In [1]:

```
!pip install nltk -U
!pip install bs4 -U
```

```
Requirement already satisfied: nltk in c:\users\ganes\anaconda3\lib\site-packages
(3.6.1)
Collecting nltk
  Downloading nltk-3.7-py3-none-any.whl (1.5 MB)
Requirement already satisfied: tqdm in c:\users\ganes\anaconda3\lib\site-packages (f
rom nltk) (4.59.0)
Requirement already satisfied: joblib in c:\users\ganes\anaconda3\lib\site-packages
(from nltk) (1.0.1)
Collecting regex>=2021.8.3
  Downloading regex-2022.3.15-cp38-cp38-win_amd64.whl (274 kB)
Requirement already satisfied: click in c:\users\ganes\anaconda3\lib\site-packages
(from nltk) (7.1.2)
Installing collected packages: regex, nltk
  Attempting uninstall: regex
    Found existing installation: regex 2021.4.4
    Uninstalling regex-2021.4.4:
      Successfully uninstalled regex-2021.4.4
  Attempting uninstall: nltk
    Found existing installation: nltk 3.6.1
    Uninstalling nltk-3.6.1:
      Successfully uninstalled nltk-3.6.1
Successfully installed nltk-3.7 regex-2022.3.15
Collecting bs4
  Downloading bs4-0.0.1.tar.gz (1.1 kB)
Requirement already satisfied: beautifulsoup4 in c:\users\ganes\anaconda3\lib\site-p
ackages (from bs4) (4.9.3)
Requirement already satisfied: soupsieve>1.2 in c:\users\ganes\anaconda3\lib\site-pa
ckages (from beautifulsoup4->bs4) (2.2.1)
Building wheels for collected packages: bs4
  Building wheel for bs4 (setup.py): started
  Building wheel for bs4 (setup.py): finished with status 'done'
  Created wheel for bs4: filename=bs4-0.0.1-py3-none-any.whl size=1273 sha256=721f8e
06d273b0ecb1434819b8f36360384fc191f4103b160e7ea3cf97e0f1db
  Stored in directory: c:\users\ganes\appdata\local\pip\cache\wheels\75\78\21\68b124
549c9bdc94f822c02fb9aa3578a669843f9767776bca
Successfully built bs4
Installing collected packages: bs4
Successfully installed bs4-0.0.1
```

In [4]:

```
import nltk
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\ganes\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\ganes\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\ganes\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\ganes\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping taggers\averaged_perceptron_tagger.zip.
```

Out[4]:

```
True
```

In [5]:

```
import nltk
```

```
 'from',
 'an',
 'adjacent',
 'fort',
 'called',
 'Torna',
 'were',
 'used',
 'to',
 'completely',
 'build',
 'and',
 'fortify',
 'the',
 'Rajgad',
 'Fort.']
```

In [9]:

```
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
```

In [10]:

```
sent=sent_tokenize(para)
```

In [11]:

```
sent[2]
```

Out[11]:

```
'[1] Treasures discovered from an adjacent fort called Torna were used to completely
build and fortify the Rajgad Fort.'
```

In [12]:

```
words=word_tokenize(para)
```

In [13]:

```
words
```

Out[13]:

```
['Rajgad',
 '(',
 'literal',
 'meaning',
 'Ruling',
 'Fort',
 ')',
 'is',
 'a',
 'hill',
 'fort',
 'situated',
 'in',
 'the',
 'Pune',
 'district',
 'of',
 'Maharashtra',
 ',',
 'India',
 '.',
 'Formerly',
 'known',
 'as',
 'Murumdev',
 ',',
 'the',
 'fort',
 'was',
 'the',
 'capital',
 'of',
 'the',
 'Maratha',
 'Empire',
 'under',
 'the',
 'rule',
 'of',
 'Chatrapati',
 'Shivaji',
 'Maharaj',
 'for',
 'almost',
 '26',
 'years',
 ',',
 'after',
 'which',
 'the',
 'capital',
 'was',
 'moved',
 'to',
 'the',
 'Raigad',
 'Fort',
 '.',
 '[',
 '1',
 ']',
 'Treasures',
 'discovered',
 'from',
 'an',
 'adjacent',
 'fort',
 'called',
 'Torna',
 'were',
 'used',
 'to',
 'completely',
 'build',
 'and',
 'fortify',
 'the',
 'Rajgad',
 'Fort',
 '.']
```

In [14]:

```python
from nltk.corpus import stopwords
```

In [15]:

```python
swords=stopwords.words('english')
```

In [16]:

```python
swords
```

Out[16]:

```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you'
 '[',
 '1',
 ']',
 'Treasures',
 'discovered',
 'adjacent',
 'fort',
 'called',
 'Torna',
 'used',
 'completely',
 'build',
 'fortify',
 'Rajgad',
 'Fort',
 '.']
```

In [21]:

```
from nltk.stem import PorterStemmer
```

In [22]:

```
ps=PorterStemmer()
```

In [24]:

```
ps.stem('working')
```

Out[24]:

```
'work'
```

In [25]:

```
y=[ps.stem(word) for word in x]
```

In [26]:

```
y
```

Out[26]:

```
['rajgad',
 '(',
 'liter',
 'mean',
 'rule',
 'fort',
 ')',
 'hill',
 'fort',
 'situat',
 'pune',
 'district',
 'maharashtra',
 ',',
 'india',
 '.',
 'formerli',
 'known',
 'murumdev',
 ',',
 'fort',
 'capit',
 'maratha',
 'empir',
 'rule',
 'chatrapati',
 'shivaji',
 'maharaj',
 'almost',
 '26',
 'year',
 '.',
 'capit',
 'move',
 'raigad',
 'fort',
 '.',
 '[',
 '1',
 ']',
 'treasur',
 'discov',
 'adjac',
 'fort',
 'call',
 'torna',
 'use',
 'complet',
 'build',
 'fortifi',
 'rajgad',
 'fort',
 '.']
```

