Bag of words

The Bag of Words (BoW) model is a simple and widely used technique in Natural Language Processing (NLP) for text representation. It transforms text into a numerical feature vector, which can then be used for tasks like machine learning, text classification, and clustering.

TF-IDF

TF-IDF stands for Term Frequency-Inverse Document Frequency, and it is a statistical technique used in Natural Language Processing (NLP) to evaluate the importance of a word in a document relative to a collection of documents (called the corpus).

In [ ]:

In [35]:
```
#import all libraries
```

In [103…
```
import nltk
import re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.stem.porter import PorterStemmer
```

In [124…
```
ps = PorterStemmer()
wordnet = WordNetLemmatizer()
```

In [105…
```
nltk.download('stopwords')
nltk.download('wordnet')
```

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

Out[105… True

In [125…
```
stop_words = set(stopwords.words('english'))
```

In [107…
```
#improt paragraph data
```

In [51]:
```
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,
               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.
               My second vision for India’s development. For fifty years we have been a developing nation.
               It is time we see ourselves as a developed nation. We are among the top 5 nations of the world
               in terms of GDP. We have a 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?
               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 mater
               I was lucky to have worked with all three of them closely and consider this the great opportunit
               I see four milestones in my career"""
```

In [52]:
```
paragraph
```

Out[52]:
```
'I have three visions for India. In 3000 years of our history, people from all over \n              the world
have come and invaded us, captured our lands, conquered our minds. \n              From Alexander onwards, the
Greeks, the Turks, the Moguls, the Portuguese, the British,\n              the French, the Dutch, all of them
came and looted us, took over what was ours. \n              Yet we have not done this to any other nation. We
have not conquered anyone. \n              We have not grabbed their land, their culture, \n              the
ir history and tried to enforce our way of life on them. \n              Why? Because we respect the freedom o
f others.That is why my \n              first vision is that of freedom. I believe that India got its first vi
sion of \n              this in 1857, when we started the War of Independence. It is this freedom that\n
we must protect and nurture and build on. If we are not free, no one will respect us.\n              My second
vision for India's development. For fifty years we have been a developing nation.\n              It is time we
see ourselves as a developed nation. We are among the top 5 nations of the world\n              in terms of GD
P. We have a 10 percent growth rate in most areas. Our poverty levels are falling.\n              Our achievem
ents are being globally recognised today. Yet we lack the self-confidence to\n              see ourselves as a
developed nation, self-reliant and self-assured. Isn't this incorrect?\n              I have a third vision. I
ndia must stand up to the world. Because I believe that unless India \n              stands up to the world, n
o one will respect us. Only strength respects strength. We must be \n              strong not only as a milita
ry power but also as an economic power. Both must go hand-in-hand. \n              My good fortune was to have
worked with three great minds. Dr. Vikram Sarabhai of the Dept. of \n              space, Professor Satish Dha
wan, who succeeded him and Dr. Brahm Prakash, father of nuclear material.\n              I was lucky to have w
orked with all three of them closely and consider this the great opportunity of my life. \n              I see
four milestones in my career'
```

In [ ]:

In [53]:
```
#create sentence tokenizatio
```

In [55]:
```
sentence_token = nltk.sent_tokenize(paragraph)
```

In [56]:
```
sentence_token
```

Out[56]:
```
['I have three visions for India.',
 'In 3000 years of our history, people from all over \n              the world have come and invaded us, captu
red our lands, conquered our minds.',
 'From Alexander onwards, the Greeks, the Turks, the Moguls, the Portuguese, the British,\n              the F
rench, 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, \n              their history and tried to enforce our way of
life on them.',
 'Why?',
 'Because we respect the freedom of others.That is why my \n              first vision is that of freedom.',
 'I believe that India got its first vision of \n              this in 1857, when we started the War of Indepe
ndence.',
 'It is this freedom that\n              we must protect and nurture and build on.',
 'If we are not free, no one will respect us.',
 'My second vision for India's development.',
 'For fifty years we have been a developing nation.',
 'It is time we see ourselves as a developed nation.',
 'We are among the top 5 nations of the world\n              in terms of GDP.',
 'We have a 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\n              see ourselves as a developed nation, self-reliant and self
-assured.',
 'Isn't this incorrect?',
 'I have a third vision.',
 'India must stand up to the world.',
 'Because I believe that unless India \n              stands up to the world, no one will respect us.',
 'Only strength respects strength.',
 'We must be \n              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 \n              space, Professor Satish Dhawan, who succeeded him and Dr. Brahm Prakash, father of nuclea
r material.',
 'I was lucky to have worked with all three of them closely and consider this the great opportunity of my life.
',
 'I see four milestones in my career']
```

