In [10]:

```
import nltk
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, WordNetLemmatizer
```

1. Copy the next paragraph and answer the questions that follow:

a) Word_tokenise and sent_tokenise

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

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.

In [21]:

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

In [22]:

```
sentences = sent_tokenize(text)
print("Sentences:")
for sentence in sentences:
    print(sentence)
```

```
Sentences:
For decades the All-India Congress under the leadership of Mohandas K. Gandhi struggled to rally the millions of British-ru
led peoples in the Indian subcontinent.
Like similar movements in other countries, it early felt the need for a distinctive symbol that could represent its nationa
list objectives.
In 1921 a university lecturer named Pingali (or Pinglay) Venkayya presented a flag design to Gandhi that consisted of the c
olours 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 whe
el, which was associated with Gandhi’s crusade to make Indians self-reliant by fabricating their own clothing from local fi
bres.Gandhi modified the flag by adding a white stripe in the centre for the other religious communities in India, thus als
o 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 o
f 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.
```

In [23]:

```
words = word_tokenize(text)
print("\nWords:")
for word in words:
    print(word)
```

```
Subhas
Chandra
Bose
used
this
flag
(
without
the
spinning
wheel
)
in
territories
his
Japanese-aided
army
had
captured
.
```

In [24]:

```
stop_words = set(stopwords.words('english'))
filtered_words = [word for word in words if word.lower() not in stop_words]
```

In [25]:

```
stemmer = PorterStemmer()
stemmed_words = [stemmer.stem(word) for word in filtered_words]
```