In [30]:

```
from nltk.stem import WordNetLemmatizer
```

In [31]:

```
wnl=WordNetLemmatizer()
```

In [32]:

```
wnl.lemmatize('working',pos='v')
#a-adjective
#n-noun
#r-adverb
```

```
---------------------------------------------------------------------------
LookupError                               Traceback (most recent call last)
~\anaconda3\lib\site-packages\nltk\corpus\util.py in __load(self)
     83                 try:
---> 84                     root = nltk.data.find(f"{self.subdir}/{zip_name}")
     85                 except LookupError:

~\anaconda3\lib\site-packages\nltk\data.py in find(resource_name, paths)
    582     resource_not_found = f"\n{sep}\n{msg}\n{sep}\n"
--> 583     raise LookupError(resource_not_found)
    584 

LookupError: 
**********************************************************************
  Resource omw-1.4 not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('omw-1.4')
  
  For more information see: https://www.nltk.org/data.html

  Attempted to load corpora/omw-1.4.zip/omw-1.4/

  Searched in:
    - 'C:\\Users\\ganes/nltk_data'
    - 'C:\\Users\\ganes\\anaconda3\\nltk_data'
    - 'C:\\Users\\ganes\\anaconda3\\share\\nltk_data'
    - 'C:\\Users\\ganes\\anaconda3\\lib\\nltk_data'
    - 'C:\\Users\\ganes\\AppData\\Roaming\\nltk_data'
    - 'C:\\nltk_data'
    - 'D:\\nltk_data'
    - 'E:\\nltk_data'
```

```
     83                  try:

~\anaconda3\lib\site-packages\nltk\data.py in find(resource_name, paths)
    581     sep = "*" * 70
    582     resource_not_found = f"\n{sep}\n{msg}\n{sep}\n"
--> 583     raise LookupError(resource_not_found)
    584 
    585 

LookupError: 
**********************************************************************
  Resource omw-1.4 not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('omw-1.4')
  
  For more information see: https://www.nltk.org/data.html

  Attempted to load corpora/omw-1.4

  Searched in:
    - 'C:\\Users\\ganes/nltk_data'
    - 'C:\\Users\\ganes\\anaconda3\\nltk_data'
    - 'C:\\Users\\ganes\\anaconda3\\share\\nltk_data'
    - 'C:\\Users\\ganes\\anaconda3\\lib\\nltk_data'
    - 'C:\\Users\\ganes\\AppData\\Roaming\\nltk_data'
    - 'C:\\nltk_data'
    - 'D:\\nltk_data'
    - 'E:\\nltk_data'
**********************************************************************
```

In [33]:

```
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\ganes\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\omw-1.4.zip.
```

Out[33]:

```
True
```

In [34]:

```
wnl.lemmatize('working',pos='v')
#a-adjective
#n-noun
#r-adverb
```

Out[34]:

```
'work'
```

In [35]:

```
print(ps.stem('went'))
print(wnl.lemmatize('went',pos='v'))
```

```
went
go
```

In [36]:

```
z=[wnl.lemmatize(word,pos='v') for word in x]
```

In [37]:

```
z
```

Out[37]:

```
['Rajgad',
 '(',
 'literal',
 'mean',
 'Ruling',
 'Fort',
 ')',
 'hill',
 'fort',
 'situate',
 'Pune',
 'district',
 'Maharashtra',
 ',',
 'India',
 '.',
 'Formerly',
 'know',
 'Murumdev',
 ',',
 'fort',
 'capital',
 'Maratha',
 'Empire',
 'rule',
 'Chatrapati',
 'Shivaji',
 'Maharaj',
 'almost',
 '26',
 'years',
 ',',
 'capital',
 'move',
 'Raigad',
 'Fort',
 '.',
 '[',
 '1',
 ']',
 'Treasures',
 'discover',
 'adjacent',
 'fort',
 'call',
 'Torna',
 'use',
 'completely',
 'build',
 'fortify',
 'Rajgad',
 'Fort',
 '.']
```

In [38]:

```
import string
```

In [39]:

```
string.punctuation
```

Out[39]:

```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

In [40]:

```
t=[word for word in words if word not in string.punctuation]
```

In [41]:

```
t
```

Out[41]:

```
['Rajgad',
 'literal',
 'meaning',
 'Ruling',
 ('and', 'CC'),
 ('fortify', 'VB'),
 ('the', 'DT'),
 ('Rajgad', 'NNP'),
 ('Fort', 'NNP')]
```

In [44]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [45]:

```
tfidf = TfidfVectorizer()
```

In [46]:

```
v=tfidf.fit_transform(t)
```

In [47]:

```
v.shape
```

Out[47]:

```
(70, 50)
```

In [48]:

```
import pandas as pd
pd.DataFrame(v)
```

Out[48]:

	0
0	(0, 35)\t1.0
1	(0, 25)\t1.0
2	(0, 29)\t1.0
3	(0, 37)\t1.0
4	(0, 17)\t1.0
...	...
65	(0, 5)\t1.0
66	(0, 18)\t1.0
67	(0, 40)\t1.0
68	(0, 35)\t1.0
69	(0, 17)\t1.0

70 rows × 1 columns

In [ ]: