In [1]:

```
import nltk
```

In [2]:

```
paragraph = """I have three visions for India. In 3000 years of our history people from
all over
the world have come and invaded us, captured our lands, conquered our minds. From Alexa
nder
onwards the Greeks, the Turks, the Moguls, the Portuguese, the British, the French, the
Dutch,
all of them came and looted us, took over what was ours. Yet we have not done this to a
ny other
nation. We have not conquered anyone. We have not grabbed their land, their culture and
their
history and tried to enforce our way of life on them. Why? Because we respect the freed
om of
others. That is why my FIRST VISION is that of FREEDOM. I believe that India got its fi
rst vision
of this in 1857, when we started the war of Independence. It is this freedom that we mu
st protect
and nurture and build on. If we are not free, no one will respect us.
We have 10 percent growth rate in most areas. Our poverty levels are falling. Our achie
vements
are being globally recognised today. Yet we lack the self-confidence to see ourselves a
s a
developed nation, self-reliant and self-assured. Isn’t this incorrect? MY SECOND VISION
for India
is DEVELOPMENT. For fifty years we have been a developing nation. It is time we see our
selves as
a developed nation. We are among top five nations in the world in terms of GDP.
I have a THIRD VISION. India must stand up to the world. Because I believe that unless
 India
stands up to the world, no one will respect us. Only strength respects strength. We mus
t be strong
not only as a military power but also as an economic power. Both must go hand-in-hand.
 My good
fortune was to have worked with three great minds. Dr.Vikram Sarabhai, of the Dept. of
 Space,
Professor Satish Dhawan, who succeeded him and Dr. Brahm Prakash, father of nuclear mat
erial.
I was lucky to have worked with all three of them closely and consider this the great o
pportunity
of my life."""
```

In [3]:

```
# First step is tokenization - either sentence or word tokenize.
word = nltk.word_tokenize(paragraph)
```

In [4]:

```
word[0:10]
```

Out[4]:

```
['I', 'have', 'three', 'visions', 'for', 'India', '.', 'In', '3000', 'year
s']
```

In [5]:

```
sentence = nltk.sent_tokenize(paragraph)
len(sentence)
```

Out[5]:

```
31
```

In [6]:

```
sentence[:5]
```

Out[6]:

```
['I have three visions for India.',
 'In 3000 years of our history people from all over\nthe world have come a
nd invaded us, captured our lands, conquered our minds.',
 'From Alexander\nonwards the Greeks, the Turks, the Moguls, the Portugues
e, the British, the French, the Dutch, \nall of them came and looted us, t
ook over what was ours.',
 'Yet we have not done this to any other\nnation.',
 'We have not conquered anyone.']
```

In [7]:

```
# Stemming : converting into base or root words
from nltk.stem import PorterStemmer
from nltk.corpus import stopwords

stopwords = stopwords.words('english')
print(stopwords)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "yo
u're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselv
es', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'hersel
f', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'the
mselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'thes
e', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'hav
e', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'th
e', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'a
t', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through',
'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down',
'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'on
ce', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both',
'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'no
t', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can',
'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll',
'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't",
'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't",
'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn',
"mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't",
'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "would
n't"]
```

In [8]:

```
stemmer = PorterStemmer()
for i in range(len(sentence)):
    word = nltk.word_tokenize(sentence[i])
    word = [stemmer.stem(words) for words in word if words not in set(stopwords)]
    sentence[i] = " ".join(word)
```

In [9]:

```
sentence[:5]
```

Out[9]:

```
['I three vision india .',
 'In 3000 year histori peopl world come invad us , captur land , conquer m
ind .',
 'from alexand onward greek , turk , mogul , portugues , british , french
, dutch , came loot us , took .',
 'yet done nation .',
 'We conquer anyon .']
```

We can see that stemming has problems such history converts to histori, capture to captur, so inorder to overcome this we use lemmatization

In [10]:

```
# Lemmatization
import nltk
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
```

In [11]:

```
paragraph = """I have three visions for India. In 3000 years of our history people from
all over
the world have come and invaded us, captured our lands, conquered our minds. From Alexa
nder
onwards the Greeks, the Turks, the Moguls, the Portuguese, the British, the French, the
Dutch,
all of them came and looted us, took over what was ours. Yet we have not done this to a
ny other
nation. We have not conquered anyone. We have not grabbed their land, their culture and
their
history and tried to enforce our way of life on them. Why? Because we respect the freed
om of
others. That is why my FIRST VISION is that of FREEDOM. I believe that India got its fi
rst vision
of this in 1857, when we started the war of Independence. It is this freedom that we mu
st protect
and nurture and build on. If we are not free, no one will respect us.
We have 10 percent growth rate in most areas. Our poverty levels are falling. Our achie
vements
are being globally recognised today. Yet we lack the self-confidence to see ourselves a
s a
developed nation, self-reliant and self-assured. Isn’t this incorrect? MY SECOND VISION
for India
is DEVELOPMENT. For fifty years we have been a developing nation. It is time we see our
selves as
a developed nation. We are among top five nations in the world in terms of GDP.
I have a THIRD VISION. India must stand up to the world. Because I believe that unless
 India
stands up to the world, no one will respect us. Only strength respects strength. We mus
t be strong
not only as a military power but also as an economic power. Both must go hand-in-hand.
 My good
fortune was to have worked with three great minds. Dr.Vikram Sarabhai, of the Dept. of
 Space,
Professor Satish Dhawan, who succeeded him and Dr. Brahm Prakash, father of nuclear mat
erial.
I was lucky to have worked with all three of them closely and consider this the great o
pportunity
of my life."""
```

In [12]:

```
sentences = nltk.sent_tokenize(paragraph)
lemmatizer = WordNetLemmatizer()
```

In [13]:

```
lemm_sent = []
for i in range(len(sentences)):
    words = nltk.word_tokenize(sentences[i])
    words = [lemmatizer.lemmatize(word) for word in words if word not in set(stopwords.
words("english"))]
    sentences[i] = " ".join(words)
    lemm_sent.append(sentences[i])
```

In [14]:

```
lemm_sent[:5]
```

Out[14]:

```
['I three vision India .',
 'In 3000 year history people world come invaded u , captured land , conqu
ered mind .',
 'From Alexander onwards Greeks , Turks , Moguls , Portuguese , British ,
French , Dutch , came looted u , took .',
 'Yet done nation .',
 'We conquered anyone .']
```

In [15]:

```
import nltk
```

In [16]:

```
paragraph = """I have three visions for India. In 3000 years of our history people from
all over
the world have come and invaded us, captured our lands, conquered our minds. From Alexa
nder
onwards the Greeks, the Turks, the Moguls, the Portuguese, the British, the French, the
Dutch,
all of them came and looted us, took over what was ours. Yet we have not done this to a
ny other
nation. We have not conquered anyone. We have not grabbed their land, their culture and
their
history and tried to enforce our way of life on them. Why? Because we respect the freed
om of
others. That is why my FIRST VISION is that of FREEDOM. I believe that India got its fi
rst vision
of this in 1857, when we started the war of Independence. It is this freedom that we mu
st protect
and nurture and build on. If we are not free, no one will respect us.
We have 10 percent growth rate in most areas. Our poverty levels are falling. Our achie
vements
are being globally recognised today. Yet we lack the self-confidence to see ourselves a
s a
developed nation, self-reliant and self-assured. Isn’t this incorrect? MY SECOND VISION
for India
is DEVELOPMENT. For fifty years we have been a developing nation. It is time we see our
selves as
a developed nation. We are among top five nations in the world in terms of GDP.
I have a THIRD VISION. India must stand up to the world. Because I believe that unless
 India
stands up to the world, no one will respect us. Only strength respects strength. We mus
t be strong
not only as a military power but also as an economic power. Both must go hand-in-hand.
 My good
fortune was to have worked with three great minds. Dr.Vikram Sarabhai, of the Dept. of
 Space,
Professor Satish Dhawan, who succeeded him and Dr. Brahm Prakash, father of nuclear mat
erial.
I was lucky to have worked with all three of them closely and consider this the great o
pportunity
of my life."""
```

In [17]:

```
# cleaning the text
import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
```

In [18]:

```
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()
sentences = nltk.sent_tokenize(paragraph)
corpus = []
for i in range(len(sentences)):
    review = re.sub("[^a-zA-z]",' ',sentences[i])
    review = review.lower()
    review = review.split()
    review = [lemmatizer.lemmatize(word) for word in review if word not in set(stopword
s.words("english"))]
    review = " ".join(review)
    corpus.append(review)
```

In [19]:

```
corpus[:10]
```

Out[19]:

```
['three vision india',
 'year history people world come invaded u captured land conquered mind',
 'alexander onwards greek turk mogul portuguese british french dutch came
looted u took',
 'yet done nation',
 'conquered anyone',
 'grabbed land culture history tried enforce way life',
 '',
 'respect freedom others',
 'first vision freedom',
 'believe india got first vision started war independence']
```

In [20]:

```
sentences[:5]
```

Out[20]:

```
['I have three visions for India.',
 'In 3000 years of our history people from all over\nthe world have come a
nd invaded us, captured our lands, conquered our minds.',
 'From Alexander\nonwards the Greeks, the Turks, the Moguls, the Portugues
e, the British, the French, the Dutch, \nall of them came and looted us, t
ook over what was ours.',
 'Yet we have not done this to any other\nnation.',
 'We have not conquered anyone.']
```

In [23]:

```
# Creating bag of word model
from sklearn.feature_extraction.text import CountVectorizer # for BOW
cv = CountVectorizer()
x = cv.fit_transform(corpus).toarray()
```

In [25]:

```
x, x.shape
```

Out[25]:

```
(array([[0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 1, 1, 0],
        [0, 1, 0, ..., 0, 0, 0],
        ...,
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0],
        [0, 0, 0, ..., 0, 0, 0]], dtype=int64), (31, 112))
```

In [ ]:

```
## Tf-idf Model ---> to overcome the disadvantages of BOW we use tfidf, BoW return spar
se matrix
# BoW disadvantages:
# 1. All words have the same importance
# 2. No sematic information Preserved
# tfidf:
# 1. Semantic information is preserved as uncommon words are given more importance than
the common
# words
# tf = term frequency
# idf = inverse documnet frequency
# tf-idf = tf*idf
# tf = (no of occurances of a word in the document) / (total number of words)
# idf = log[(no of document or sentences) / (no of documnet containing the word)]
```

In [26]:

```
import nltk
import re
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [27]:

```
paragraph = """I have three visions for India. In 3000 years of our history people from all over
the world have come and invaded us, captured our lands, conquered our minds. From Alexander
onwards the Greeks, the Turks, the Moguls, the Portuguese, the British, the French, the Dutch,
all of them came and looted us, took over what was ours. Yet we have not done this to any other
nation. We have not conquered anyone. We have not grabbed their land, their culture and their
history and tried to enforce our way of life on them. Why? Because we respect the freedom of
others. That is why my FIRST VISION is that of FREEDOM. I believe that India got its first vision
of this in 1857, when we started the war of Independence. It is this freedom that we must protect
and nurture and build on. If we are not free, no one will respect us.
We have 10 percent growth rate in most areas. Our poverty levels are falling. Our achievements
are being globally recognised today. Yet we lack the self-confidence to see ourselves as a
developed nation, self-reliant and self-assured. Isn’t this incorrect? MY SECOND VISION for India
is DEVELOPMENT. For fifty years we have been a developing nation. It is time we see ourselves as
a developed nation. We are among top five nations in the world in terms of GDP.
I have a THIRD VISION. India must stand up to the world. Because I believe that unless India
stands up to the world, no one will respect us. Only strength respects strength. We must be strong
not only as a military power but also as an economic power. Both must go hand-in-hand. My good
fortune was to have worked with three great minds. Dr.Vikram Sarabhai, of the Dept. of Space,
Professor Satish Dhawan, who succeeded him and Dr. Brahm Prakash, father of nuclear material.
I was lucky to have worked with all three of them closely and consider this the great opportunity
of my life."""
```

In [28]:

```
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()
sentences = nltk.sent_tokenize(paragraph)
corpus = []
for i in range(len(sentences)):
    review = re.sub("[^a-zA-Z]",' ',sentences[i])
    review = review.lower()
    review = nltk.word_tokenize(review)
    review = [lemmatizer.lemmatize(word) for word in review if word not in set(stopwords.words("english"))]
    review = " ".join(review)
    corpus.append(review)
```

In [29]:

```
corpus[:5]
```

Out[29]:

```
['three vision india',
 'year history people world come invaded u captured land conquered mind',
 'alexander onwards greek turk mogul portuguese british french dutch came
looted u took',
 'yet done nation',
 'conquered anyone']
```

In [32]:

```
tfidf = TfidfVectorizer()
y = tfidf.fit_transform(corpus).toarray()
```

In [33]:

```
y
```

Out[33]:

```
array([[0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.25883507, 0.30512561,
        0.        ],
       [0.        , 0.28867513, 0.        , ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ]])
```