In [26]:

```
lemmatizer = WordNetLemmatizer()
lemmatized_words = [lemmatizer.lemmatize(word) for word in filtered_words]

print("Original Words:", words)
print("\nFiltered Words (without stop words):", filtered_words)
print("\nStemmed Words:", stemmed_words)
print("\nLemmatized Words:", lemmatized_words)
```

```
Original Words: ['For', 'decades', 'the', 'All-India', 'Congress', 'under', 'the', 'leadership', 'of', 'Mohandas', 'K.', 'G
andhi', 'struggled', 'to', 'rally', 'the', 'millions', 'of', 'British-ruled', 'peoples', 'in', 'the', 'Indian', 'subcontine
nt', '.', 'Like', 'similar', 'movements', 'in', 'other', 'countries', ',', 'it', 'early', 'felt', 'the', 'need', 'for',
'a', 'distinctive', 'symbol', 'that', 'could', 'represent', 'its', 'nationalist', 'objectives', '.', 'In', '1921', 'a', 'un
iversity', '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', 'horizonta
lly', 'divided', 'flag', ',', 'Lala', 'Hans', 'Raj', 'Sondhi', 'suggested', 'the', 'addition', 'of', 'the', 'traditional',
'spinning', 'wheel', ',', 'which', 'was', 'associated', 'with', 'Gandhi', '’', 's', 'crusade', 'to', 'make', 'Indians', 'se
lf-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', 'In
dia', ',', 'thus', 'also', 'providing', 'a', 'clearly', 'visible', 'background', 'for', 'the', 'spinning', 'wheel', '.', 'I
n', 'May', '1923', 'at', 'Nagpur', ',', 'during', 'peaceful', 'protests', 'against', 'British', 'rule', ',', 'the', 'flag',
'was', 'carried', 'by', 'thousands', 'of', 'people', ',', 'hundreds', 'of', 'whom', 'were', 'arrested', '.', 'The', 'Congre
ss', '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', 'ti
me', ',', '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', 'arm
y', 'had', 'captured', '.']

Filtered Words (without stop words): ['decades', 'All-India', 'Congress', 'leadership', 'Mohandas', 'K.', 'Gandhi', 'strugg
led', 'rally', 'millions', 'British-ruled', 'peoples', 'Indian', 'subcontinent', '.', 'Like', 'similar', 'movements', 'coun
tries', ',', 'early', 'felt', 'need', 'distinctive', 'symbol', 'could', 'represent', 'nationalist', 'objectives', '.', '192
1', 'university', 'lecturer', 'named', 'Pingali', '(', 'Pinglay', ')', 'Venkayya', 'presented', 'flag', 'design', 'Gandhi',
'consisted', 'colours', 'associated', 'two', 'principal', 'religions', ',', 'red', 'Hindus', 'green', 'Muslims', '.', 'cent
re', 'horizontally', 'divided', 'flag', ',', 'Lala', 'Hans', 'Raj', 'Sondhi', 'suggested', 'addition', 'traditional', 'spin
ning', 'wheel', ',', 'associated', 'Gandhi', '’', 'crusade', 'make', 'Indians', 'self-reliant', 'fabricating', 'clothing',
'local', 'fibres.Gandhi', 'modified', 'flag', 'adding', 'white', 'stripe', 'centre', 'religious', 'communities', 'India',
',', 'thus', 'also', 'providing', 'clearly', 'visible', 'background', 'spinning', 'wheel', '.', 'May', '1923', 'Nagpur',
',', 'peaceful', 'protests', 'British', 'rule', ',', 'flag', 'carried', 'thousands', 'people', ',', 'hundreds', 'arrested',
'.', 'Congress', 'flag', 'came', 'associated', 'nationhood', 'India', ',', 'officially', 'recognized', 'annual', 'meeting',
'party', 'August', '1931', '.', 'time', ',', 'current', 'arrangement', 'stripes', 'use', 'deep', 'saffron', 'instead', 're
d', 'approved', '.', 'avoid', 'sectarian', 'associations', 'original', 'proposal', ',', 'new', 'attributions', 'associate
d', 'saffron', ',', 'white', ',', 'green', 'stripes', '.', 'said', 'stand', ',', 'respectively', ',', 'courage', 'sacrific
e', ',', 'peace', 'truth', ',', 'faith', 'chivalry', '.', 'World', 'War', 'II', 'Subhas', 'Chandra', 'Bose', 'used', 'fla
g', '(', 'without', 'spinning', 'wheel', ')', 'territories', 'Japanese-aided', 'army', 'captured', '.']