In [58]:
```
len(sentence_token)
```

Out[58]:
```
31
```

In [59]:
```
#there are total 31 sentences
```

In [ ]:

In [61]:
```
corpus = []
```

In [62]:
```
#p aragraph data cleaning
```

In [126…
```
for i in range(len(sentence_token)):
    review = re.sub('[^a-zA-Z]', ' ',sentence_token[i])
    review = review.lower()
    review = review.split()
    review = [wordnet.lemmatize(word) for word in review if word not in stop_words]
    review = ' '.join(review)
    corpus.append(review)
```

In [128…
```
corpus
```

Out[128…
```
['i have three visions for india',
 'in 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 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 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',
 'my second vision for india s development',
 'for fifty years we have been a developing nation',
 'it is time we see ourselves as a developed nation',
 'we are among the top nations of the world in terms of gdp',
 'we have a 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',
 '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',
 'i see four milestones in my career',
 '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',
 'freedom must protect nurture build',
 'free one respect u',
 'second vision india development',
 'fifty year developing nation',
 'time see developed nation',
 'among top nation world term gdp',
 'percent growth rate area',
 'poverty level falling',
 'achievement globally recognised today',
 'yet lack self confidence see developed nation self reliant self assured',
 'incorrect',
 'third vision',
 'india must stand world',
 'believe unless india stand world one respect u',
 'strength respect strength',
 'must strong military power also economic power',
 'must go hand hand',
 'good fortune worked three great mind',
 'dr vikram sarabhai dept',
 'space professor satish dhawan succeeded dr brahm prakash father nuclear material',
 'lucky worked three closely consider great opportunity life',
 'see four milestone career',
 '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',
'freedom must protect nurture build',
'free one respect u',
'second vision india development',
'fifty year developing nation',
'time see developed nation',
'among top nation world term gdp',
'percent growth rate area',
'poverty level falling',
'achievement globally recognised today',
'yet lack self confidence see developed nation self reliant self assured',
'incorrect',
'third vision',
'india must stand world',
'believe unless india stand world one respect u',
'strength respect strength',
'must strong military power also economic power',
'must go hand hand',
'good fortune worked three great mind',
'dr vikram sarabhai dept',
'space professor satish dhawan succeeded dr brahm prakash father nuclear material',
'lucky worked three closely consider great opportunity life',
'see four milestone career',
'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',
'freedom must protect nurture build',
'free one respect u',
'second vision india development',
'fifty year developing nation',
'time see developed nation',
'among top nation world term gdp',
'percent growth rate area',
'poverty level falling',
'achievement globally recognised today',
'yet lack self confidence see developed nation self reliant self assured',
'incorrect',
'third vision',
'india must stand world',
'believe unless india stand world one respect u',
'strength respect strength',
'must strong military power also economic power',
'must go hand hand',
'good fortune worked three great mind',
'dr vikram sarabhai dept',
'space professor satish dhawan succeeded dr brahm prakash father nuclear material',
'lucky worked three closely consider great opportunity life',
'see four milestone career']
```

In [139…
```
#create bag of word model
from sklearn.feature_extraction.text import CountVectorizer
cv1 = CountVectorizer()
x1 = cv1.fit_transform(corpus).toarray()
```

In [140…
```
x1
```

Out[140…
```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 1, 0],
       [0, 0, 1, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

In [ ]:

In [134…
```
#create TF-IDF model
```

In [137…
```
from sklearn.feature_extraction.text import TfidfVectorizer
cv = TfidfVectorizer()
x = cv.fit_transform(corpus).toarray()
```

In [138…
```
x
```

Out[138…

```
array([[0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.21472037,
        0.        ],
       [0.        , 0.        , 0.13230068, ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ]])
```

In [ ]: