False


1412

False False True
```

```
False

However 1414 However True False False Xxxxx True

, 1421 ,  False True False , False

he 1423 he  True False False xx True

said 1426 said  True False False xxxx False

that 1431 that True False False xxxx True

  1435   False False True   False

about 1436 about  True False False xxxx True

80 1442 80 False False False dd False

% 1444 %  False True False % False

of 1446 of  True False False xx True

the 1449 the  True False False xxx True

saplings 1453 saplings  True False False xxxx False

planted 1462 planted  True False False xxxx False

in 1470 in  True False False xx True

the 1473 the  True False False xxx True

last 1477 last  True False False xxxx True

four 1482 four  True False False xxxx True

annual 1487 annual  True False False xxxx False

drives 1494 drives  True False False xxxx False

have 1501 have  True False False xxxx True

survived 1506 survived True False False xxxx False

. 1514 . False True False . False


1515

False False True

False

" 1517 " False True False " False

All 1518 All  True False False Xxx True

the 1522 the  True False False xxx True

regions 1526 regions  True False False xxxx False

where 1534 where  True False False xxxx True
```

```
plantation 1540 plantation  True False False xxxx False

is 1551 is  True False False xx True

being 1554 being  True False False xxxx True

carried 1560 carried  True False False xxxx False

out 1568 out  True False False xxx True

have 1572 have  True False False xxxx True

been 1577 been  True False False xxxx True

geo 1582 geo True False False xxx False

- 1585 - False True False - False

tagged 1586 tagged  True False False xxxx False

so 1593 so  True False False xx True

that 1596 that  True False False xxxx True

we 1601 we  True False False xx True

can 1604 can  True False False xxx True

ascertain 1608 ascertain  True False False xxxx False

what 1618 what  True False False xxxx True

exactly 1623 exactly  True False False xxxx False

happened 1631 happened True False False xxxx False

, 1639 , False True False , False

" 1640 "  False True False " False

Chauhan 1642 Chauhan  True False False Xxxxx False

told 1650 told  True False False xxxx False

The 1655 The  True False False Xxx True

Pioneer 1659 Pioneer  True False False Xxxxx False

newspaper 1667 newspaper True False False xxxx False

. 1676 . False True False . False


1677

False False True

False

" 1679 " False True False " False

These 1680 These  True False False Xxxxx True
```

```
saplings 1686 saplings  True False False xxxx False

carry 1695 carry  True False False xxxx False

QR 1701 QR  True False False XX False

codes 1704 codes  True False False xxxx False

so 1710 so  True False False xx True

that 1713 that  True False False xxxx True

officials 1718 officials  True False False xxxx False

can 1728 can  True False False xxx True

maintain 1732 maintain  True False False xxxx False

a 1741 a  True False False x True

record 1743 record  True False False xxxx False

and 1750 and  True False False xxx True

verify 1754 verify  True False False xxxx False

whether 1761 whether  True False False xxxx True

the 1769 the  True False False xxx True

saplings 1773 saplings  True False False xxxx False

survived 1782 survived  True False False xxxx False

or 1791 or  True False False xx True

not 1794 not True False False xxx True

. 1797 .  False True False . False

Besides 1799 Besides True False False Xxxxx True

, 1806 ,  False True False , False

teams 1808 teams  True False False xxxx False

have 1814 have  True False False xxxx True

been 1819 been  True False False xxxx True

formed 1824 formed  True False False xxxx False

to 1831 to  True False False xx True

monitor 1834 monitor  True False False xxxx False

progress 1842 progress  True False False xxxx False

of 1851 of  True False False xx True

the 1854 the  True False False xxx True

plantation 1858 plantation  True False False xxxx False
```

```
drive 1869 drive True False False xxxx False

, 1874 , False True False , False

" 1875 "  False True False " False

he 1877 he  True False False xx True

said 1880 said True False False xxxx False

. 1884 . False True False . False


1885

False False True

False

What 1887 What  True False False Xxxx True

is 1892 is  True False False xx True

the 1895 the  True False False xxx True

extent 1899 extent  True False False xxxx False

of 1906 of  True False False xx True

India 1909 India True False False Xxxxx False

's 1914 's  False False False 'x True

tree 1917 tree  True False False xxxx False

planting 1922 planting  True False False xxxx False

project 1931 project True False False xxxx False

? 1938 ? False True False ? False


1939

False False True

False

India 1941 India  True False False Xxxxx False

has 1947 has  True False False xxx True

vowed 1951 vowed  True False False xxxx False

to 1957 to  True False False xx True

have 1960 have  True False False xxxx True

a 1965 a  True False False x True

third 1967 third  True False False xxxx True
```

```
of 1973 of  True False False xx True

its 1976 its  True False False xxx True

total 1980 total  True False False xxxx False

land 1986 land  True False False xxxx False

area 1991 area True False False xxxx False

, 1995 ,  False True False , False

or 1997 or  True False False xx True

95 2000 95  False False False dd False

million 2003 million  True False False xxxx False

hectares 2011 hectares True False False xxxx False

, 2019 ,  False True False , False

under 2021 under  True False False xxxx True

forest 2027 forest  True False False xxxx False

and 2034 and  True False False xxx True

tree 2038 tree  True False False xxxx False

cover 2043 cover  True False False xxxx False

by 2049 by  True False False xx True

2030 2052 2030 False False False dddd False

. 2056 . False True False . False


2057

False False True

False

The 2059 The  True False False Xxx True

government 2063 government  True False False xxxx False

has 2074 has  True False False xxx True

allocated 2078 allocated  True False False xxxx False

$ 2088 $ False False False $ False

6.2 2089 6.2  False False False d.d False

billion 2093 billion  True False False xxxx False

( 2101 ( False True False ( False

€ 2102 € False False False € False
```

```
5.2 2103 5.2  False False False d.d False

billion 2107 billion True False False xxxx False

) 2114 )  False True False ) False

for 2116 for  True False False xxx True

the 2120 the  True False False xxx True

tree 2124 tree True False False xxxx False

- 2128 - False True False - False

planting 2129 planting  True False False xxxx False

across 2138 across  True False False xxxx True

the 2145 the  True False False xxx True

country 2149 country True False False xxxx False

. 2156 . False True False . False


2157

False False True

False

However 2159 However True False False Xxxxx True

, 2166 ,  False True False , False

industrial 2168 industrial  True False False xxxx False

development 2179 development  True False False xxxx False

and 2191 and  True False False xxx True

a 2195 a  True False False x True

rapidly 2197 rapidly  True False False xxxx False

growing 2205 growing  True False False xxxx False

population 2213 population  True False False xxxx False

has 2224 has  True False False xxx True

put 2228 put  True False False xxx True

further 2232 further  True False False xxxx True

stress 2240 stress  True False False xxxx False

on 2247 on  True False False xx True

the 2250 the  True False False xxx True

land 2254 land True False False xxxx False
```

```
. 2258 . False True False . False
```

In [294…

```
spacy_stopwords = spacy.lang.en.stop_words.STOP_WORDS

len(spacy_stopwords)
```

Out[294…

```
326
```

In [295…

```
words = [word.lemma_ for word in doc]
print(words)
```

```
['million', 'of', 'people', 'in', 'India', 'take', 'part', 'in', 'an', 'annual', 'tr
ee', 'planting', 'drive', 'Sunday', '.', 'More', 'than', '250', 'million', 'saplin
g', 'be', 'plant', 'in', 'a', 'single', 'day', 'across', 'the', 'country', "'s", 'mo
st', '-', 'populous', 'state', '.', '\n\n', 'the', 'campaign', 'be', 'lead', 'by',
'Uttar', 'Pradesh', 'state', 'government', 'official', ',', 'lawmaker', ',', 'and',
'activist', ',', 'in', 'a', 'bid', 'to', 'reduce', 'carbon', 'emission', 'and', 'com
bat', 'climate', 'change', '.', '\n\n', 'where', 'be', 'the', 'tree', 'plant', '?',
'\n\n', 'the', 'sapling', 'be', 'plant', 'by', 'volunteer', 'in', 'forest', ',', 'fa
rm', ',', 'school', ',', 'and', 'along', 'riverbank', 'and', 'highway', '.', '\n\n',
'"', 'we', 'be', 'committed', 'to', 'increase', 'the', 'forest', 'cover', 'of', 'Utt
ar', 'Pradesh', 'to', 'over', '15', '%', 'of', 'the', 'total', 'land', 'area', 'in',
'the', 'next', 'five', 'year', ',', "'", 'say', 'state', 'forest', 'official', 'Man
oj', 'Singh', '.', '\n\n', 'accord', 'to', 'another', 'government', 'official', ',',
'the', 'forest', 'cover', 'of', 'the', 'state', 'have', 'increase', 'over', 'the',
'last', 'few', 'year', '.', '\n\n', '"', 'there', 'have', 'be', 'an', 'increase', 'o
f', '127', 'sqare', 'kilometer', '[', '79', 'sqare', 'mile', ']', '\xa0', 'in', 'th
e', 'forest', 'cover', 'in', 'Uttar', 'Pradesh', 'as', 'compare', 'to', '2017', ',',
'"', 'a', 'state', 'government', 'spokesperson', 'be', 'quote', 'as', 'say', 'in',
'\xa0', 'the', 'Indian', 'Express', 'newspaper', '.', '\n\n', '"', 'there', 'have',
'also', 'be', 'an', 'increase', 'in', 'tree', 'and', 'plant', '.', 'the', 'tree', 'c
over', 'have', 'increase', 'to', '3.05', '%', ',', 'as', 'compare', 'to', 'the', 'na
tional', 'average', 'of', '2.89', '%', ',', '"', 'the', 'official', 'say', ',', 'cit
e', 'the', '2019', 'Forest', 'Survey', 'of', 'India', 'report', '.', '\n\n', 'how',
'many', 'sapling', 'survive', '?', '\n\n', 'Uttar', 'Pradesh', 'State', 'Forest', 'M
inister', 'Dara', 'Singh', 'say', 'the', 'long', '-', 'term', 'survival', 'of', 'th
e', 'sapling', 'remain', 'a', 'concern', ',', 'add', 'that', 'usually', 'only', '6
0', '%', 'of', 'the', 'sapling', 'survive', '.', 'the', 'rest', 'succumb', 'to', 'di
sease', 'or', 'lack', 'of', 'water', '.', '\n\n', 'however', ',', 'he', 'say', 'tha
t', '\xa0', 'about', '80', '%', 'of', 'the', 'sapling', 'plant', 'in', 'the', 'las
t', 'four', 'annual', 'drive', 'have', 'survive', '.', '\n\n', '"', 'all', 'the', 'r
egion', 'where', 'plantation', 'be', 'be', 'carry', 'out', 'have', 'be', 'geo', '-',
'tag', 'so', 'that', 'we', 'can', 'ascertain', 'what', 'exactly', 'happen', ',',
'"', 'Chauhan', 'tell', 'the', 'Pioneer', 'newspaper', '.', '\n\n', '"', 'these', 's
apling', 'carry', 'qr', 'code', 'so', 'that', 'official', 'can', 'maintain', 'a', 'r
ecord', 'and', 'verify', 'whether', 'the', 'sapling', 'survive', 'or', 'not', '.',
'besides', ',', 'team', 'have', 'be', 'form', 'to', 'monitor', 'progress', 'of', 'th
e', 'plantation', 'drive', ',', '"', 'he', 'say', '.', '\n\n', 'what', 'be', 'the',
'extent', 'of', 'India', "'s", 'tree', 'planting', 'project', '?', '\n\n', 'India',
'have', 'vow', 'to', 'have', 'a', 'third', 'of', 'its', 'total', 'land', 'area',
',', 'or', '95', 'million', 'hectare', ',', 'under', 'forest', 'and', 'tree', 'cove
r', 'by', '2030', '.', '\n\n', 'the', 'government', 'have', 'allocate', '$', '6.2',
'billion', '(', '€', '5.2', 'billion', ')', 'for', 'the', 'tree', '-', 'planting',
'across', 'the', 'country', '.', '\n\n', 'however', ',', 'industrial', 'developmen
t', 'and', 'a', 'rapidly', 'grow', 'population', 'have', 'put', 'further', 'stress',
'on', 'the', 'land', '.']
```

In [ ]:

In [296…

