In [ ]:
```
#8. Text Analytics
#1.Extract Sample document and apply following document preprocessing methods:
#Stemming andLemmatization.
#2.Create representation of document by calculating Term Frequency and Inverse

#no dataset
```

In [39]:
```
import nltk
```

In [40]:
```
nltk.download('stopwords')
nltk.download('words')
nltk.download('wordnet')
nltk.download('averged_perception_tagger')
nltk.download('punkt')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\arifa\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package words to
[nltk_data]     C:\Users\arifa\AppData\Roaming\nltk_data...
[nltk_data]   Package words is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\arifa\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Error loading averged_perception_tagger: Package
[nltk_data]     'averged_perception_tagger' not found in index
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\arifa\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[40]:
```
True
```

In [41]:
```
import pandas as pd
import numpy as np
```

In [42]:
```
sent= "They told that thier eges are 20
23 and 27 respectively"
```

In [43]:
```
add=[]
```

In [44]:
```
for word in sent.split():
    if word.isdigit():
        add.append(int(word))
```

In [45]:
```
print ("Ave", sum(add)/len(add))
```

```
Ave 23.333333333333332
```

In [46]:
```
from nltk.tokenize import word_tokenize, sent_tokenize
```

In [47]:
```
sent= "Hello all! how are you? Welcome to pun "
```

In [48]:
```
sent_tokenize(sent)
```

Out[48]:
```
['Hello all!', 'how are you?', 'Welcome to pun']
```

In [49]:
```
word_tokenize(sent)
```

Out[49]:
```
['Hello', 'all', '!', 'how', 'are', 'you', '?', 'Welcome', 'to', 'pun']
```

In [50]:
```
from nltk.tokenize import SpaceTokenizer
tk=SpaceTokenizer()
tk.tokenize(sent)
```

Out[50]:
```
['Hello', 'all!', 'how', 'are', 'you?', 'Welcome', 'to', 'pun', '']
```

In [51]:
```
sent='Hello all!\tHow are u?\tto pune'
```

In [52]:
```
print(sent)
```

```
Hello all!      How are u?      to pune
```

In [53]:
```
s1='ctas','catlike','catty','cat'
s2='stemmer','stemming','stemmed','stem'
s3='fishing','fished','fisher','fish'
s4='argue','argued','argues','argus'
```

In [54]:
```
from nltk.stem import PorterStemmer
```

In [55]:
```
ps=PorterStemmer()
```

In [56]:
```
ps.stem(s3[0])
```

Out[56]:
```
'fish'
```

In [57]:
```
ps=PorterStemmer()
print(ps.stem(word))
```

```
respect
```

In [58]:
```
# Lemmatization
```

In [59]:
```
word='playing'
```

In [60]:
```
from nltk.stem import WordNetLemmatizer
```

In [61]:
```
wnl=WordNetLemmatizer()
```

In [62]:
```
print(wnl.lemmatize(word,'n')) # noun
print(wnl.lemmatize(word,'v')) # verb
print(wnl.lemmatize(word,'a')) # adjective
print(wnl.lemmatize(word,'r')) # adverb
```

```
playing
play
playing
playing
```

In [63]:
```
word='went'
```

In [64]:
```
wnl=WordNetLemmatizer()
print(wnl.lemmatize(word,'n')) # noun
print(wnl.lemmatize(word,'v')) # verb
print(wnl.lemmatize(word,'a')) # adjective
print(wnl.lemmatize(word,'r')) # adverb
```

```
went
go
went
went
```

In [65]:
```
# POS tagging
```

In [66]:
```
from nltk import pos_tag
```

In [67]:
```
import nltk
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\arifa\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

Out[67]:
```
True
```

In [68]:

```
sents='Rajgad (literal meaning Ruling Fort) is a hill fort
situated in the Pune district of Maharashtra, India. Formerly
known as Murumde'
```

In [69]:

```
print(sents)
```

```
Rajgad (literal meaning Ruling Fort) is a hill fort situated in the Pune dis
trict of Maharashtra, India. Formerly known as Murumde
```

In [70]:

```
words=word_tokenize(sents)
```

In [71]:

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

```
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\arifa\AppData\Roaming\nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

Out[71]:

```
True
```

In [72]:

```
pos_tag(words)
```

Out[72]:

```
[('Rajgad', 'NNP'),
 ('(', '('),
 ('literal', 'JJ'),
 ('meaning', 'NN'),
 ('Ruling', 'NNP'),
 ('Fort', 'NNP'),
 (')', ')'),
 ('is', 'VBZ'),
 ('a', 'DT'),
 ('hill', 'NN'),
 ('fort', 'NN'),
 ('situated', 'VBN'),
 ('in', 'IN'),
 ('the', 'DT'),
 ('Pune', 'NNP'),
 ('district', 'NN'),
 ('of', 'IN'),
 ('Maharashtra', 'NNP'),
 (',', ','),
 ('India', 'NNP'),
 ('.', '.'),
 ('Formerly', 'RB'),
 ('known', 'VBN'),
 ('as', 'IN'),
 ('Murumde', 'NNP')]
```

In [73]:

```
tags=pos_tag(words)
```

In [74]:
```
for word in tags:
    if word[1].startswith('V'):
        print(word[0])
```

```
is
situated
known
```

In [75]:
```
# spell correction
```

In [76]:
```
# spell correction
from textblob import TextBlob
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_1432\601046462.py in <module>
      1 # spell correction
----> 2 from textblob import TextBlob

ModuleNotFoundError: No module named 'textblob'
```

In [77]:
```
t=TextBlob('computoor')
print(t.correct())
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_1432\2196745319.py in <module>
----> 1 t=TextBlob('computoor')
      2 print(t.correct())

NameError: name 'TextBlob' is not defined
```

In [78]:
```
t=TextBlob('nead')
print(t.correct())
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_1432\3224985225.py in <module>
----> 1 t=TextBlob('nead')
      2 print(t.correct())

NameError: name 'TextBlob' is not defined
```

In [ ]: