In [1]:
```
#### Simple program to paragraph into vector using BOW(Bag of words and TFIDF)
import nltk
```

In [2]:
```
paragraph ="""Shivaji I (Shivaji Shahaji Bhonsale, Marathi pronunciation: [ʃiˈʊaːdʑiː ˈbʱos(ə)le]; c. 19 Februa

Over the course of his life, Shivaji engaged in both alliances and hostilities with the Mughal Empire, the Sulta
"""
```

In [3]:
```
paragraph
```

Out[3]:
```
"Shivaji I (Shivaji Shahaji Bhonsale, Marathi pronunciation: [ʃiˈʊaːdʑiː ˈbʱos(ə)le]; c.\u200919 February 1630
– 3 April 1680)[5] was an Indian ruler and a member of the Bhonsle dynasty.[6] Shivaji carved out his own indep
endent kingdom from the Sultanate of Bijapur that formed the genesis of the Maratha Confederacy. In 1674, he wa
s formally crowned the Chhatrapati of his realm at Raigad Fort.[7]\n\nOver the course of his life, Shivaji enga
ged in both alliances and hostilities with the Mughal Empire, the Sultanate of Golconda, the Sultanate of Bijap
ur and the European colonial powers. Following the Battle of Purandar, Shivaji entered into vassalage with the
Mughal empire, assuming the role of a Mughal chief and undertaking military expeditions on behalf of the empire
for a brief duration.[8] Shivaji's military forces expanded the Maratha sphere of influence, capturing and buil
ding forts, and forming a Maratha navy.\n"
```

In [6]:
```
#### convert the paragrapg in to sentance
sentance = nltk.sent_tokenize(paragraph)
```

In [7]:
```
sentance
```

Out[7]:
```
['Shivaji I (Shivaji Shahaji Bhonsale, Marathi pronunciation: [ʃiˈʊaːdʑiː ˈbʱos(ə)le]; c.\u200919 February 1630
– 3 April 1680)[5] was an Indian ruler and a member of the Bhonsle dynasty.',
 '[6] Shivaji carved out his own independent kingdom from the Sultanate of Bijapur that formed the genesis of t
he Maratha Confederacy.',
 'In 1674, he was formally crowned the Chhatrapati of his realm at Raigad Fort.',
 '[7]\n\nOver the course of his life, Shivaji engaged in both alliances and hostilities with the Mughal Empire,
the Sultanate of Golconda, the Sultanate of Bijapur and the European colonial powers.',
 'Following the Battle of Purandar, Shivaji entered into vassalage with the Mughal empire, assuming the role of
a Mughal chief and undertaking military expeditions on behalf of the empire for a brief duration.',
 "[8] Shivaji's military forces expanded the Maratha sphere of influence, capturing and building forts, and for
ming a Maratha navy."]
```

In [11]:
```
#### then we need to remove the unwanted words into the sentance
import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
```

In [14]:
```
corpus=[]
for i in range(len(sentance)):
    normal = re.sub('[^a-zA-Z]' ,' ' ,sentance[i])
    normal=normal.lower()
    corpus.append(normal)
```

In [15]:
```
corpus
```

Out[15]:
```
['shivaji i  shivaji shahaji bhonsale  marathi pronunciation    i  a d i   b os   le   c     february
april          was an indian ruler and a member of the bhonsle dynasty ',
 '    shivaji carved out his own independent kingdom from the sultanate of bijapur that formed the genesis of t
he maratha confederacy ',
 'in       he was formally crowned the chhatrapati of his realm at raigad fort ',
 '     over the course of his life  shivaji engaged in both alliances and hostilities with the mughal empire  t
he sultanate of golconda  the sultanate of bijapur and the european colonial powers ',
 'following the battle of purandar  shivaji entered into vassalage with the mughal empire  assuming the role of
a mughal chief and undertaking military expeditions on behalf of the empire for a brief duration ',
 '    shivaji s military forces expanded the maratha sphere of influence  capturing and building forts  and for
ming a maratha navy ']
```