```
vocabulary = []
vocabulary =  " ".join([word.lemma_ for word in doc if word not in spacy_stopwords])
print(vocabulary)
```

```
million of people in India take part in an annual tree planting drive Sunday . More
than 250 million sapling be plant in a single day across the country 's most - popul
ous state .

 the campaign be lead by Uttar Pradesh state government official , lawmaker , and ac
tivist , in a bid to reduce carbon emission and combat climate change .

 where be the tree plant ?

 the sapling be plant by volunteer in forest , farm , school , and along riverbank a
nd highway .

 " we be committed to increase the forest cover of Uttar Pradesh to over 15 % of the
total land area in the next five year , '' say state forest official Manoj Singh .

 accord to another government official , the forest cover of the state have increase
over the last few year .

 " there have be an increase of 127 sqare kilometer [ 79 sqare mile ]   in the fores
t cover in Uttar Pradesh as compare to 2017 , " a state government spokesperson be q
uote as say in   the Indian Express newspaper .

 " there have also be an increase in tree and plant . the tree cover have increase t
o 3.05 % , as compare to the national average of 2.89 % , " the official say , cite
the 2019 Forest Survey of India report .

 how many sapling survive ?

 Uttar Pradesh State Forest Minister Dara Singh say the long - term survival of the
sapling remain a concern , add that usually only 60 % of the sapling survive . the r
est succumb to disease or lack of water .

 however , he say that   about 80 % of the sapling plant in the last four annual dri
ve have survive .

 " all the region where plantation be be carry out have be geo - tag so that we can
ascertain what exactly happen , " Chauhan tell the Pioneer newspaper .

 " these sapling carry qr code so that official can maintain a record and verify whe
ther the sapling survive or not . besides , team have be form to monitor progress of
the plantation drive , " he say .

 what be the extent of India 's tree planting project ?

 India have vow to have a third of its total land area , or 95 million hectare , und
er forest and tree cover by 2030 .

 the government have allocate $ 6.2 billion ( € 5.2 billion ) for the tree - plantin
g across the country .

 however , industrial development and a rapidly grow population have put further str
ess on the land .
```

In [297…

```
for token in doc:
    print(token, token.pos_)
```

```
Millions NOUN
of ADP
```

```
people NOUN
in ADP
India PROPN
took VERB
part NOUN
in ADP
an DET
annual ADJ
tree NOUN
planting NOUN
drive NOUN
Sunday PROPN
. PUNCT
More ADJ
than ADP
250 NUM
million NUM
saplings NOUN
were AUX
planted VERB
in ADP
a DET
single ADJ
day NOUN
across ADP
the DET
country NOUN
's PART
most ADV
- PUNCT
populous ADJ
state NOUN
. PUNCT


 SPACE
The DET
campaign NOUN
was AUX
led VERB
by ADP
Uttar PROPN
Pradesh PROPN
state NOUN
government NOUN
officials NOUN
, PUNCT
lawmakers NOUN
, PUNCT
and CCONJ
activists NOUN
, PUNCT
in ADP
a DET
bid NOUN
to PART
reduce VERB
carbon NOUN
emissions NOUN
and CCONJ
combat NOUN
climate NOUN
change NOUN
. PUNCT
```

```
 SPACE
Where SCONJ
were AUX
the DET
trees NOUN
planted VERB
? PUNCT


 SPACE
The DET
saplings NOUN
were AUX
planted VERB
by ADP
volunteers NOUN
in ADP
forests NOUN
, PUNCT
farms NOUN
, PUNCT
schools NOUN
, PUNCT
and CCONJ
along ADP
riverbanks NOUN
and CCONJ
highways NOUN
. PUNCT


 SPACE
" PUNCT
We PRON
are AUX
committed ADJ
to ADP
increasing VERB
the DET
forest NOUN
cover NOUN
of ADP
Uttar PROPN
Pradesh PROPN
to ADP
over ADP
15 NUM
% NOUN
of ADP
the DET
total ADJ
land NOUN
area NOUN
in ADP
the DET
next ADJ
five NUM
years NOUN
, PUNCT
'' PUNCT
said VERB
state NOUN
```

```
forest NOUN
official NOUN
Manoj PROPN
Singh PROPN
. PUNCT


 SPACE
According VERB
to ADP
another DET
government NOUN
official NOUN
, PUNCT
the DET
forest NOUN
cover NOUN
of ADP
the DET
state NOUN
has AUX
increased VERB
over ADP
the DET
last ADJ
few ADJ
years NOUN
. PUNCT


 SPACE
" PUNCT
There PRON
has AUX
been AUX
an DET
increase NOUN
of ADP
127 NUM
sqare NOUN
kilometers NOUN
[ PUNCT
79 NUM
sqare NOUN
miles NOUN
] PUNCT
  SPACE
in ADP
the DET
forest NOUN
cover NOUN
in ADP
Uttar PROPN
Pradesh PROPN
as SCONJ
compared VERB
to ADP
2017 NUM
, PUNCT
" PUNCT
a DET
state NOUN
government NOUN
spokesperson NOUN
```

```
was AUX
quoted VERB
as ADP
saying VERB
in ADP
  SPACE
The DET
Indian PROPN
Express PROPN
newspaper NOUN
. PUNCT


 SPACE
" PUNCT
There PRON
has AUX
also ADV
been AUX
an DET
increase NOUN
in ADP
trees NOUN
and CCONJ
plants NOUN
. PUNCT
The DET
tree NOUN
cover NOUN
has AUX
increased VERB
to ADP
3.05 NUM
% NOUN
, PUNCT
as SCONJ
compared VERB
to ADP
the DET
national ADJ
average NOUN
of ADP
2.89 NUM
% NOUN
, PUNCT
" PUNCT
the DET
official NOUN
said VERB
, PUNCT
citing VERB
the DET
2019 NUM
Forest PROPN
Survey PROPN
of ADP
India PROPN
report NOUN
. PUNCT


 SPACE
How SCONJ
many ADJ
```

```
saplings NOUN
survive VERB
? PUNCT


 SPACE
Uttar PROPN
Pradesh PROPN
State PROPN
Forest PROPN
Minister PROPN
Dara PROPN
Singh PROPN
said VERB
the DET
long ADJ
- PUNCT
term NOUN
survival NOUN
of ADP
the DET
saplings NOUN
remains VERB
a DET
concern NOUN
, PUNCT
adding VERB
that SCONJ
usually ADV
only ADV
60 NUM
% NOUN
of ADP
the DET
saplings NOUN
survive VERB
. PUNCT
The DET
rest NOUN
succumb NOUN
to PART
disease VERB
or CCONJ
lack NOUN
of ADP
water NOUN
. PUNCT


 SPACE
However ADV
, PUNCT
he PRON
said VERB
that SCONJ
  SPACE
about ADV
80 NUM
% NOUN
of ADP
the DET
saplings NOUN
planted VERB
in ADP
```

```
the DET
last ADJ
four NUM
annual ADJ
drives NOUN
have AUX
survived VERB
. PUNCT


 SPACE
" PUNCT
All DET
the DET
regions NOUN
where SCONJ
plantation NOUN
is AUX
being AUX
carried VERB
out ADP
have AUX
been AUX
geo NOUN
- PUNCT
tagged VERB
so SCONJ
that SCONJ
we PRON
can AUX
ascertain VERB
what PRON
exactly ADV
happened VERB
, PUNCT
" PUNCT
Chauhan PROPN
told VERB
The DET
Pioneer PROPN
newspaper NOUN
. PUNCT


 SPACE
" PUNCT
These DET
saplings NOUN
carry VERB
QR NOUN
codes NOUN
so SCONJ
that SCONJ
officials NOUN
can AUX
maintain VERB
a DET
record NOUN
and CCONJ
verify VERB
whether SCONJ
the DET
saplings NOUN
survived VERB
```

```
or CCONJ
not PART
. PUNCT
Besides ADV
, PUNCT
teams NOUN
have AUX
been AUX
formed VERB
to PART
monitor VERB
progress NOUN
of ADP
the DET
plantation NOUN
drive NOUN
, PUNCT
" PUNCT
he PRON
said VERB
. PUNCT


 SPACE
What PRON
is AUX
the DET
extent NOUN
of ADP
India PROPN
's PART
tree NOUN
planting NOUN
project NOUN
? PUNCT


 SPACE
India PROPN
has AUX
vowed VERB
to PART
have VERB
a DET
third NOUN
of ADP
its PRON
total ADJ
land NOUN
area NOUN
, PUNCT
or CCONJ
95 NUM
million NUM
hectares NOUN
, PUNCT
under ADP
forest NOUN
and CCONJ
tree NOUN
cover NOUN
by ADP
2030 NUM
. PUNCT
```

```
 SPACE
The DET
government NOUN
has AUX
allocated VERB
$ SYM
6.2 NUM
billion NUM
( PUNCT
€ SYM
5.2 NUM
billion NUM
) PUNCT
for ADP
the DET
tree NOUN
- PUNCT
planting NOUN
across ADP
the DET
country NOUN
. PUNCT


 SPACE
However ADV
, PUNCT
industrial ADJ
development NOUN
and CCONJ
a DET
rapidly ADV
growing VERB
population NOUN
has AUX
put VERB
further ADJ
stress NOUN
on ADP
the DET
land NOUN
. PUNCT
```

In [298…

```
verbs = [token.text for token in doc if token.pos_ == "VERB"]
nouns = [token.text for token in doc if token.pos_ == "NOUN"]
print('Verbs ',len(verbs),'Nouns ',len(nouns))
print('Verbs ',verbs)
```

```
Verbs  43 Nouns  113
Verbs  ['took', 'planted', 'led', 'reduce', 'planted', 'planted', 'increasing', 'sai
d', 'According', 'increased', 'compared', 'quoted', 'saying', 'increased', 'compare
d', 'said', 'citing', 'survive', 'said', 'remains', 'adding', 'survive', 'disease',
'said', 'planted', 'survived', 'carried', 'tagged', 'ascertain', 'happened', 'told',
'carry', 'maintain', 'verify', 'survived', 'formed', 'monitor', 'said', 'vowed', 'ha
ve', 'allocated', 'growing', 'put']
```

In [299…

```
for token in doc:
    print(token, token.lemma_)
```

```
Millions million
of of
```

```
people people
in in
India India
took take
part part
in in
an an
annual annual
tree tree
planting planting
drive drive
Sunday Sunday
. .
More More
than than
250 250
million million
saplings sapling
were be
planted plant
in in
a a
single single
day day
across across
the the
country country
's 's
most most
- -
populous populous
state state
. .




