NLP Assessment

Q1). Copy the next paragraph and answer the questions that follow:

For decades the All-India Congress under the leadership of Mohandas K. Gandhi struggled to rally the millions of British-ruled peoples in the Indian subcontinent. Like similar movements in other countries, it early felt the need for a distinctive symbol that could represent its nationalist objectives. In 1921 a university lecturer named Pingali (or Pinglay) Venkayya presented a flag design to Gandhi that consisted of the colours associated with the two principal religions, red for the Hindus and green for the Muslims. To the centre of the horizontally divided flag, Lala Hans Raj Sondhi suggested the addition of the traditional spinning wheel, which was associated with Gandhi's crusade to make Indians self-reliant by fabricating their own clothing from local fibres.
Gandhi modified the flag by adding a white stripe in the centre for the other religious communities in India, thus also providing a clearly visible background for the spinning wheel. In May 1923 at Nagpur, during peaceful protests against British rule, the flag was carried by thousands of people, hundreds of whom were arrested. The Congress flag came to be associated with nationhood for India, and it was officially recognized at the annual meeting of the party in August 1931. At the same time, the current arrangement of stripes and the use of deep saffron instead of red were approved. To avoid the sectarian associations of the original proposal, new attributions were associated with the saffron, white, and green stripes. They were said to stand for, respectively, courage and sacrifice, peace and truth, and faith and chivalry. During World War II Subhas Chandra Bose used this flag (without the spinning wheel) in territories his Japanese-aided army had captured.

a). Word_tokenise and sent_tokenise

In [4]:
```
import nltk
nltk.download('punkt')
```

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

Out[4]:
```
True
```

In [8]:
```
from nltk.tokenize import word_tokenize

corpus="""For decades the All-India Congress under the leadership of Mohandas K. Gandhi struggled to rally the millions of British-
```

```
Gandhi modified the flag by adding a white stripe in the centre for the other religious communities in India, thus also providing a

#word_tokenize
print(word_tokenize(corpus))
```

```
['For', 'decades', 'the', 'All-India', 'Congress', 'under', 'the', 'leadership', 'of', 'Mohandas', 'K.', 'Gandhi', 'struggled', 't
o', 'rally', 'the', 'millions', 'of', 'British-ruled', 'peoples', 'in', 'the', 'Indian', 'subcontinent', '.', 'Like', 'similar', 'm
ovements', 'in', 'other', 'countries', ',', 'it', 'early', 'felt', 'the', 'need', 'for', 'a', 'distinctive', 'symbol', 'that', 'cou
ld', 'represent', 'its', 'nationalist', 'objectives', '.', 'In', '1921', 'a', 'university', 'lecturer', 'named', 'Pingali', '(', 'o
r', 'Pinglay', ')', 'Venkayya', 'presented', 'a', 'flag', 'design', 'to', 'Gandhi', 'that', 'consisted', 'of', 'the', 'colours', 'a
ssociated', 'with', 'the', 'two', 'principal', 'religions', ',', 'red', 'for', 'the', 'Hindus', 'and', 'green', 'for', 'the', 'Musl
ims', '.', 'To', 'the', 'centre', 'of', 'the', 'horizontally', 'divided', 'flag', ',', 'Lala', 'Hans', 'Raj', 'Sondhi', 'suggeste
d', 'the', 'addition', 'of', 'the', 'traditional', 'spinning', 'wheel', ',', 'which', 'was', 'associated', 'with', 'Gandhi', '’',
's', 'crusade', 'to', 'make', 'Indians', 'self-reliant', 'by', 'fabricating', 'their', 'own', 'clothing', 'from', 'local', 'fibre
s', '.', 'Gandhi', 'modified', 'the', 'flag', 'by', 'adding', 'a', 'white', 'stripe', 'in', 'the', 'centre', 'for', 'the', 'other',
'religious', 'communities', 'in', 'India', ',', 'thus', 'also', 'providing', 'a', 'clearly', 'visible', 'background', 'for', 'the',
'spinning', 'wheel', '.', 'In', 'May', '1923', 'at', 'Nagpur', ',', 'during', 'peaceful', 'protests', 'against', 'British', 'rule',
',', 'the', 'flag', 'was', 'carried', 'by', 'thousands', 'of', 'people', ',', 'hundreds', 'of', 'whom', 'were', 'arrested', '.', 'T
he', 'Congress', 'flag', 'came', 'to', 'be', 'associated', 'with', 'nationhood', 'for', 'India', ',', 'and', 'it', 'was', 'official
ly', 'recognized', 'at', 'the', 'annual', 'meeting', 'of', 'the', 'party', 'in', 'August', '1931', '.', 'At', 'the', 'same', 'tim
e', ',', 'the', 'current', 'arrangement', 'of', 'stripes', 'and', 'the', 'use', 'of', 'deep', 'saffron', 'instead', 'of', 'red', 'w
ere', 'approved', '.', 'To', 'avoid', 'the', 'sectarian', 'associations', 'of', 'the', 'original', 'proposal', ',', 'new', 'attribu
tions', 'were', 'associated', 'with', 'the', 'saffron', ',', 'white', ',', 'and', 'green', 'stripes', '.', 'They', 'were', 'said',
'to', 'stand', 'for', ',', 'respectively', ',', 'courage', 'and', 'sacrifice', ',', 'peace', 'and', 'truth', ',', 'and', 'faith',
'and', 'chivalry', '.', 'During', 'World', 'War', 'II', 'Subhas', 'Chandra', 'Bose', 'used', 'this', 'flag', '(', 'without', 'the',
'spinning', 'wheel', ')', 'in', 'territories', 'his', 'Japanese-aided', 'army', 'had', 'captured', '.']
```