In [16]:
```
stemmmer = PorterStemmer()
```

In [17]:
```
### then we apply stemming and convert sentance into words
for i in corpus:
    words = nltk.word_tokenize(i)
    for word in words:
        if word not in set(stopwords.words('english')):
            print(stemmmer.stem(word))
```

```
shivaji
shivaji
shahaji
bhonsal
marathi
pronunci
b
os
le
c
februari
april
indian
ruler
member
bhonsl
dynasti
shivaji
carv
independ
kingdom
sultan
bijapur
form
genesi
maratha
confederaci
formal
crown
chhatrapati
realm
raigad
fort
cours
life
shivaji
engag
allianc
hostil
mughal
empir
sultan
golconda
sultan
bijapur
european
coloni
power
follow
battl
purandar
shivaji
enter
vassalag
mughal
empir
assum
role
mughal
chief
undertak
militari
expedit
behalf
empir
brief
durat
shivaji
militari
forc
expand
maratha
sphere
influenc
captur
build
fort
form
maratha
navi
```

In [25]:
```
### then we need to convert words into the vector
### BOW(Bag Of Words)
```

```
from sklearn.feature_extraction.text import CountVectorizer
model = CountVectorizer(binary=True)
```

In [26]:
```
X = model.fit_transform(corpus)
```

In [27]:
```
X
```

Out[27]:
```
<6x82 sparse matrix of type '<class 'numpy.int64'>'
        with 110 stored elements in Compressed Sparse Row format>
```

In [29]:
```
model.vocabulary_
```

Out[29]:

```
{'shivaji': 73,
 'shahaji': 72,
 'bhonsale': 8,
 'marathi': 54,
 'pronunciation': 66,
 'os': 61,
 'le': 51,
 'february': 30,
 'april': 3,
 'was': 80,
 'an': 1,
 'indian': 47,
 'ruler': 71,
 'and': 2,
 'member': 55,
 'of': 59,
 'the': 77,
 'bhonsle': 9,
 'dynasty': 23,
 'carved': 15,
 'out': 62,
 'his': 43,
 'own': 64,
 'independent': 46,
 'kingdom': 50,
 'from': 39,
 'sultanate': 75,
 'bijapur': 10,
 'that': 76,
 'formed': 35,
 'genesis': 40,
 'maratha': 53,
 'confederacy': 19,
 'in': 45,
 'he': 42,
 'formally': 34,
 'crowned': 21,
 'chhatrapati': 16,
 'realm': 69,
 'at': 5,
 'raigad': 68,
 'fort': 37,
 'over': 63,
 'course': 20,
 'life': 52,
 'engaged': 25,
 'both': 11,
 'alliances': 0,
 'hostilities': 44,
 'with': 81,
 'mughal': 57,
 'empire': 24,
 'golconda': 41,
 'european': 27,
 'colonial': 18,
 'powers': 65,
 'following': 31,
 'battle': 6,
 'purandar': 67,
 'entered': 26,
 'into': 49,
 'vassalage': 79,
 'assuming': 4,
 'role': 70,
 'chief': 17,
 'undertaking': 78,
 'military': 56,
 'expeditions': 29,
 'on': 60,
 'behalf': 7,
 'for': 32,
 'brief': 12,
 'duration': 22,
 'forces': 33,
 'expanded': 28,
 'sphere': 74,
 'influence': 48,
 'capturing': 14,
 'building': 13,
 'forts': 38,
 'forming': 36,
 'navy': 58}
```

In [30]:
```
corpus[0]
```

Out[30]:
```
'shivaji i  shivaji shahaji bhonsale  marathi pronunciation    i  a d i   b os   le   c     february          a
pril          was an indian ruler and a member of the bhonsle dynasty '
```

In [31]:
```
X[0].toarray()
```

Out[31]:
```
array([[0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,
        1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0]], dtype=int64)
```

In [32]:
```
#### then we apply TFIDF
from sklearn.feature_extraction.text import TfidfVectorizer
model1 = TfidfVectorizer()
```

In [33]:
```
X1= model.fit_transform(corpus)
```

In [34]:
```
X1[0].toarray()
```

Out[34]:
```
array([[0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,
        1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0]], dtype=int64)
```

In [35]:
```
X1[4].toarray()
```

Out[35]:
```
array([[0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0,
        1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0,
        0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1]], dtype=int64)
```

In [ ]:
```

```