The the
campaign campaign
was be
led lead
by by
Uttar Uttar
Pradesh Pradesh
state state
government government
officials official
, ,
lawmakers lawmaker
, ,
and and
activists activist
, ,
in in
a a
bid bid
to to
reduce reduce
carbon carbon
emissions emission
and and
combat combat
climate climate
```

```
change change
. .



Where where
were be
the the
trees tree
planted plant
? ?



The the
saplings sapling
were be
planted plant
by by
volunteers volunteer
in in
forests forest
, ,
farms farm
, ,
schools school
, ,
and and
along along
riverbanks riverbank
and and
highways highway
. .



" "
We we
are be
committed committed
to to
increasing increase
the the
forest forest
cover cover
of of
Uttar Uttar
Pradesh Pradesh
to to
over over
15 15
% %
of of
the the
total total
land land
area area
in in
```

```
the the
next next
five five
years year
, ,
'' ''
said say
state state
forest forest
official official
Manoj Manoj
Singh Singh
. .



According accord
to to
another another
government government
official official
, ,
the the
forest forest
cover cover
of of
the the
state state
has have
increased increase
over over
the the
last last
few few
years year
. .



" "
There there
has have
been be
an an
increase increase
of of
127 127
sqare sqare
kilometers kilometer
[ [
79 79
sqare sqare
miles mile
] ]

in in
the the
forest forest
cover cover
in in
```

```
Uttar Uttar
Pradesh Pradesh
as as
compared compare
to to
2017 2017
, ,
" "
a a
state state
government government
spokesperson spokesperson
was be
quoted quote
as as
saying say
in in

The the
Indian Indian
Express Express
newspaper newspaper
. .




" "
There there
has have
also also
been be
an an
increase increase
in in
trees tree
and and
plants plant
. .
The the
tree tree
cover cover
has have
increased increase
to to
3.05 3.05
% %
, ,
as as
compared compare
to to
the the
national national
average average
of of
2.89 2.89
% %
, ,
" "
the the
official official
said say
, ,
```

```
citing cite
the the
2019 2019
Forest Forest
Survey Survey
of of
India India
report report
. .




How how
many many
saplings sapling
survive survive
? ?




Uttar Uttar
Pradesh Pradesh
State State
Forest Forest
Minister Minister
Dara Dara
Singh Singh
said say
the the
long long
- -
term term
survival survival
of of
the the
saplings sapling
remains remain
a a
concern concern
, ,
adding add
that that
usually usually
only only
60 60
% %
of of
the the
saplings sapling
survive survive
. .
The the
rest rest
succumb succumb
to to
disease disease
or or
lack lack
of of
water water
```

```
. .



However however
, ,
he he
said say
that that

about about
80 80
% %
of of
the the
saplings sapling
planted plant
in in
the the
last last
four four
annual annual
drives drive
have have
survived survive
. .



" "
All all
the the
regions region
where where
plantation plantation
is be
being be
carried carry
out out
have have
been be
geo geo
- -
tagged tag
so so
that that
we we
can can
ascertain ascertain
what what
exactly exactly
happened happen
, ,
" "
Chauhan Chauhan
told tell
The the
Pioneer Pioneer
newspaper newspaper
. .
```

" "
These these
saplings sapling
carry carry
QR qr
codes code
so so
that that
officials official
can can
maintain maintain
a a
record record
and and
verify verify
whether whether
the the
saplings sapling
survived survive
or or
not not
. .
Besides besides
, ,
teams team
have have
been be
formed form
to to
monitor monitor
progress progress
of of
the the
plantation plantation
drive drive
, ,
" "
he he
said say
. .

What what
is be
the the
extent extent
of of
India India
's 's
tree tree
planting planting
project project
? ?

```
India India
has have
vowed vow
to to
have have
a a
third third
of of
its its
total total
land land
area area
, ,
or or
95 95
million million
hectares hectare
, ,
under under
forest forest
and and
tree tree
cover cover
by by
2030 2030
. .
```

```
The the
government government
has have
allocated allocate
$ $
6.2 6.2
billion billion
( (
€ €
5.2 5.2
billion billion
) )
for for
the the
tree tree
- -
planting planting
across across
the the
country country
. .
```

```
However however
, ,
industrial industrial
development development
and and
```

```
a a
rapidly rapidly
growing grow
population population
has have
put put
further further
stress stress
on on
the the
land land
. .
```

In [300…

```
import nltk
from nltk.stem import WordNetLemmatizer,PorterStemmer
```

In [301…

```
words = []

words = " ".join(token.text for token in doc)
words
```

Out[301…

```
'Millions of people in India took part in an annual tree planting drive Sunday . Mor
e than 250 million saplings were planted in a single day across the country \'s most
- populous state . \n\n The campaign was led by Uttar Pradesh state government offic
ials , lawmakers , and activists , in a bid to reduce carbon emissions and combat cl
imate change . \n\n Where were the trees planted ? \n\n The saplings were planted by
volunteers in forests , farms , schools , and along riverbanks and highways . \n\n "
We are committed to increasing the forest cover of Uttar Pradesh to over 15 % of the
total land area in the next five years , \'\' said state forest official Manoj Singh
. \n\n According to another government official , the forest cover of the state has
increased over the last few years . \n\n " There has been an increase of 127 sqare k
ilometers [ 79 sqare miles ] \xa0 in the forest cover in Uttar Pradesh as compared t
o 2017 , " a state government spokesperson was quoted as saying in \xa0 The Indian E
xpress newspaper . \n\n " There has also been an increase in trees and plants . The
tree cover has increased to 3.05 % , as compared to the national average of 2.89 % ,
" the official said , citing the 2019 Forest Survey of India report . \n\n How many
saplings survive ? \n\n Uttar Pradesh State Forest Minister Dara Singh said the long
- term survival of the saplings remains a concern , adding that usually only 60 % of
the saplings survive . The rest succumb to disease or lack of water . \n\n However ,
he said that \xa0 about 80 % of the saplings planted in the last four annual drives
have survived . \n\n " All the regions where plantation is being carried out have be
en geo - tagged so that we can ascertain what exactly happened , " Chauhan told The
Pioneer newspaper . \n\n " These saplings carry QR codes so that officials can maint
ain a record and verify whether the saplings survived or not . Besides , teams have
been formed to monitor progress of the plantation drive , " he said . \n\n What is t
he extent of India \'s tree planting project ? \n\n India has vowed to have a third
of its total land area , or 95 million hectares , under forest and tree cover by 203
0 . \n\n The government has allocated $ 6.2 billion ( € 5.2 billion ) for the tree -
planting across the country . \n\n However , industrial development and a rapidly gr
owing population has put further stress on the land .'
```

In [302…

```
ps = PorterStemmer()
stemmed_words = []
for token in doc:
    stemmed_words.append(ps.stem(token.text))
```

In [303…

```
stemmed_words
```

Out[303…

```
['million',
 'of',
 'peopl',
 'in',
 'india',
 'took',
 'part',
 'in',
 'an',
 'annual',
 'tree',
 'plant',
 'drive',
 'sunday',
 '.',
 'more',
 'than',
 '250',
 'million',
 'sapl',
 'were',
 'plant',
 'in',
 'a',
 'singl',
 'day',
 'across',
 'the',
 'countri',
 "'s",
 'most',
 '-',
 'popul',
 'state',
 '.',
 '\n\n',
 'the',
 'campaign',
 'wa',
 'led',
 'by',
 'uttar',
 'pradesh',
 'state',
 'govern',
 'offici',
 ',',
 'lawmak',
 ',',
 'and',
 'activist',
 ',',
 'in',
 'a',
 'bid',
 'to',
 'reduc',
 'carbon',
 'emiss',
 'and',
 'combat',
 'climat',
 'chang',
 '.',
```

```
 '\n\n',
 'where',
 'were',
 'the',
 'tree',
 'plant',
 '?',
 '\n\n',
 'the',
 'sapl',
 'were',
 'plant',
 'by',
 'volunt',
 'in',
 'forest',
 ',',
 'farm',
 ',',
 'school',
 ',',
 'and',
 'along',
 'riverbank',
 'and',
 'highway',
 '.',
 '\n\n',
 '"',
 'We',
 'are',
 'commit',
 'to',
 'increas',
 'the',
 'forest',
 'cover',
 'of',
 'uttar',
 'pradesh',
 'to',
 'over',
 '15',
 '%',
 'of',
 'the',
 'total',
 'land',
 'area',
 'in',
 'the',
 'next',
 'five',
 'year',
 ',',
 "''",
 'said',
 'state',
 'forest',
 'offici',
 'manoj',
 'singh',
 '.',
 '\n\n',
```

```
 'accord',
 'to',
 'anoth',
 'govern',
 'offici',
 ',',
 'the',
 'forest',
 'cover',
 'of',
 'the',
 'state',
 'ha',
 'increas',
 'over',
 'the',
 'last',
 'few',
 'year',
 '.',
 '\n\n',
 '"',
 'there',
 'ha',
 'been',
 'an',
 'increas',
 'of',
 '127',
 'sqare',
 'kilomet',
 '[',
 '79',
 'sqare',
 'mile',
 ']',
 '\xa0',
 'in',
 'the',
 'forest',
 'cover',
 'in',
 'uttar',
 'pradesh',
 'as',
 'compar',
 'to',
 '2017',
 ',',
 '"',
 'a',
 'state',
 'govern',
 'spokesperson',
 'wa',
 'quot',
 'as',
 'say',
 'in',
 '\xa0',
 'the',
 'indian',
 'express',
 'newspap',
```

```
'.',
'\n\n',
'"',
'there',
'ha',
'also',
'been',
'an',
'increas',
'in',
'tree',
'and',
'plant',
'.',
'the',
'tree',
'cover',
'ha',
'increas',
'to',
'3.05',
'%',
',',
'as',
'compar',
'to',
'the',
'nation',
'averag',
'of',
'2.89',
'%',
',',
'"',
'the',
'offici',
'said',
',',
'cite',
'the',
'2019',
'forest',
'survey',
'of',
'india',
'report',
'.',
'\n\n',
'how',
'mani',
'sapl',
'surviv',
'?',
'\n\n',
'uttar',
'pradesh',
'state',
'forest',
'minist',
'dara',
'singh',
'said',
'the',
'long',
```

```
'-',
'term',
'surviv',
'of',
'the',
'sapl',
'remain',
'a',
'concern',
',',
'ad',
'that',
'usual',
'onli',
'60',
'%',
'of',
'the',
'sapl',
'surviv',
'.',
'the',
'rest',
'succumb',
'to',
'diseas',
'or',
'lack',
'of',
'water',
'.',
'\n\n',
'howev',
',',
'he',
'said',
'that',
'\xa0',
'about',
'80',
'%',
'of',
'the',
'sapl',
'plant',
'in',
'the',
'last',
'four',
'annual',
'drive',
'have',
'surviv',
'.',
'\n\n',
'"',
'all',
'the',
'region',
'where',
'plantat',
'is',
'be',
'carri',
```

```
 'out',
 'have',
 'been',
 'geo',
 '-',
 'tag',
 'so',
 'that',
 'we',
 'can',
 'ascertain',
 'what',
 'exactli',
 'happen',
 ',',
 '"',
 'chauhan',
 'told',
 'the',
 'pioneer',
 'newspap',
 '.',
 '\n\n',
 '"',
 'these',
 'sapl',
 'carri',
 'QR',
 'code',
 'so',
 'that',
 'offici',
 'can',
 'maintain',
 'a',
 'record',
 'and',
 'verifi',
 'whether',
 'the',
 'sapl',
 'surviv',
 'or',
 'not',
 '.',
 'besid',
 ',',
 'team',
 'have',
 'been',
 'form',
 'to',
 'monitor',
 'progress',
 'of',
 'the',
 'plantat',
 'drive',
 ',',
 '"',
 'he',
 'said',
 '.',
 '\n\n',
```

```
 'what',
 'is',
 'the',
 'extent',
 'of',
 'india',
 "'s",
 'tree',
 'plant',
 'project',
 '?',
 '\n\n',
 'india',
 'ha',
 'vow',
 'to',
 'have',
 'a',
 'third',
 'of',
 'it',
 'total',
 'land',
 'area',
 ',',
 'or',
 '95',
 'million',
 'hectar',
 ',',
 'under',
 'forest',
 'and',
 'tree',
 'cover',
 'by',
 '2030',
 '.',
 '\n\n',
 'the',
 'govern',
 'ha',
 'alloc',
 '$',
 '6.2',
 'billion',
 '(',
 '€',
 '5.2',
 'billion',
 ')',
 'for',
 'the',
 'tree',
 '-',
 'plant',
 'across',
 'the',
 'countri',
 '.',
 '\n\n',
 'howev',
 ',',
 'industri',
```

```
'develop',
'and',
'a',
'rapidli',
'grow',
'popul',
'ha',
'put',
'further',
'stress',
'on',
'the',
'land',
'.']
```

In [304…

```
wl = WordNetLemmatizer()
wl_stemmed_words = []
for token in doc:
    wl_stemmed_words.append(wl.lemmatize(token.text))
wl_stemmed_words
```

Out[304…

```
['Millions',
 'of',
 'people',
 'in',
 'India',
 'took',
 'part',
 'in',
 'an',
 'annual',
 'tree',
 'planting',
 'drive',
 'Sunday',
 '.',
 'More',
 'than',
 '250',
 'million',
 'sapling',
 'were',
 'planted',
 'in',
 'a',
 'single',
 'day',
 'across',
 'the',
 'country',
 "'s",
 'most',
 '-',
 'populous',
 'state',
 '.',
 '\n\n',
 'The',
 'campaign',
 'wa',
 'led',
 'by',
 'Uttar',
```

```
 'Pradesh',
 'state',
 'government',
 'official',
 ',',
 'lawmaker',
 ',',
 'and',
 'activist',
 ',',
 'in',
 'a',
 'bid',
 'to',
 'reduce',
 'carbon',
 'emission',
 'and',
 'combat',
 'climate',
 'change',
 '.',
 '\n\n',
 'Where',
 'were',
 'the',
 'tree',
 'planted',
 '?',
 '\n\n',
 'The',
 'sapling',
 'were',
 'planted',
 'by',
 'volunteer',
 'in',
 'forest',
 ',',
 'farm',
 ',',
 'school',
 ',',
 'and',
 'along',
 'riverbank',
 'and',
 'highway',
 '.',
 '\n\n',
 '"',
 'We',
 'are',
 'committed',
 'to',
 'increasing',
 'the',
 'forest',
 'cover',
 'of',
 'Uttar',
 'Pradesh',
 'to',
 'over',
```

```
 '15',
 '%',
 'of',
 'the',
 'total',
 'land',
 'area',
 'in',
 'the',
 'next',
 'five',
 'year',
 ',',
 "''",
 'said',
 'state',
 'forest',
 'official',
 'Manoj',
 'Singh',
 '.',
 '\n\n',
 'According',
 'to',
 'another',
 'government',
 'official',
 ',',
 'the',
 'forest',
 'cover',
 'of',
 'the',
 'state',
 'ha',
 'increased',
 'over',
 'the',
 'last',
 'few',
 'year',
 '.',
 '\n\n',
 '"',
 'There',
 'ha',
 'been',
 'an',
 'increase',
 'of',
 '127',
 'sqare',
 'kilometer',
 '[',
 '79',
 'sqare',
 'mile',
 ']',
 '\xa0',
 'in',
 'the',
 'forest',
 'cover',
 'in',
```

```
 'Uttar',
 'Pradesh',
 'a',
 'compared',
 'to',
 '2017',
 ',',
 '"',
 'a',
 'state',
 'government',
 'spokesperson',
 'wa',
 'quoted',
 'a',
 'saying',
 'in',
 '\xa0',
 'The',
 'Indian',
 'Express',
 'newspaper',
 '.',
 '\n\n',
 '"',
 'There',
 'ha',
 'also',
 'been',
 'an',
 'increase',
 'in',
 'tree',
 'and',
 'plant',
 '.',
 'The',
 'tree',
 'cover',
 'ha',
 'increased',
 'to',
 '3.05',
 '%',
 ',',
 'a',
 'compared',
 'to',
 'the',
 'national',
 'average',
 'of',
 '2.89',
 '%',
 ',',
 '"',
 'the',
 'official',
 'said',
 ',',
 'citing',
 'the',
 '2019',
 'Forest',
```

```
 'Survey',
 'of',
 'India',
 'report',
 '.',
 '\n\n',
 'How',
 'many',
 'sapling',
 'survive',
 '?',
 '\n\n',
 'Uttar',
 'Pradesh',
 'State',
 'Forest',
 'Minister',
 'Dara',
 'Singh',
 'said',
 'the',
 'long',
 '-',
 'term',
 'survival',
 'of',
 'the',
 'sapling',
 'remains',
 'a',
 'concern',
 ',',
 'adding',
 'that',
 'usually',
 'only',
 '60',
 '%',
 'of',
 'the',
 'sapling',
 'survive',
 '.',
 'The',
 'rest',
 'succumb',
 'to',
 'disease',
 'or',
 'lack',
 'of',
 'water',
 '.',
 '\n\n',
 'However',
 ',',
 'he',
 'said',
 'that',
 '\xa0',
 'about',
 '80',
 '%',
 'of',
```

```
 'the',
 'sapling',
 'planted',
 'in',
 'the',
 'last',
 'four',
 'annual',
 'drive',
 'have',
 'survived',
 '.',
 '\n\n',
 '"',
 'All',
 'the',
 'region',
 'where',
 'plantation',
 'is',
 'being',
 'carried',
 'out',
 'have',
 'been',
 'geo',
 '-',
 'tagged',
 'so',
 'that',
 'we',
 'can',
 'ascertain',
 'what',
 'exactly',
 'happened',
 ',',
 '"',
 'Chauhan',
 'told',
 'The',
 'Pioneer',
 'newspaper',
 '.',
 '\n\n',
 '"',
 'These',
 'sapling',
 'carry',
 'QR',
 'code',
 'so',
 'that',
 'official',
 'can',
 'maintain',
 'a',
 'record',
 'and',
 'verify',
 'whether',
 'the',
 'sapling',
 'survived',
```

```
 'or',
 'not',
 '.',
 'Besides',
 ',',
 'team',
 'have',
 'been',
 'formed',
 'to',
 'monitor',
 'progress',
 'of',
 'the',
 'plantation',
 'drive',
 ',',
 '"',
 'he',
 'said',
 '.',
 '\n\n',
 'What',
 'is',
 'the',
 'extent',
 'of',
 'India',
 "'s",
 'tree',
 'planting',
 'project',
 '?',
 '\n\n',
 'India',
 'ha',
 'vowed',
 'to',
 'have',
 'a',
 'third',
 'of',
 'it',
 'total',
 'land',
 'area',
 ',',
 'or',
 '95',
 'million',
 'hectare',
 ',',
 'under',
 'forest',
 'and',
 'tree',
 'cover',
 'by',
 '2030',
 '.',
 '\n\n',
 'The',
 'government',
 'ha',
```

```
'allocated',
'$',
'6.2',
'billion',
'(',
'€',
'5.2',
'billion',
')',
'for',
'the',
'tree',
'-',
'planting',
'across',
'the',
'country',
'.',
'\n\n',
'However',
',',
'industrial',
'development',
'and',
'a',
'rapidly',
'growing',
'population',
'ha',
'put',
'further',
'stress',
'on',
'the',
'land',
'.']
```

In [305…

```
corpus = [text,text2]
def termfreq(corpus):
    dic={}

    for doc in corpus:
        #words = " ".join([token.text for token in doc])
        for word in doc.split():
            if word in dic:
                dic[word]+=1
            else:
                dic[word]=1

    for word,freq in dic.items():
        print(word,freq)
        dic[word]=freq/len(doc)
    print('Document size in number of words',len(doc))
    return dic
termfreq(corpus)
```

```
Millions 1
of 36
people 1
in 22
India 3
took 1
part 2
```

```
an 4
annual 2
tree 6
planting 5
drive 5
Sunday. 1
More 1
than 1
250 1
million 2
saplings 14
were 6
planted 7
a 15
single 2
day 1
across 3
the 64
country's 1
most-populous 1
state. 1
The 12
campaign 1
was 6
led 1
by 5
Uttar 4
Pradesh 4
state 6
government 8
officials, 1
lawmakers, 1
and 26
activists, 1
bid 1
to 23
reduce 1
carbon 1
emissions 1
combat 1
climate 1
change. 1
Where 1
trees 7
planted? 1
volunteers 3
forests, 1
farms, 1
schools, 1
along 2
riverbanks 1
highways. 1
"We 1
are 1
committed 1
increasing 1
forest 10
cover 5
over 3
15% 1
total 3
land 2
area 2
next 1
```

```
five 1
years,'' 1
said 3
official 2
Manoj 1
Singh. 1
According 1
another 1
official, 1
has 12
increased 2
last 2
few 1
years. 1
"There 2
been 6
increase 2
127 1
sqare 2
kilometers 1
[79 1
miles] 1
as 8
compared 2
2017," 1
spokesperson 1
quoted 1
saying 1
Indian 2
Express 2
newspaper. 2
also 1
plants. 1
3.05%, 1
national 1
average 1
2.89%," 1
said, 1
citing 1
2019 1
Forest 8
Survey 1
report. 1
How 1
many 2
survive? 1
State 2
Minister 1
Dara 1
Singh 1
long-term 1
survival 2
remains 1
concern, 1
adding 1
that 5
usually 1
only 1
60% 1
survive. 1
rest 1
succumb 1
disease 1
or 3
```

```
lack 1
water. 1
However, 3
he 2
about 3
80% 1
four 1
drives 2
have 4
survived. 1
"All 1
regions 1
where 1
plantation 10
is 3
being 1
carried 1
out 2
geo-tagged 1
so 2
we 1
can 3
ascertain 1
what 1
exactly 1
happened," 1
Chauhan 1
told 1
Pioneer 1
"These 1
carry 1
QR 1
codes 1
officials 1
maintain 2
record 3
verify 1
whether 1
survived 1
not. 1
Besides, 1
teams 1
formed 1
monitor 1
progress 1
drive," 1
said. 1
What 1
extent 1
India's 1
project? 1
vowed 1
third 1
its 3
area, 1
95 1
hectares, 1
under 4
2030. 1
allocated 1
$6.2 1
billion 1
(€5.2 1
billion) 1
```

```
for 5
tree-planting 1
country. 1
industrial 1
development 1
rapidly 1
growing 1
population 1
put 1
further 1
stress 1
on 8
land. 3
Between 1
2016 3
2019, 2
department 2
BJP 1
had 7
launched 1
‘Green 2
Maharashtra’ 1
with 7
aim 2
plant 3
50 3
crore 11
four-year 1
period. 1
In 6
October 1
claimed 2
it 3
surpassed 1
target 3
33 3
July-September 1
2019. 2
found 1
non-forest 1
agencies 2
– 6
such 2
gram 2
panchayats 2
which 7
tasked 1
not 3
uploaded 2
mandatory 1
audio-visual 1
proof 1
specially 1
created 2
portal. 1
Pune 5
Revenue 1
Division, 1
1.7 1
saplings; 1
however, 1
no 1
evidence 1
87 1
```

```
per 1
cent 1
(1.49 1
crore) 1
saplings. 2
Also, 1
59 1
involved 1
38 1
submitted 1
reports 1
This 1
year, 1
targets 1
set 4
comparatively 1
modest. 1
For 1
example, 1
Circle 2
comprises 2
three 2
divisions 1
Solapur 1
district 1
planned 1
17 1
lakh 3
may 1
meet 1
due 1
unavailability 1
funds. 1
Last 1
year 2
70 1
Division 1
six 1
talukas 1
namely 1
Maval, 1
Mulshi, 1
Daund, 1
Indapur, 1
Baramati 1
Havveli 1
preparations 1
done 1
4 4
special 1
emphasis 1
teakwood. 1
National 2
Policy 1
aims 2
emphasizes 1
at 2
maintaining 1
33% 1
country’s 1
geographical 1
green 1
cover. 1
view 1
```

```
this 3
programme 1
within 1
Maharashtra, 1
Maharashtra 2
Department 4
all 2
between 1
1st 4
July 5
7th, 1
2017 2
celebrate 1
‘Vanmohotsav’. 1
programme, 1
announced 1
2 1
resounding 1
success 1
final 1
reported 1
figure 1
2.82 1
day. 1
To 1
consistency 1
platform 1
without 1
affecting 1
momentum, 1
crore, 1
13 1
mission 1
shall 1
be 3
accomplished 1
consecutive 1
years 1
viz. 1
2017, 1
2018 1
will 2
during 1
Vanmohotsav, 1
7th 2
state-wide 1
involvement 1
departments 1
Students 1
Schools 1
Colleges, 1
NSS, 1
NCC, 1
CSR, 1
NGOs, 1
Railways, 1
Highways, 1
Defence, 1
NABARD 1
other 1
stakeholders 2
Society. 1
first 1
kind, 1
```

```
24-hour 1
toll 1
free 1
helpline 1
number 1
1926 1
called 2
‘Hello 1
Forest’ 1
up 1
provide 1
information 1
regarding 1
plantation, 2
protection 1
mass 1
awareness. 1
mobile 1
application 2
‘My 1
Plants’ 1
details 1
numbers, 1
species 1
location 1
into 1
computer 1
system 1
Department. 1
All 1
individual, 1
collective 1
organizational 1
level 1
should 1
download 1
use 1
their 1
work 1
through 1
application, 1
operational 1
from 2
July. 1
consonance 1
public 1
participation, 1
initiated 1
‘Maharashtra 1
Harit 1
Sena’/ 1
Army’ 1
body 1
dedicated 1
participate 1
protection, 1
activities 1
forest, 1
wildlife, 1
related 1
sectors 1
around 1
year. 1
Individuals 1
```

```
organisations 1
interested 1
volunteering 1
register 1
Green 1
Army 1
website 1
www.greenarmy.mahaforest.gov.in 1
An 1
integrated 1
place 1
ensure 1
seamless 1
successful 1
participation 1
society, 1
especially 1
public. 1
Document size in number of words 3790
```

Out[305…

```
{'Millions': 0.0002638522427440633,
 'of': 0.00949868073878628,
 'people': 0.0002638522427440633,
 'in': 0.005804749340369393,
 'India': 0.0007915567282321899,
 'took': 0.0002638522427440633,
 'part': 0.0005277044854881266,
 'an': 0.0010554089709762533,
 'annual': 0.0005277044854881266,
 'tree': 0.0015831134564643799,
 'planting': 0.0013192612137203166,
 'drive': 0.0013192612137203166,
 'Sunday.': 0.0002638522427440633,
 'More': 0.0002638522427440633,
 'than': 0.0002638522427440633,
 '250': 0.0002638522427440633,
 'million': 0.0005277044854881266,
 'saplings': 0.0036939313984168864,
 'were': 0.0015831134564643799,
 'planted': 0.0018469656992084432,
 'a': 0.00395778364116095,
 'single': 0.0005277044854881266,
 'day': 0.0002638522427440633,
 'across': 0.0007915567282321899,
 'the': 0.016886543535620052,
 "country's": 0.0002638522427440633,
 'most-populous': 0.0002638522427440633,
 'state.': 0.0002638522427440633,
 'The': 0.0031662269129287598,
 'campaign': 0.0002638522427440633,
 'was': 0.0015831134564643799,
 'led': 0.0002638522427440633,
 'by': 0.0013192612137203166,
 'Uttar': 0.0010554089709762533,
 'Pradesh': 0.0010554089709762533,
 'state': 0.0015831134564643799,
 'government': 0.0021108179419525065,
 'officials,': 0.0002638522427440633,
 'lawmakers,': 0.0002638522427440633,
 'and': 0.006860158311345646,
 'activists,': 0.0002638522427440633,
 'bid': 0.0002638522427440633,
 'to': 0.006068601583113457,
 'reduce': 0.0002638522427440633,
 'carbon': 0.0002638522427440633,
```

```
 'emissions': 0.0002638522427440633,
 'combat': 0.0002638522427440633,
 'climate': 0.0002638522427440633,
 'change.': 0.0002638522427440633,
 'Where': 0.0002638522427440633,
 'trees': 0.0018469656992084432,
 'planted?': 0.0002638522427440633,
 'volunteers': 0.0007915567282321899,
 'forests,': 0.0002638522427440633,
 'farms,': 0.0002638522427440633,
 'schools,': 0.0002638522427440633,
 'along': 0.0005277044854881266,
 'riverbanks': 0.0002638522427440633,
 'highways.': 0.0002638522427440633,
 '"We': 0.0002638522427440633,
 'are': 0.0002638522427440633,
 'committed': 0.0002638522427440633,
 'increasing': 0.0002638522427440633,
 'forest': 0.002638522427440633,
 'cover': 0.0013192612137203166,
 'over': 0.0007915567282321899,
 '15%': 0.0002638522427440633,
 'total': 0.0007915567282321899,
 'land': 0.0005277044854881266,
 'area': 0.0005277044854881266,
 'next': 0.0002638522427440633,
 'five': 0.0002638522427440633,
"years,'"": 0.0002638522427440633,
 'said': 0.0007915567282321899,
 'official': 0.0005277044854881266,
 'Manoj': 0.0002638522427440633,
 'Singh.': 0.0002638522427440633,
 'According': 0.0002638522427440633,
 'another': 0.0002638522427440633,
 'official,': 0.0002638522427440633,
 'has': 0.0031662269129287598,
 'increased': 0.0005277044854881266,
 'last': 0.0005277044854881266,
 'few': 0.0002638522427440633,
 'years.': 0.0002638522427440633,
 '"There': 0.0005277044854881266,
 'been': 0.0015831134564643799,
 'increase': 0.0005277044854881266,
 '127': 0.0002638522427440633,
 'sqare': 0.0005277044854881266,
 'kilometers': 0.0002638522427440633,
 '[79': 0.0002638522427440633,
 'miles]': 0.0002638522427440633,
 'as': 0.0021108179419525065,
 'compared': 0.0005277044854881266,
 '2017,"': 0.0002638522427440633,
 'spokesperson': 0.0002638522427440633,
 'quoted': 0.0002638522427440633,
 'saying': 0.0002638522427440633,
 'Indian': 0.0005277044854881266,
 'Express': 0.0005277044854881266,
 'newspaper.': 0.0005277044854881266,
 'also': 0.0002638522427440633,
 'plants.': 0.0002638522427440633,
 '3.05%,': 0.0002638522427440633,
 'national': 0.0002638522427440633,
 'average': 0.0002638522427440633,
 '2.89%,"': 0.0002638522427440633,
 'said,': 0.0002638522427440633,
```