In [9]:

```
#sent_tokenize
from nltk.tokenize import sent_tokenize
print(sent_tokenize(corpus))
```

```
['For decades the All-India Congress under the leadership of Mohandas K. Gandhi struggled to rally the millions of British-ruled pe
oples in the Indian subcontinent.', 'Like similar movements in other countries, it early felt the need for a distinctive symbol tha
t could represent its nationalist objectives.', 'In 1921 a university lecturer named Pingali (or Pinglay) Venkayya presented a flag
design to Gandhi that consisted of the colours associated with the two principal religions, red for the Hindus and green for the Mu
slims.', 'To the centre of the horizontally divided flag, Lala Hans Raj Sondhi suggested the addition of the traditional spinning w
heel, which was associated with Gandhi’s crusade to make Indians self-reliant by fabricating their own clothing from local fibre
s.', 'Gandhi modified the flag by adding a white stripe in the centre for the other religious communities in India, thus also provi
ding a clearly visible background for the spinning wheel.', 'In May 1923 at Nagpur, during peaceful protests against British rule,
the flag was carried by thousands of people, hundreds of whom were arrested.', 'The Congress flag came to be associated with nation
hood for India, and it was officially recognized at the annual meeting of the party in August 1931.', 'At the same time, the curren
t arrangement of stripes and the use of deep saffron instead of red were approved.', 'To avoid the sectarian associations of the or
iginal proposal, new attributions were associated with the saffron, white, and green stripes.', 'They were said to stand for, respe
ctively, courage and sacrifice, peace and truth, and faith and chivalry.', 'During World War II Subhas Chandra Bose used this flag
(without the spinning wheel) in territories his Japanese-aided army had captured.']
```

b) Using stop words eliminate most common words, do stemming and lemmatization.

In [11]:
```
#stopwords### b) Using stop words eliminate most common words, do stemming and lemmatization.
nltk.download('stopwords')
```

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

Out[11]:
```
True
```

In [16]:
```
from nltk.corpus import stopwords
stop_words=set(stopwords.words('english'))
filtered=[]
for word in word_tokens:
    if word not in stop_words:
            filtered.append(word)
print("Filtered Sentence:")
print(" ".join(filtered))
```

```
Filtered Sentence:
For decades All-India Congress leadership Mohandas K. Gandhi struggled rally millions British-ruled peoples Indian subcontinent . L
ike similar movements countries , early felt need distinctive symbol could represent nationalist objectives . In 1921 university le
cturer named Pingali ( Pinglay ) Venkayya presented flag design Gandhi consisted colours associated two principal religions , red H
indus green Muslims . To centre horizontally divided flag , Lala Hans Raj Sondhi suggested addition traditional spinning wheel , as
sociated Gandhi ’ crusade make Indians self-reliant fabricating clothing local fibres . Gandhi modified flag adding white stripe ce
ntre religious communities India , thus also providing clearly visible background spinning wheel . In May 1923 Nagpur , peaceful pr
otests British rule , flag carried thousands people , hundreds arrested . The Congress flag came associated nationhood India , offi
cially recognized annual meeting party August 1931 . At time , current arrangement stripes use deep saffron instead red approved .
To avoid sectarian associations original proposal , new attributions associated saffron , white , green stripes . They said stand ,
respectively , courage sacrifice , peace truth , faith chivalry . During World War II Subhas Chandra Bose used flag ( without spinn
ing wheel ) territories Japanese-aided army captured .
```