Stemmed Words: ['decad', 'all-india', 'congress', 'leadership', 'mohanda', 'K.', 'gandhi', 'struggl', 'ralli', 'million',
'british-rul', 'peopl', 'indian', 'subcontin', '.', 'like', 'similar', 'movement', 'countri', ',', 'earli', 'felt', 'need',
'distinct', 'symbol', 'could', 'repres', 'nationalist', 'object', '.', '1921', 'univers', 'lectur', 'name', 'pingali', '(',
'pinglay', ')', 'venkayya', 'present', 'flag', 'design', 'gandhi', 'consist', 'colour', 'associ', 'two', 'princip', 'religi
on', ',', 'red', 'hindu', 'green', 'muslim', '.', 'centr', 'horizont', 'divid', 'flag', ',', 'lala', 'han', 'raj', 'sondh
i', 'suggest', 'addit', 'tradit', 'spin', 'wheel', ',', 'associ', 'gandhi', '’', 'crusad', 'make', 'indian', 'self-reli',
'fabric', 'cloth', 'local', 'fibres.gandhi', 'modifi', 'flag', 'ad', 'white', 'stripe', 'centr', 'religi', 'commun', 'indi
a', ',', 'thu', 'also', 'provid', 'clearli', 'visibl', 'background', 'spin', 'wheel', '.', 'may', '1923', 'nagpur', ',', 'p
eac', 'protest', 'british', 'rule', ',', 'flag', 'carri', 'thousand', 'peopl', ',', 'hundr', 'arrest', '.', 'congress', 'fl
ag', 'came', 'associ', 'nationhood', 'india', ',', 'offici', 'recogn', 'annual', 'meet', 'parti', 'august', '1931', '.', 't
ime', ',', 'current', 'arrang', 'stripe', 'use', 'deep', 'saffron', 'instead', 'red', 'approv', '.', 'avoid', 'sectarian',
'associ', 'origin', 'propos', ',', 'new', 'attribut', 'associ', 'saffron', ',', 'white', ',', 'green', 'stripe', '.', 'sai
d', 'stand', ',', 'respect', ',', 'courag', 'sacrific', ',', 'peac', 'truth', ',', 'faith', 'chivalri', '.', 'world', 'wa
r', 'II', 'subha', 'chandra', 'bose', 'use', 'flag', '(', 'without', 'spin', 'wheel', ')', 'territori', 'japanese-aid', 'ar
mi', 'captur', '.']

Lemmatized Words: ['decade', 'All-India', 'Congress', 'leadership', 'Mohandas', 'K.', 'Gandhi', 'struggled', 'rally', 'mill
ion', 'British-ruled', 'people', 'Indian', 'subcontinent', '.', 'Like', 'similar', 'movement', 'country', ',', 'early', 'fe
lt', 'need', 'distinctive', 'symbol', 'could', 'represent', 'nationalist', 'objective', '.', '1921', 'university', 'lecture
r', 'named', 'Pingali', '(', 'Pinglay', ')', 'Venkayya', 'presented', 'flag', 'design', 'Gandhi', 'consisted', 'colour', 'a
ssociated', 'two', 'principal', 'religion', ',', 'red', 'Hindus', 'green', 'Muslims', '.', 'centre', 'horizontally', 'divid
ed', 'flag', ',', 'Lala', 'Hans', 'Raj', 'Sondhi', 'suggested', 'addition', 'traditional', 'spinning', 'wheel', ',', 'assoc
iated', 'Gandhi', '’', 'crusade', 'make', 'Indians', 'self-reliant', 'fabricating', 'clothing', 'local', 'fibres.Gandhi',
'modified', 'flag', 'adding', 'white', 'stripe', 'centre', 'religious', 'community', 'India', ',', 'thus', 'also', 'providi
ng', 'clearly', 'visible', 'background', 'spinning', 'wheel', '.', 'May', '1923', 'Nagpur', ',', 'peaceful', 'protest', 'Br
itish', 'rule', ',', 'flag', 'carried', 'thousand', 'people', ',', 'hundred', 'arrested', '.', 'Congress', 'flag', 'came',
'associated', 'nationhood', 'India', ',', 'officially', 'recognized', 'annual', 'meeting', 'party', 'August', '1931', '.',
'time', ',', 'current', 'arrangement', 'stripe', 'use', 'deep', 'saffron', 'instead', 'red', 'approved', '.', 'avoid', 'sec
tarian', 'association', 'original', 'proposal', ',', 'new', 'attribution', 'associated', 'saffron', ',', 'white', ',', 'gre
en', 'stripe', '.', 'said', 'stand', ',', 'respectively', ',', 'courage', 'sacrifice', ',', 'peace', 'truth', ',', 'faith',
'chivalry', '.', 'World', 'War', 'II', 'Subhas', 'Chandra', 'Bose', 'used', 'flag', '(', 'without', 'spinning', 'wheel',
')', 'territory', 'Japanese-aided', 'army', 'captured', '.']
```

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

In [27]:

```
from sklearn.feature_extraction.text import CountVectorizer
```

In [31]:

```
text = "Construction of the mausoleum was essentially completed in 1643. but work continued on other phases of the project for another 10
```

In [32]:

```
sentences = sent_tokenize(text)
```

In [33]:

```
tokenized_sentences = [word_tokenize(sentence) for sentence in sentences]
```

In [34]:

```
sentence_strings = [' '.join(tokens) for tokens in tokenized_sentences]
```

In [35]:

```
vectorizer = CountVectorizer()
```

In [36]:

```
bag_of_words_vectors = vectorizer.fit_transform(sentence_strings)
```

In [39]:

```
bag_of_words_array = bag_of_words_vectors.toarray()
```

In [40]:

```
for i, sentence in enumerate(sentences):
    print(f"Sentence {i + 1}: {sentence}")
    print("Bag-of-Words Vector:", bag_of_words_array[i])
    print()
```

```
Sentence 1: Construction of the mausoleum was essentially completed in 1643. but work continued on other phases of the proj
ect for another 10 years.
Bag-of-Words Vector: [0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 0 0 0 0 0 1 0 1 0 0 1
 0 0 0 0 0 1 0 0 2 1 1 1 1 0 0 0 0 2 0 0 0 0 0 0 1 1 1]

Sentence 2: The Taj Mahal complex is believed to have been completed in its entirety in 1653 at a cost estimated at the tim
e to be around 32 million.
Bag-of-Words Vector: [0 0 0 1 0 1 0 0 0 0 0 1 0 2 1 0 1 1 0 0 0 1 1 0 0 1 0 0 0 0 1 0 1 0 0 1 2
 1 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 2 1 2 0 0 0 0 0 0 0]

Sentence 3: The construction project employed some 20,000 artisans under the guidance of a board of architects led by the c
ourt architect to the emperor, Ustad Ahmad Lahauri.
Bag-of-Words Vector: [1 0 0 0 1 0 1 0 0 1 1 0 1 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 1 0 0
 0 0 1 1 0 0 0 0 2 0 0 0 1 0 1 0 0 4 0 1 0 1 1 0 0 0 0]

Sentence 4: Various types of symbolism have been employed in the Taj to reflect natural beauty and divinity.
Bag-of-Words Vector: [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1
 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 1 1 0 1 1 0 0 1 0 0 0]
```

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 impo rtant 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 [42]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [43]:

```
: is famous for its boat races, backwater holidays, beaches, marine products and coir industry. Alappuzha Beach is a popular picnic spot."
```

In [44]:

```
sentences = sent_tokenize(paragraph)
```

In [45]:

```
tokenized_sentences = [word_tokenize(sentence) for sentence in sentences]
```

In [46]:

```
sentence_strings = [' '.join(tokens) for tokens in tokenized_sentences]
```

In [47]:

```
vectorizer = TfidfVectorizer()
```

In [48]:

```
tfidf_vectors = vectorizer.fit_transform(sentence_strings)
```

In [50]:

```
tfidf_array = tfidf_vectors.toarray()
```

In [51]:

```
for i, sentence in enumerate(sentences):
    print(f"Sentence {i + 1}: {sentence}")
    print("TF-IDF Feature Vector:", tfidf_array[i])
    print()
```

```
Sentence 1: Referred to as the Venice of the East, Alappuzha has always enjoyed an important place in the maritime history
of Kerala.
TF-IDF Feature Vector: [0.14226399 0.18706005 0.18706005 0.         0.18706005 0.
 0.         0.         0.         0.         0.18706005 0.18706005
 0.         0.         0.18706005 0.18706005 0.         0.18706005
 0.18706005 0.         0.         0.         0.         0.18706005
 0.         0.18706005 0.3741201  0.         0.18706005 0.
 0.         0.         0.18706005 0.         0.56118015 0.18706005
 0.         0.18706005]

Sentence 2: Today, it is famous for its boat races, backwater holidays, beaches, marine products and coir industry.
TF-IDF Feature Vector: [0.         0.         0.         0.25336031 0.         0.25336031
 0.         0.25336031 0.25336031 0.25336031 0.         0.
 0.25336031 0.25336031 0.         0.         0.25336031 0.
 0.         0.25336031 0.19268705 0.25336031 0.25336031 0.
 0.25336031 0.         0.         0.         0.         0.
 0.25336031 0.25336031 0.         0.         0.         0.
 0.25336031 0.        ]

Sentence 3: Alappuzha Beach is a popular picnic spot.
TF-IDF Feature Vector: [0.3349067  0.         0.         0.         0.         0.
 0.44036207 0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.3349067  0.         0.         0.
 0.         0.         0.44036207 0.         0.44036207
 0.         0.         0.         0.44036207 0.         0.
 0.         0.        ]
```

In [ ]:

```
```