```
'citing': 0.0002638522427440633,
'2019': 0.0002638522427440633,
'Forest': 0.0021108179419525065,
'Survey': 0.0002638522427440633,
'report.': 0.0002638522427440633,
'How': 0.0002638522427440633,
'many': 0.0005277044854881266,
'survive?': 0.0002638522427440633,
'State': 0.0005277044854881266,
'Minister': 0.0002638522427440633,
'Dara': 0.0002638522427440633,
'Singh': 0.0002638522427440633,
'long-term': 0.0002638522427440633,
'survival': 0.0005277044854881266,
'remains': 0.0002638522427440633,
'concern,': 0.0002638522427440633,
'adding': 0.0002638522427440633,
'that': 0.0013192612137203166,
'usually': 0.0002638522427440633,
'only': 0.0002638522427440633,
'60%': 0.0002638522427440633,
'survive.': 0.0002638522427440633,
'rest': 0.0002638522427440633,
'succumb': 0.0002638522427440633,
'disease': 0.0002638522427440633,
'or': 0.0007915567282321899,
'lack': 0.0002638522427440633,
'water.': 0.0002638522427440633,
'However,': 0.0007915567282321899,
'he': 0.0005277044854881266,
'about': 0.0007915567282321899,
'80%': 0.0002638522427440633,
'four': 0.0002638522427440633,
'drives': 0.0005277044854881266,
'have': 0.0010554089709762533,
'survived.': 0.0002638522427440633,
'"All': 0.0002638522427440633,
'regions': 0.0002638522427440633,
'where': 0.0002638522427440633,
'plantation': 0.002638522427440633,
'is': 0.0007915567282321899,
'being': 0.0002638522427440633,
'carried': 0.0002638522427440633,
'out': 0.0005277044854881266,
'geo-tagged': 0.0002638522427440633,
'so': 0.0005277044854881266,
'we': 0.0002638522427440633,
'can': 0.0007915567282321899,
'ascertain': 0.0002638522427440633,
'what': 0.0002638522427440633,
'exactly': 0.0002638522427440633,
'happened,"': 0.0002638522427440633,
'Chauhan': 0.0002638522427440633,
'told': 0.0002638522427440633,
'Pioneer': 0.0002638522427440633,
'"These': 0.0002638522427440633,
'carry': 0.0002638522427440633,
'QR': 0.0002638522427440633,
'codes': 0.0002638522427440633,
'officials': 0.0002638522427440633,
'maintain': 0.0005277044854881266,
'record': 0.0007915567282321899,
'verify': 0.0002638522427440633,
'whether': 0.0002638522427440633,
```

```
 'survived': 0.0002638522427440633,
 'not.': 0.0002638522427440633,
 'Besides,': 0.0002638522427440633,
 'teams': 0.0002638522427440633,
 'formed': 0.0002638522427440633,
 'monitor': 0.0002638522427440633,
 'progress': 0.0002638522427440633,
 'drive,"': 0.0002638522427440633,
 'said.': 0.0002638522427440633,
 'What': 0.0002638522427440633,
 'extent': 0.0002638522427440633,
 "India's": 0.0002638522427440633,
 'project?': 0.0002638522427440633,
 'vowed': 0.0002638522427440633,
 'third': 0.0002638522427440633,
 'its': 0.0007915567282321899,
 'area,': 0.0002638522427440633,
 '95': 0.000263852242744063,
 'hectares,': 0.0002638522427440633,
 'under': 0.0010554089709762533,
 '2030.': 0.0002638522427440633,
 'allocated': 0.0002638522427440633,
 '$6.2': 0.000263852242744063,
 'billion': 0.0002638522427440633,
 '(€5.2': 0.0002638522427440633,
 'billion)': 0.0002638522427440633,
 'for': 0.0013192612137203166,
 'tree-planting': 0.0002638522427440633,
 'country.': 0.000263852242744063,
 'industrial': 0.0002638522427440633,
 'development': 0.0002638522427440633,
 'rapidly': 0.0002638522427440633,
 'growing': 0.0002638522427440633,
 'population': 0.0002638522427440633,
 'put': 0.0002638522427440633,
 'further': 0.0002638522427440633,
 'stress': 0.0002638522427440633,
 'on': 0.0021108179419525065,
 'land.': 0.0007915567282321899,
 'Between': 0.0002638522427440633,
 '2016': 0.0007915567282321899,
 '2019,': 0.0005277044854881266,
 'department': 0.0005277044854881266,
 'BJP': 0.0002638522427440633,
 'had': 0.0018469656992084432,
 'launched': 0.0002638522427440633,
 '‘Green': 0.0005277044854881266,
 'Maharashtra’': 0.0002638522427440633,
 'with': 0.0018469656992084432,
 'aim': 0.0005277044854881266,
 'plant': 0.0007915567282321899,
 '50': 0.0007915567282321899,
 'crore': 0.0029023746701846965,
 'four-year': 0.0002638522427440633,
 'period.': 0.0002638522427440633,
 'In': 0.0015831134564643799,
 'October': 0.0002638522427440633,
 'claimed': 0.0005277044854881266,
 'it': 0.0007915567282321899,
 'surpassed': 0.0002638522427440633,
 'target': 0.0007915567282321899,
 '33': 0.0007915567282321899,
 'July-September': 0.0002638522427440633,
 '2019.': 0.0005277044854881266,
```

'found': 0.0002638522427440633,
'non-forest': 0.0002638522427440633,
'agencies': 0.0005277044854881266,
'—': 0.0015831134564643799,
'such': 0.0005277044854881266,
'gram': 0.0005277044854881266,
'panchayats': 0.0005277044854881266,
'which': 0.0018469656992084432,
'tasked': 0.0002638522427440633,
'not': 0.0007915567282321899,
'uploaded': 0.0005277044854881266,
'mandatory': 0.0002638522427440633,
'audio-visual': 0.0002638522427440633,
'proof': 0.0002638522427440633,
'specially': 0.0002638522427440633,
'created': 0.0005277044854881266,
'portal.': 0.0002638522427440633,
'Pune': 0.0013192612137203166,
'Revenue': 0.0002638522427440633,
'Division,': 0.0002638522427440633,
'1.7': 0.0002638522427440633,
'saplings;': 0.0002638522427440633,
'however,': 0.0002638522427440633,
'no': 0.0002638522427440633,
'evidence': 0.0002638522427440633,
'87': 0.0002638522427440633,
'per': 0.0002638522427440633,
'cent': 0.0002638522427440633,
'(1.49': 0.0002638522427440633,
'crore)': 0.0002638522427440633,
'saplings.': 0.0005277044854881266,
'Also,': 0.0002638522427440633,
'59': 0.0002638522427440633,
'involved': 0.0002638522427440633,
'38': 0.0002638522427440633,
'submitted': 0.0002638522427440633,
'reports': 0.0002638522427440633,
'This': 0.0002638522427440633,
'year,': 0.0002638522427440633,
'targets': 0.0002638522427440633,
'set': 0.0010554089709762533,
'comparatively': 0.0002638522427440633,
'modest.': 0.0002638522427440633,
'For': 0.0002638522427440633,
'example,': 0.0002638522427440633,
'Circle': 0.0005277044854881266,
'comprises': 0.0005277044854881266,
'three': 0.0005277044854881266,
'divisions': 0.0002638522427440633,
'Solapur': 0.0002638522427440633,
'district': 0.0002638522427440633,
'planned': 0.0002638522427440633,
'17': 0.0002638522427440633,
'lakh': 0.0007915567282321899,
'may': 0.0002638522427440633,
'meet': 0.0002638522427440633,
'due': 0.0002638522427440633,
'unavailability': 0.0002638522427440633,
'funds.': 0.0002638522427440633,
'Last': 0.0002638522427440633,
'year': 0.0005277044854881266,
'70': 0.0002638522427440633,
'Division': 0.0002638522427440633,
'six': 0.0002638522427440633,

```
'talukas': 0.0002638522427440633,
'namely': 0.0002638522427440633,
'Maval,': 0.0002638522427440633,
'Mulshi,': 0.0002638522427440633,
'Daund,': 0.0002638522427440633,
'Indapur,': 0.0002638522427440633,
'Baramati': 0.0002638522427440633,
'Havveli': 0.0002638522427440633,
'preparations': 0.0002638522427440633,
'done': 0.0002638522427440633,
'4': 0.0010554089709762533,
'special': 0.0002638522427440633,
'emphasis': 0.0002638522427440633,
'teakwood.': 0.0002638522427440633,
'National': 0.0005277044854881266,
'Policy': 0.0002638522427440633,
'aims': 0.0005277044854881266,
'emphasizes': 0.0002638522427440633,
'at': 0.0005277044854881266,
'maintaining': 0.0002638522427440633,
'33%': 0.0002638522427440633,
'country’s': 0.0002638522427440633,
'geographical': 0.0002638522427440633,
'green': 0.0002638522427440633,
'cover.': 0.0002638522427440633,
'view': 0.0002638522427440633,
'this': 0.0007915567282321899,
'programme': 0.0002638522427440633,
'within': 0.0002638522427440633,
'Maharashtra,': 0.0002638522427440633,
'Maharashtra': 0.0005277044854881266,
'Department': 0.0010554089709762533,
'all': 0.0005277044854881266,
'between': 0.0002638522427440633,
'1st': 0.0010554089709762533,
'July': 0.0013192612137203166,
'7th,': 0.0002638522427440633,
'2017': 0.0005277044854881266,
'celebrate': 0.0002638522427440633,
'‘Vanmohotsav’.': 0.0002638522427440633,
'programme,': 0.0002638522427440633,
'announced': 0.0002638522427440633,
'2': 0.0002638522427440633,
'resounding': 0.0002638522427440633,
'success': 0.0002638522427440633,
'final': 0.0002638522427440633,
'reported': 0.0002638522427440633,
'figure': 0.0002638522427440633,
'2.82': 0.0002638522427440633,
'day.': 0.0002638522427440633,
'To': 0.0002638522427440633,
'consistency': 0.0002638522427440633,
'platform': 0.0002638522427440633,
'without': 0.0002638522427440633,
'affecting': 0.0002638522427440633,
'momentum,': 0.0002638522427440633,
'crore,': 0.0002638522427440633,
'13': 0.0002638522427440633,
'mission': 0.0002638522427440633,
'shall': 0.0002638522427440633,
'be': 0.0007915567282321899,
'accomplished': 0.0002638522427440633,
'consecutive': 0.0002638522427440633,
'years': 0.0002638522427440633,
```

```
'viz.': 0.0002638522427440633,
'2017,': 0.0002638522427440633,
'2018': 0.0002638522427440633,
'will': 0.0005277044854881266,
'during': 0.0002638522427440633,
'Vanmohotsav,': 0.0002638522427440633,
'7th': 0.0005277044854881266,
'state-wide': 0.0002638522427440633,
'involvement': 0.0002638522427440633,
'departments': 0.0002638522427440633,
'Students': 0.0002638522427440633,
'Schools': 0.0002638522427440633,
'Colleges,': 0.0002638522427440633,
'NSS,': 0.0002638522427440633,
'NCC,': 0.0002638522427440633,
'CSR,': 0.0002638522427440633,
'NGOs,': 0.0002638522427440633,
'Railways,': 0.0002638522427440633,
'Highways,': 0.0002638522427440633,
'Defence,': 0.0002638522427440633,
'NABARD': 0.0002638522427440633,
'other': 0.0002638522427440633,
'stakeholders': 0.0005277044854881266,
'Society.': 0.0002638522427440633,
'first': 0.0002638522427440633,
'kind,': 0.0002638522427440633,
'24-hour': 0.0002638522427440633,
'toll': 0.0002638522427440633,
'free': 0.0002638522427440633,
'helpline': 0.0002638522427440633,
'number': 0.0002638522427440633,
'1926': 0.0002638522427440633,
'called': 0.0005277044854881266,
'‘Hello': 0.0002638522427440633,
'Forest’': 0.0002638522427440633,
'up': 0.0002638522427440633,
'provide': 0.0002638522427440633,
'information': 0.0002638522427440633,
'regarding': 0.0002638522427440633,
'plantation,': 0.0005277044854881266,
'protection': 0.0002638522427440633,
'mass': 0.0002638522427440633,
'awareness.': 0.0002638522427440633,
'mobile': 0.0002638522427440633,
'application': 0.0005277044854881266,
'‘My': 0.0002638522427440633,
'Plants’': 0.0002638522427440633,
'details': 0.0002638522427440633,
'numbers,': 0.0002638522427440633,
'species': 0.0002638522427440633,
'location': 0.0002638522427440633,
'into': 0.0002638522427440633,
'computer': 0.0002638522427440633,
'system': 0.0002638522427440633,
'Department.': 0.0002638522427440633,
'All': 0.0002638522427440633,
'individual,': 0.0002638522427440633,
'collective': 0.0002638522427440633,
'organizational': 0.0002638522427440633,
'level': 0.0002638522427440633,
'should': 0.0002638522427440633,
'download': 0.0002638522427440633,
'use': 0.0002638522427440633,
'their': 0.0002638522427440633,
```