In [34]:
```
nltk.download('wordnet')
```

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

Out[34]:
```
True
```

In [42]:
```
from nltk.stem import PorterStemmer
stem_words=[]
porste=PorterStemmer()
for j in filtered:
    rootWord=porste.stem(j)
    stem_words.append(rootWord)
print(filtered)
print(stem_words)
```

```
['For', 'decades', 'All-India', 'Congress', 'leadership', 'Mohandas', 'K.', 'Gandhi', 'struggled', 'rally', 'millions', 'British-ru
led', 'peoples', 'Indian', 'subcontinent', '.', 'Like', 'similar', 'movements', 'countries', ',', 'early', 'felt', 'need', 'distinc
tive', 'symbol', 'could', 'represent', 'nationalist', 'objectives', '.', 'In', '1921', 'university', 'lecturer', 'named', 'Pingal
i', '(', 'Pinglay', ')', 'Venkayya', 'presented', 'flag', 'design', 'Gandhi', 'consisted', 'colours', 'associated', 'two', 'princip
al', 'religions', ',', 'red', 'Hindus', 'green', 'Muslims', '.', 'To', 'centre', 'horizontally', 'divided', 'flag', ',', 'Lala', 'H
ans', 'Raj', 'Sondhi', 'suggested', 'addition', 'traditional', 'spinning', 'wheel', ',', 'associated', 'Gandhi', '’', 'crusade', 'm
ake', 'Indians', 'self-reliant', 'fabricating', 'clothing', 'local', 'fibres', '.', 'Gandhi', 'modified', 'flag', 'adding', 'whit
e', 'stripe', 'centre', 'religious', 'communities', 'India', ',', 'thus', 'also', 'providing', 'clearly', 'visible', 'background',
'spinning', 'wheel', '.', 'In', 'May', '1923', 'Nagpur', ',', 'peaceful', 'protests', 'British', 'rule', ',', 'flag', 'carried', 't
housands', 'people', ',', 'hundreds', 'arrested', '.', 'The', 'Congress', 'flag', 'came', 'associated', 'nationhood', 'India', ',',
'officially', 'recognized', 'annual', 'meeting', 'party', 'August', '1931', '.', 'At', 'time', ',', 'current', 'arrangement', 'stri
pes', 'use', 'deep', 'saffron', 'instead', 'red', 'approved', '.', 'To', 'avoid', 'sectarian', 'associations', 'original', 'proposa
l', ',', 'new', 'attributions', 'associated', 'saffron', ',', 'white', ',', 'green', 'stripes', '.', 'They', 'said', 'stand', ',',
'respectively', ',', 'courage', 'sacrifice', ',', 'peace', 'truth', ',', 'faith', 'chivalry', '.', 'During', 'World', 'War', 'II',
'Subhas', 'Chandra', 'Bose', 'used', 'flag', '(', 'without', 'spinning', 'wheel', ')', 'territories', 'Japanese-aided', 'army', 'ca
ptured', '.']
['for', 'decad', 'all-india', 'congress', 'leadership', 'mohanda', 'k.', 'gandhi', 'struggl', 'ralli', 'million', 'british-rul', 'p
eopl', 'indian', 'subcontin', '.', 'like', 'similar', 'movement', 'countri', ',', 'earli', 'felt', 'need', 'distinct', 'symbol', 'c
ould', 'repres', 'nationalist', 'object', '.', 'in', '1921', 'univers', 'lectur', 'name', 'pingali', '(', 'pinglay', ')', 'venkayy
a', 'present', 'flag', 'design', 'gandhi', 'consist', 'colour', 'associ', 'two', 'princip', 'religion', ',', 'red', 'hindu', 'gree
n', 'muslim', '.', 'to', 'centr', 'horizont', 'divid', 'flag', ',', 'lala', 'han', 'raj', 'sondhi', 'suggest', 'addit', 'tradit',
'spin', 'wheel', ',', 'associ', 'gandhi', '’', 'crusad', 'make', 'indian', 'self-reli', 'fabric', 'cloth', 'local', 'fibr', '.', 'g
andhi', 'modifi', 'flag', 'ad', 'white', 'stripe', 'centr', 'religi', 'commun', 'india', ',', 'thu', 'also', 'provid', 'clearli',
'visibl', 'background', 'spin', 'wheel', '.', 'in', 'may', '1923', 'nagpur', ',', 'peac', 'protest', 'british', 'rule', ',', 'fla
g', 'carri', 'thousand', 'peopl', ',', 'hundr', 'arrest', '.', 'the', 'congress', 'flag', 'came', 'associ', 'nationhood', 'india',
',', 'offici', 'recogn', 'annual', 'meet', 'parti', 'august', '1931', '.', 'at', 'time', ',', 'current', 'arrang', 'stripe', 'use',
'deep', 'saffron', 'instead', 'red', 'approv', '.', 'to', 'avoid', 'sectarian', 'associ', 'origin', 'propos', ',', 'new', 'attribu
t', 'associ', 'saffron', ',', 'white', ',', 'green', 'stripe', '.', 'they', 'said', 'stand', ',', 'respect', ',', 'courag', 'sacrif
ic', ',', 'peac', 'truth', ',', 'faith', 'chivalri', '.', 'dure', 'world', 'war', 'ii', 'subha', 'chandra', 'bose', 'use', 'flag',
'(', 'without', 'spin', 'wheel', ')', 'territori', 'japanese-aid', 'armi', 'captur', '.']
```

In [43]:
```
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')
```

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

Out[43]:
```
True
```

In [45]:
```
lemma_word=[]
wordnet_lemmatizer=WordNetLemmatizer()
for w in filtered:
    word1 = wordnet_lemmatizer.lemmatize(w, pos = "n")
```

```
    word2 = wordnet_lemmatizer.lemmatize(word1, pos = "v")
    word3 = wordnet_lemmatizer.lemmatize(word2, pos = ("a"))
    lemma_word.append(word3)
print(lemma_word)
```

```
['For', 'decade', 'All-India', 'Congress', 'leadership', 'Mohandas', 'K.', 'Gandhi', 'struggle', 'rally', 'million', 'British-rule
d', 'people', 'Indian', 'subcontinent', '.', 'Like', 'similar', 'movement', 'country', ',', 'early', 'felt', 'need', 'distinctive',
'symbol', 'could', 'represent', 'nationalist', 'objective', '.', 'In', '1921', 'university', 'lecturer', 'name', 'Pingali', '(', 'P
inglay', ')', 'Venkayya', 'present', 'flag', 'design', 'Gandhi', 'consist', 'colour', 'associate', 'two', 'principal', 'religion',
',', 'red', 'Hindus', 'green', 'Muslims', '.', 'To', 'centre', 'horizontally', 'divide', 'flag', ',', 'Lala', 'Hans', 'Raj', 'Sondh
i', 'suggest', 'addition', 'traditional', 'spin', 'wheel', ',', 'associate', 'Gandhi', '’', 'crusade', 'make', 'Indians', 'self-rel
iant', 'fabricate', 'clothe', 'local', 'fibre', '.', 'Gandhi', 'modify', 'flag', 'add', 'white', 'stripe', 'centre', 'religious',
'community', 'India', ',', 'thus', 'also', 'provide', 'clearly', 'visible', 'background', 'spin', 'wheel', '.', 'In', 'May', '192
3', 'Nagpur', ',', 'peaceful', 'protest', 'British', 'rule', ',', 'flag', 'carry', 'thousand', 'people', ',', 'hundred', 'arrest',
'.', 'The', 'Congress', 'flag', 'come', 'associate', 'nationhood', 'India', ',', 'officially', 'recognize', 'annual', 'meet', 'part
y', 'August', '1931', '.', 'At', 'time', ',', 'current', 'arrangement', 'stripe', 'use', 'deep', 'saffron', 'instead', 'red', 'appr
ove', '.', 'To', 'avoid', 'sectarian', 'association', 'original', 'proposal', ',', 'new', 'attribution', 'associate', 'saffron',
',', 'white', ',', 'green', 'stripe', '.', 'They', 'say', 'stand', ',', 'respectively', ',', 'courage', 'sacrifice', ',', 'peace',
'truth', ',', 'faith', 'chivalry', '.', 'During', 'World', 'War', 'II', 'Subhas', 'Chandra', 'Bose', 'use', 'flag', '(', 'without',
'spin', 'wheel', ')', 'territory', 'Japanese-aided', 'army', 'capture', '.']
```

Q2). Copy the paragraph and apply the bag-of-words approach; Also, identify the bag-of vector for each sentence.