```
'work': 0.0002638522427440633,
'through': 0.0002638522427440633,
'application,': 0.0002638522427440633,
'operational': 0.0002638522427440633,
'from': 0.0005277044854881266,
'July.': 0.0002638522427440633,
'consonance': 0.0002638522427440633,
'public': 0.0002638522427440633,
'participation,': 0.0002638522427440633,
'initiated': 0.0002638522427440633,
'‘Maharashtra': 0.0002638522427440633,
'Harit': 0.0002638522427440633,
'Sena’/': 0.0002638522427440633,
'Army’': 0.0002638522427440633,
'body': 0.0002638522427440633,
'dedicated': 0.0002638522427440633,
'participate': 0.0002638522427440633,
'protection,': 0.0002638522427440633,
'activities': 0.0002638522427440633,
'forest,': 0.0002638522427440633,
'wildlife,': 0.0002638522427440633,
'related': 0.0002638522427440633,
'sectors': 0.0002638522427440633,
'around': 0.0002638522427440633,
'year.': 0.0002638522427440633,
'Individuals': 0.0002638522427440633,
'organisations': 0.0002638522427440633,
'interested': 0.0002638522427440633,
'volunteering': 0.0002638522427440633,
'register': 0.0002638522427440633,
'Green': 0.0002638522427440633,
'Army': 0.0002638522427440633,
'website': 0.0002638522427440633,
'www.greenarmy.mahaforest.gov.in': 0.0002638522427440633,
'An': 0.0002638522427440633,
'integrated': 0.0002638522427440633,
'place': 0.0002638522427440633,
'ensure': 0.0002638522427440633,
'seamless': 0.0002638522427440633,
'successful': 0.0002638522427440633,
'participation': 0.0002638522427440633,
'society,': 0.0002638522427440633,
'especially': 0.0002638522427440633,
'public.': 0.0002638522427440633}
```

In [306…

```
#print(text)
```

In [307…

```
#print(text2)
```

IDF

TF-IDF stands for “Term Frequency — Inverse Data Frequency”. First, we will learn what this term means mathematically.

Term Frequency (tf): gives us the frequency of the word in each document in the corpus. It is the ratio of number of times the word appears in a document compared to the total number of words in that document. It increases as the number of occurrences of that word within the document increases. Each document has its own tf.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

Inverse Data Frequency (idf): used to calculate the weight of rare words across all documents in the corpus. The words that occur rarely in the corpus have a high IDF score. It is given by the equation below.

$$idf(w) = log(\frac{N}{df_t})$$

Combining these two we come up with the TF-IDF score (w) for a word in a document in the corpus. It is the product of tf and idf:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

In [308…

```
import re
import nltk
from nltk.corpus import stopwords
doc_text = " "

def preprocess_docs(text):
    text = str(text).lower()
    #print(text)
    text = re.sub('[^a-zA-z0-9\s]','',str(text))
    text = text.split()
    #print(text)
    text = [wl.lemmatize(word) for word in text if not word in stopwords.words('engl
    new_text = ' '.join(text)
    #print("\ndoc : ", new_text)
        #doc_text = doc_text + new_text
            #doc_text = doc_text + new_text
        #print(new_text)
    return new_text
corpus = [text,text2]
#print(text)
text1 = preprocess_docs(text)
text2 = preprocess_docs(text2)
print("\n doc1",text1,"\nDoc 2",text2)
```

```
 doc1 million people india took part annual tree planting drive sunday 250 million s
apling planted single day across country mostpopulous state campaign led uttar prade
sh state government official lawmaker activist bid reduce carbon emission combat cli
mate change tree planted sapling planted volunteer forest farm school along riverban
```

k highway committed increasing forest cover uttar pradesh 15 total land area next fi ve year said state forest official manoj singh according another government official forest cover state increased last year increase 127 sqare kilometer [79 sqare miles] forest cover uttar pradesh compared 2017 state government spokesperson quoted saying indian express newspaper also increase tree plant tree cover increased 305 compared national average 289 official said citing 2019 forest survey india report many sapli ng survive uttar pradesh state forest minister dara singh said longterm survival sap ling remains concern adding usually 60 sapling survive rest succumb disease lack wat er however said 80 sapling planted last four annual drive survived region plantation carried geotagged ascertain exactly happened chauhan told pioneer newspaper sapling carry qr code official maintain record verify whether sapling survived besides team formed monitor progress plantation drive said extent india tree planting project ind ia vowed third total land area 95 million hectare forest tree cover 2030 government allocated 62 billion 52 billion treeplanting across country however industrial devel opment rapidly growing population put stress land
Doc 2 2016 2019 state forest department bjp government launched green maharashtra dr ive aim plant 50 crore tree across state fouryear period october 2019 government cla imed surpassed target planting 33 crore tree julyseptember 2019 indian express found nonforest agency gram panchayat tasked planting tree uploaded mandatory audiovisual proof tree plantation drive specially created portal pune revenue division claimed g ram panchayat planted 17 crore sapling however evidence uploaded 87 per cent 149 cro re sapling also 59 government agency involved drive many 38 submitted survival repor t sapling year target set forest department comparatively modest example pune circle comprises three division pune solapur district planned plant 17 lakh sapling forest land however may meet target due unavailability fund last year pune circle planted 7 0 lakh sapling forest land pune division comprises six talukas namely maval mulshi d aund indapur baramati havveli preparation done plantation 4 lakh tree special emphas is teakwood national forest policy aim emphasizes maintaining 33 country geographica l area forest green cover view part 50 crore plantation programme within maharashtra maharashtra forest department aim plant 4 crore sapling state 1st july 7th 2017 cele brate vanmohotsav plantation programme announced 2016 aim planting 2 crore tree 1st july 2016 resounding success final total reported figure 282 crore sapling planted s ingle day maintain consistency platform without affecting momentum forest department set target plantation 4 crore 13 crore 33 crore sapling mission 50 crore plantation shall accomplished three consecutive year viz 2017 2018 2019 4 crore sapling year 20 17 planted vanmohotsav july 1st july 7th statewide drive involvement 33 government d epartment along student school college n ncc csr ngo railway national highway defenc e nabard stakeholder society first kind 24hour toll free helpline number 1926 called hello forest set provide information regarding plantation protection mass awareness forest department created mobile application called plant record detail plantation n umber specie location computer system forest department volunteer individual collect ive organizational level download use application record tree plantation work applic ation operational 1st july 7th july consonance public participation maharashtra fore st department initiated maharashtra harit sena green army body dedicated volunteer p articipate plantation protection activity forest wildlife related sector around year individual organisation interested volunteering register green army website wwwgreen armymahaforestgovin integrated drive set place ensure seamless successful participat ion stakeholder society especially public

In [309…

```
first= text1
second= text2
#split so each word have their own string
first = first.split(" ")
second= second.split(" ")
#print(first,second)
total= set(first).union(set(second))
#print(total)
wordDictA = dict.fromkeys(total, 0)
wordDictB = dict.fromkeys(total, 0)
for word in first:
    wordDictA[word]+=1

for word in second:
```

```
    wordDictB[word]+=1
#put them in a dataframe and then view the result:
pd.DataFrame([wordDictA, wordDictB])
```

Out[309…

	15	forest	celebrate	citing	many	agency	planting	six	62	operational	...	resounding	mob
0	1	8	0	1	1	0	2	0	1	0	...	0	
1	0	13	1	0	1	2	3	1	0	1	...	1	

2 rows × 321 columns

In [310…

```
def computeTF(wordDict, bow):
    tfDict = {}
    bowCount = len(bow)
    for word, count in wordDict.items():
        tfDict[word] = count/float(bowCount)
    return tfDict
#running our sentences through the tf function:
tfFirst = computeTF(wordDictA, first)
tfSecond = computeTF(wordDictB, second)
#Converting to dataframe for visualization
tf_df= pd.DataFrame([tfFirst, tfSecond])
tf_df.head()
```

Out[310…

	15	forest	celebrate	citing	many	agency	planting	six	62	operatio
0	0.004566	0.036530	0.000000	0.004566	0.004566	0.00000	0.009132	0.000000	0.004566	0.0000
1	0.000000	0.035422	0.002725	0.000000	0.002725	0.00545	0.008174	0.002725	0.000000	0.0027

2 rows × 321 columns

In [311…

```
def computeIDF(docList):
    idfDict = {}
    N = len(docList)

    idfDict = dict.fromkeys(docList[0].keys(), 0)
    for doc in docList:
        for word, val in doc.items():
            if val > 0:
                idfDict[word] += 1

    for word, val in idfDict.items():
        idfDict[word] = math.log10(N / float(val))

    return idfDict
#inputing our sentences in the log file
idfs = computeIDF([wordDictA, wordDictB])
```

In [312…

```
idfs
```

Out[312…

```
{'15': 0.3010299956639812,
 'forest': 0.0,
 'celebrate': 0.3010299956639812,
 'citing': 0.3010299956639812,
 'many': 0.0,
```

```
'agency': 0.3010299956639812,
'planting': 0.0,
'six': 0.3010299956639812,
'62': 0.3010299956639812,
'operational': 0.3010299956639812,
'dedicated': 0.3010299956639812,
'figure': 0.3010299956639812,
'allocated': 0.3010299956639812,
'sapling': 0.0,
'department': 0.3010299956639812,
'body': 0.3010299956639812,
'mulshi': 0.3010299956639812,
'teakwood': 0.3010299956639812,
'division': 0.3010299956639812,
'crore': 0.3010299956639812,
'bjp': 0.3010299956639812,
'district': 0.3010299956639812,
'unavailability': 0.3010299956639812,
'sqare': 0.3010299956639812,
'uploaded': 0.3010299956639812,
'consecutive': 0.3010299956639812,
'kilometer': 0.3010299956639812,
'mandatory': 0.3010299956639812,
'exactly': 0.3010299956639812,
'free': 0.3010299956639812,
'consistency': 0.3010299956639812,
'record': 0.0,
'individual': 0.3010299956639812,
'created': 0.3010299956639812,
'place': 0.3010299956639812,
'hectare': 0.3010299956639812,
'2017': 0.0,
'planned': 0.3010299956639812,
'circle': 0.3010299956639812,
'detail': 0.3010299956639812,
'interested': 0.3010299956639812,
'increase': 0.3010299956639812,
'305': 0.3010299956639812,
'harit': 0.3010299956639812,
'cent': 0.3010299956639812,
'people': 0.3010299956639812,
'farm': 0.3010299956639812,
'concern': 0.3010299956639812,
'download': 0.3010299956639812,
'region': 0.3010299956639812,
'within': 0.3010299956639812,
'13': 0.3010299956639812,
'related': 0.3010299956639812,
'success': 0.3010299956639812,
'sector': 0.3010299956639812,
'said': 0.3010299956639812,
'policy': 0.3010299956639812,
'survived': 0.3010299956639812,
'level': 0.3010299956639812,
'increased': 0.3010299956639812,
'planted': 0.0,
'organizational': 0.3010299956639812,
'maharashtra': 0.3010299956639812,
'project': 0.3010299956639812,
'succumb': 0.3010299956639812,
'127': 0.3010299956639812,
'indian': 0.0,
'lack': 0.3010299956639812,
'toll': 0.3010299956639812,
```