Construction of the mausoleum was essentially completed in 1643. but work continued on other phases of the project for another 10 years. The Taj Mahal complex is believed to have been completed in its entirety in 1653 at a cost estimated at the time to be around 32 million. The construction project employed some 20,000 artisans under the guidance of a board of architects led by the court architect to the emperor, Ustad Ahmad Lahauri. Various types of symbolism have been employed in the Taj to reflect natural beauty and divinity.

In [1]:
```
# import libraries
import nltk
import re
import numpy as np
nltk.download('punkt')
```

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

Out[1]:
```
True
```

In [17]:
```
text="""Construction of the mausoleum was essentially completed in 1643. but work continued on
other phases of the project for another 10 years. The Taj Mahal complex is believed to have
been completed in its entirety in 1653 at a cost estimated at the time to be around 32 million.
The construction project employed some 20,000 artisans under the guidance of a board of
architects led by the court architect to the emperor, Ustad Ahmad Lahauri. Various types of
symbolism have been employed in the Taj to reflect natural beauty and divinity."""

data=nltk.sent_tokenize(text)

#converting the text into lower case and removing non-word characters and punctuations also
for i in range(len(data)):
    data[i] = data[i].lower()
    data[i] = re.sub(r'\W', ' ', data[i])
    data[i] = re.sub(r'\s+', ' ', data[i])
data
```

Out[17]:
```
['construction of the mausoleum was essentially completed in 1643 but work continued on other phases of the project for another 10
years ',
 'the taj mahal complex is believed to have been completed in its entirety in 1653 at a cost estimated at the time to be around 32
million ',
 'the construction project employed some 20 000 artisans under the guidance of a board of architects led by the court architect to
the emperor ustad ahmad lahauri ',
 'various types of symbolism have been employed in the taj to reflect natural beauty and divinity ']
```

In [20]:
```
# creating a bag words to hold the words and their counts
word_count={}
for d in data:
    words=nltk.word_tokenize(d)
    for word in words:
        if word not in word_count.keys():
            word_count[word]=1
        else:
            word_count[word]+=1
word_count
```

Out[20]:

```
{'construction': 2,
 'of': 5,
 'the': 9,
 'mausoleum': 1,
 'was': 1,
 'essentially': 1,
 'completed': 2,
 'in': 4,
 '1643': 1,
 'but': 1,
 'work': 1,
 'continued': 1,
 'on': 1,
 'other': 1,
 'phases': 1,
 'project': 2,
 'for': 1,
 'another': 1,
 '10': 1,
 'years': 1,
 'taj': 2,
 'mahal': 1,
 'complex': 1,
 'is': 1,
 'believed': 1,
 'to': 4,
 'have': 2,
 'been': 2,
 'its': 1,
 'entirety': 1,
 '1653': 1,
 'at': 2,
 'a': 2,
 'cost': 1,
 'estimated': 1,
 'time': 1,
 'be': 1,
 'around': 1,
 '32': 1,
 'million': 1,
 'employed': 2,
 'some': 1,
 '20': 1,
 '000': 1,
```

```
 'artisans': 1,
 'under': 1,
 'guidance': 1,
 'board': 1,
 'architects': 1,
 'led': 1,
 'by': 1,
 'court': 1,
 'architect': 1,
 'emperor': 1,
 'ustad': 1,
 'ahmad': 1,
 'lahauri': 1,
 'various': 1,
 'types': 1,
 'symbolism': 1,
 'reflect': 1,
 'natural': 1,
 'beauty': 1,
 'and': 1,
 'divinity': 1}
```

In [22]:
```
#count the items
len(word_count)
```

Out[22]:
```
65
```

In [23]:
```
# taking the most frequently used words
import heapq

# lets choose 30 most frequent words

freq_words=heapq.nlargest(30, word_count, key=word_count.get)
freq_words
```

Out[23]:
```
['the',
 'of',
 'in',
 'to',
 'construction',
 'completed',
 'project',
 'taj',
 'have',
 'been',
 'at',
 'a',
 'employed',
 'mausoleum',
 'was',
 'essentially',
 '1643',
 'but',
 'work',
 'continued',
 'on',
 'other',
 'phases',
 'for',
 'another',
 '10',
 'years',
 'mahal',
 'complex',
 'is']
```

In [26]:
```
# finding the bag of vector
x=[]
for d in data:
    vector=[]
    for word in freq_words:
        if word in nltk.word_tokenize(d):
            vector.append(1)
        else:
            vector.append(0)
    x.append(vector)
x=np.asarray(x)
x
```

Out[26]:
```
array([[1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 0, 0, 0],
       [1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 1, 1, 1],
       [1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0],
       [1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0]])
```

Q3). Copy the paragraph and apply TFIDF method and find the feature vector for each sentence.