```
'95': 0.3010299956639812,
'without': 0.3010299956639812,
'70': 0.3010299956639812,
'1926': 0.3010299956639812,
'day': 0.0,
'viz': 0.3010299956639812,
'carbon': 0.3010299956639812,
'minister': 0.3010299956639812,
'school': 0.0,
'specially': 0.3010299956639812,
'code': 0.3010299956639812,
'team': 0.3010299956639812,
'17': 0.3010299956639812,
'provide': 0.3010299956639812,
'single': 0.0,
'quoted': 0.3010299956639812,
'vowed': 0.3010299956639812,
'33': 0.3010299956639812,
'specie': 0.3010299956639812,
'audiovisual': 0.3010299956639812,
'country': 0.0,
'289': 0.3010299956639812,
'change': 0.3010299956639812,
'revenue': 0.3010299956639812,
'1st': 0.3010299956639812,
'singh': 0.3010299956639812,
'havveli': 0.3010299956639812,
'four': 0.3010299956639812,
'wildlife': 0.3010299956639812,
'regarding': 0.3010299956639812,
'computer': 0.3010299956639812,
'application': 0.3010299956639812,
'drive': 0.0,
'[79': 0.3010299956639812,
'society': 0.3010299956639812,
'increasing': 0.3010299956639812,
'60': 0.3010299956639812,
'railway': 0.3010299956639812,
'portal': 0.3010299956639812,
'also': 0.0,
'whether': 0.3010299956639812,
'protection': 0.3010299956639812,
'army': 0.3010299956639812,
'national': 0.0,
'282': 0.3010299956639812,
'stress': 0.3010299956639812,
'formed': 0.3010299956639812,
'149': 0.3010299956639812,
'told': 0.3010299956639812,
'comprises': 0.3010299956639812,
'kind': 0.3010299956639812,
'however': 0.0,
'special': 0.3010299956639812,
'50': 0.3010299956639812,
'volunteering': 0.3010299956639812,
'done': 0.3010299956639812,
'report': 0.0,
'reported': 0.3010299956639812,
'stakeholder': 0.3010299956639812,
'survey': 0.3010299956639812,
'launched': 0.3010299956639812,
'saying': 0.3010299956639812,
'n': 0.3010299956639812,
'announced': 0.3010299956639812,
```

```
'called': 0.3010299956639812,
'next': 0.3010299956639812,
'activity': 0.3010299956639812,
'participate': 0.3010299956639812,
'uttar': 0.3010299956639812,
'year': 0.0,
'register': 0.3010299956639812,
'nabard': 0.3010299956639812,
'happened': 0.3010299956639812,
'52': 0.3010299956639812,
'took': 0.3010299956639812,
'emission': 0.3010299956639812,
'rest': 0.3010299956639812,
'fouryear': 0.3010299956639812,
'water': 0.3010299956639812,
'2030': 0.3010299956639812,
'set': 0.3010299956639812,
'last': 0.0,
'80': 0.3010299956639812,
'activist': 0.3010299956639812,
'campaign': 0.3010299956639812,
'led': 0.3010299956639812,
'growing': 0.3010299956639812,
'programme': 0.3010299956639812,
'awareness': 0.3010299956639812,
'besides': 0.3010299956639812,
'five': 0.3010299956639812,
'2019': 0.0,
'disease': 0.3010299956639812,
'average': 0.3010299956639812,
'participation': 0.3010299956639812,
'vanmohotsav': 0.3010299956639812,
'especially': 0.3010299956639812,
'initiated': 0.3010299956639812,
'website': 0.3010299956639812,
'sena': 0.3010299956639812,
'maintaining': 0.3010299956639812,
'student': 0.3010299956639812,
'location': 0.3010299956639812,
'reduce': 0.3010299956639812,
'proof': 0.3010299956639812,
'defence': 0.3010299956639812,
'emphasis': 0.3010299956639812,
'survive': 0.3010299956639812,
'cover': 0.0,
'according': 0.3010299956639812,
'successful': 0.3010299956639812,
'dara': 0.3010299956639812,
'use': 0.3010299956639812,
'july': 0.3010299956639812,
'adding': 0.3010299956639812,
'remains': 0.3010299956639812,
'express': 0.0,
'250': 0.3010299956639812,
'longterm': 0.3010299956639812,
'involvement': 0.3010299956639812,
'verify': 0.3010299956639812,
'million': 0.3010299956639812,
'billion': 0.3010299956639812,
'ngo': 0.3010299956639812,
'due': 0.3010299956639812,
'state': 0.0,
'first': 0.3010299956639812,
'evidence': 0.3010299956639812,
```

```
'official': 0.3010299956639812,
'october': 0.3010299956639812,
'example': 0.3010299956639812,
'bid': 0.3010299956639812,
'gram': 0.3010299956639812,
'ensure': 0.3010299956639812,
'submitted': 0.3010299956639812,
'2016': 0.3010299956639812,
'consonance': 0.3010299956639812,
'tree': 0.0,
'sunday': 0.3010299956639812,
'public': 0.3010299956639812,
'newspaper': 0.3010299956639812,
'committed': 0.3010299956639812,
'development': 0.3010299956639812,
'helpline': 0.3010299956639812,
'plant': 0.0,
'talukas': 0.3010299956639812,
'lawmaker': 0.3010299956639812,
'shall': 0.3010299956639812,
'87': 0.3010299956639812,
'baramati': 0.3010299956639812,
'annual': 0.3010299956639812,
'progress': 0.3010299956639812,
'government': 0.0,
'aim': 0.3010299956639812,
'chauhan': 0.3010299956639812,
'integrated': 0.3010299956639812,
'qr': 0.3010299956639812,
'found': 0.3010299956639812,
'college': 0.3010299956639812,
'comparatively': 0.3010299956639812,
'modest': 0.3010299956639812,
'land': 0.0,
'usually': 0.3010299956639812,
'pioneer': 0.3010299956639812,
'monitor': 0.3010299956639812,
'total': 0.0,
'manoj': 0.3010299956639812,
'extent': 0.3010299956639812,
'claimed': 0.3010299956639812,
'may': 0.3010299956639812,
'around': 0.3010299956639812,
'tasked': 0.3010299956639812,
'involved': 0.3010299956639812,
'climate': 0.3010299956639812,
'volunteer': 0.0,
'hello': 0.3010299956639812,
'geographical': 0.3010299956639812,
'third': 0.3010299956639812,
'put': 0.3010299956639812,
'panchayat': 0.3010299956639812,
'preparation': 0.3010299956639812,
'area': 0.0,
'lakh': 0.3010299956639812,
'along': 0.0,
'seamless': 0.3010299956639812,
'daund': 0.3010299956639812,
'highway': 0.0,
'38': 0.3010299956639812,
'per': 0.3010299956639812,
'wwwgreenarmymahaforestgovin': 0.3010299956639812,
'mostpopulous': 0.3010299956639812,
'spokesperson': 0.3010299956639812,
```

```
'population': 0.3010299956639812,
'carried': 0.3010299956639812,
'julyseptember': 0.3010299956639812,
'plantation': 0.0,
'work': 0.3010299956639812,
'csr': 0.3010299956639812,
'2': 0.3010299956639812,
'emphasizes': 0.3010299956639812,
'period': 0.3010299956639812,
'combat': 0.3010299956639812,
'accomplished': 0.3010299956639812,
'view': 0.3010299956639812,
'final': 0.3010299956639812,
'compared': 0.3010299956639812,
'4': 0.3010299956639812,
'24hour': 0.3010299956639812,
'information': 0.3010299956639812,
'number': 0.3010299956639812,
'ascertain': 0.3010299956639812,
'solapur': 0.3010299956639812,
'pradesh': 0.3010299956639812,
'riverbank': 0.3010299956639812,
'collective': 0.3010299956639812,
'three': 0.3010299956639812,
'fund': 0.3010299956639812,
'treeplanting': 0.3010299956639812,
'platform': 0.3010299956639812,
'across': 0.0,
'industrial': 0.3010299956639812,
'2018': 0.3010299956639812,
'7th': 0.3010299956639812,
'survival': 0.0,
'surpassed': 0.3010299956639812,
'momentum': 0.3010299956639812,
'organisation': 0.3010299956639812,
'miles]': 0.3010299956639812,
'green': 0.3010299956639812,
'meet': 0.3010299956639812,
'mission': 0.3010299956639812,
'mass': 0.3010299956639812,
'maval': 0.3010299956639812,
'part': 0.0,
'affecting': 0.3010299956639812,
'another': 0.3010299956639812,
'maintain': 0.0,
'nonforest': 0.3010299956639812,
'system': 0.3010299956639812,
'india': 0.3010299956639812,
'rapidly': 0.3010299956639812,
'indapur': 0.3010299956639812,
'resounding': 0.3010299956639812,
'mobile': 0.3010299956639812,
'59': 0.3010299956639812,
'namely': 0.3010299956639812,
'ncc': 0.3010299956639812,
'statewide': 0.3010299956639812,
'geotagged': 0.3010299956639812,
'carry': 0.3010299956639812,
'target': 0.3010299956639812,
'pune': 0.3010299956639812}
```

In [313…

```
def computeTFIDF(tfBow, idfs):
    tfidf = {}
```

```
    for word, val in tfBow.items():
        tfidf[word] = val*idfs[word]
    return tfidf
#running our two sentences through the IDF:
idfFirst = computeTFIDF(tfFirst, idfs)
idfSecond = computeTFIDF(tfSecond, idfs)
#putting it in a dataframe
idf= pd.DataFrame([idfFirst, idfSecond])
```

In [314…

```
idf.transpose()
```

Out[314…

	0	1
15	0.001375	0.000000
forest	0.000000	0.000000
celebrate	0.000000	0.000820
citing	0.001375	0.000000
many	0.000000	0.000000
...	...	...
statewide	0.000000	0.000820
geotagged	0.001375	0.000000
carry	0.001375	0.000000
target	0.000000	0.003281
pune	0.000000	0.004101

321 rows × 2 columns