Referred to as the Venice of the East, Alappuzha has always enjoyed an important place in the maritime history of Kerala. Today, it is famous for its boat races, backwater holidays, beaches, marine products and coir industry. Alappuzha Beach is a popular picnic spot.

In [27]:
```
# import ibraries
import pandas as pd
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [30]:
```
text="""Referred to as the Venice of the East, Alappuzha has always enjoyed an important place
in the maritime history of Kerala. Today, it is famous for its boat races, backwater holidays,
beaches, marine products and coir industry. Alappuzha Beach is a popular picnic spot."""

#converting to lower case
lower_text=text.lower()
sents=sent_tokenize(lower_text)
sents
```

Out[30]:
```
['referred to as the venice of the east, alappuzha has always enjoyed an important place \nin the maritime history of kerala.',
 'today, it is famous for its boat races, backwater holidays, \nbeaches, marine products and coir industry.',
 'alappuzha beach is a popular picnic spot.']
```

In [32]:
```
for sent in sents:
    words=word_tokenize(lower_text)   # word tokenizing the text
    words=[word for word in words if word not in stopwords.words("english")]  # removing stop words
    lower_text= " ".join(words)
lower_text=[lower_text]
lower_text
```

Out[32]:
```
['referred venice east , alappuzha always enjoyed important place maritime history kerala . today , famous boat races , backwater h
olidays , beaches , marine products coir industry . alappuzha beach popular picnic spot .']
```

In [33]:
```
vectorizer = TfidfVectorizer()
tfidf_model = vectorizer.fit_transform(lower_text)
print(tfidf_model)
```

```
  (0, 23)       0.18569533817705186
  (0, 17)       0.18569533817705186
  (0, 19)       0.18569533817705186
  (0, 3)        0.18569533817705186
  (0, 13)       0.18569533817705186
  (0, 6)        0.18569533817705186
  (0, 20)       0.18569533817705186
  (0, 15)       0.18569533817705186
  (0, 4)        0.18569533817705186
  (0, 11)       0.18569533817705186
  (0, 2)        0.18569533817705186
  (0, 21)       0.18569533817705186
  (0, 5)        0.18569533817705186
  (0, 9)        0.18569533817705186
  (0, 24)       0.18569533817705186
  (0, 14)       0.18569533817705186
  (0, 10)       0.18569533817705186
  (0, 16)       0.18569533817705186
  (0, 18)       0.18569533817705186
  (0, 12)       0.18569533817705186
  (0, 8)        0.18569533817705186
  (0, 1)        0.18569533817705186
  (0, 0)        0.3713906763541037
  (0, 7)        0.18569533817705186
  (0, 25)       0.18569533817705186
  (0, 22)       0.18569533817705186
```

In [34]:
```
print(tfidf_model.toarray())
```

```
[[0.37139068 0.18569534 0.18569534 0.18569534 0.18569534 0.18569534
  0.18569534 0.18569534 0.18569534 0.18569534 0.18569534 0.18569534
  0.18569534 0.18569534 0.18569534 0.18569534 0.18569534 0.18569534
  0.18569534 0.18569534 0.18569534 0.18569534 0.18569534 0.18569534
  0.18569534 0.18569534]]
```

In [40]:
```
#to dataframe
import pandas as pd
```

```
# Assuming tfidf_model and vectorizer are already defined
tfidf_matrix=tfidf_model.toarray()
feature_names=vectorizer.get_feature_names_out()

# Convert the TF-IDF matrix to a DataFrame
tfidf_df=pd.DataFrame(tfidf_matrix, columns=feature_names)
tfidf_df
```

Out[40]:

	alappuzha	always	backwater	beach	beaches	boat	coir	east	enjoyed	famous	...	maritime	picnic	place	popular	products
0	0.371391	0.185695	0.185695	0.185695	0.185695	0.185695	0.185695	0.185695	0.185695	0.185695	...	0.185695	0.185695	0.185695	0.185695	0.185695

1 rows × 26 columns

In [ ]: