Practical No. 7

Part A

In [33]:

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

In [ ]:

```
import nltk
nltk.download('all')
```

In [20]:

```
document = """Natural language processing (NLP) is a subfield of artificial intelli
```

In [ ]:

```
# Tokenization
"""
In Python tokenization basically refers to splitting up a larger body of text into
"""
tokens = word_tokenize(document)
```

In [ ]:

```
# POS Tagging
"""
POS Tagging Parts of speech Tagging is responsible for reading the text in a langua
"""
pos_tags = pos_tag(tokens)
```

In [ ]:

```
# Stop words removal
"""
Stop words removal in Python is a common preprocessing step in Natural Language Pro
Stop words are words that do not add much meaning to a sentence and are pre-defined
"""
stop_words = set(stopwords.words('english'))
filtered_tokens = [token for token in tokens if token.lower() not in stop_words]
```

In [24]:

```
# Stemming
stemmer = PorterStemmer()
stemmed_tokens = [stemmer.stem(token) for token in filtered_tokens]
```

In [25]:

```
# Lemmatization
lemmatizer = WordNetLemmatizer()
lemmatized_tokens = [lemmatizer.lemmatize(token) for token in filtered_tokens]
```

In [ ]:

```
print("Original Document:\n", document)
print("\nTokens:\n", tokens)
print("\nPOS Tags:\n", pos_tags)
print("\nFiltered Tokens (after stop words removal):\n", filtered_tokens)
print("\nStemmed Tokens:\n", stemmed_tokens)
print("\nLemmatized Tokens:\n", lemmatized_tokens)
```

```
Original Document:
 Natural language processing (NLP) is a subfield of artificial intelligence (AI) tha
t focuses on the interaction between computers and humans using natural language. It
involves the analysis, understanding, and generation of human language, enabling mac
hines to process and comprehend text in a meaningful way. NLP techniques are widely
used in various applications such as sentiment analysis, machine translation, chatbo
ts, and information retrieval. Preprocessing is an essential step in NLP, which invo
lves tokenization, part-of-speech tagging, stop words removal, stemming, and lemmati
zation.

Tokens:
 ['Natural', 'language', 'processing', '(', 'NLP', ')', 'is', 'a', 'subfield', 'of',
'artificial', 'intelligence', '(', 'AI', ')', 'that', 'focuses', 'on', 'the', 'inter
action', 'between', 'computers', 'and', 'humans', 'using', 'natural', 'language',
'.', 'It', 'involves', 'the', 'analysis', ',', 'understanding', ',', 'and', 'generat
ion', 'of', 'human', 'language', ',', 'enabling', 'machines', 'to', 'process', 'an
d', 'comprehend', 'text', 'in', 'a', 'meaningful', 'way', '.', 'NLP', 'techniques',
'are', 'widely', 'used', 'in', 'various', 'applications', 'such', 'as', 'sentiment',
'analysis', ',', 'machine', 'translation', ',', 'chatbots', ',', 'and', 'informatio
n', 'retrieval', '.', 'Preprocessing', 'is', 'an', 'essential', 'step', 'in', 'NLP',
',', 'which', 'involves', 'tokenization', ',', 'part-of-speech', 'tagging', ',', 'st
op', 'words', 'removal', ',', 'stemming', ',', 'and', 'lemmatization', '.']

POS Tags:
 [('Natural', 'JJ'), ('language', 'NN'), ('processing', 'NN'), ('(', '('), ('NLP',
'NNP'), (')', ')'), ('is', 'VBZ'), ('a', 'DT'), ('subfield', 'NN'), ('of', 'IN'),
('artificial', 'JJ'), ('intelligence', 'NN'), ('(', '('), ('AI', 'NNP'), (')', ')'),
('that', 'WDT'), ('focuses', 'VBZ'), ('on', 'IN'), ('the', 'DT'), ('interaction', 'N
N'), ('between', 'IN'), ('computers', 'NNS'), ('and', 'CC'), ('humans', 'NNS'), ('us
ing', 'VBG'), ('natural', 'JJ'), ('language', 'NN'), ('.', '.'), ('It', 'PRP'), ('in
volves', 'VBZ'), ('the', 'DT'), ('analysis', 'NN'), (',', ','), ('understanding', 'N
N'), (',', ','), ('and', 'CC'), ('generation', 'NN'), ('of', 'IN'), ('human', 'JJ'),
('language', 'NN'), (',', ','), ('enabling', 'VBG'), ('machines', 'NNS'), ('to', 'T
O'), ('process', 'VB'), ('and', 'CC'), ('comprehend', 'VB'), ('text', 'NN'), ('in',
'IN'), ('a', 'DT'), ('meaningful', 'JJ'), ('way', 'NN'), ('.', '.'), ('NLP', 'NNP'),
('techniques', 'NNS'), ('are', 'VBP'), ('widely', 'RB'), ('used', 'VBN'), ('in', 'I
N'), ('various', 'JJ'), ('applications', 'NNS'), ('such', 'JJ'), ('as', 'IN'), ('sen
timent', 'NN'), ('analysis', 'NN'), (',', ','), ('machine', 'NN'), ('translation',
'NN'), (',', ','), ('chatbots', 'NNS'), (',', ','), ('and', 'CC'), ('information',
'NN'), ('retrieval', 'NN'), ('.', '.'), ('Preprocessing', 'NNP'), ('is', 'VBZ'), ('a
n', 'DT'), ('essential', 'JJ'), ('step', 'NN'), ('in', 'IN'), ('NLP', 'NNP'), (',',
','), ('which', 'WDT'), ('involves', 'VBZ'), ('tokenization', 'NN'), (',', ','), ('p
art-of-speech', 'JJ'), ('tagging', 'NN'), (',', ','), ('stop', 'VB'), ('words', 'NN
S'), ('removal', 'JJ'), (',', ','), ('stemming', 'VBG'), (',', ','), ('and', 'CC'),
('lemmatization', 'NN'), ('.', '.')]

Filtered Tokens (after stop words removal):
 ['Natural', 'language', 'processing', '(', 'NLP', ')', 'subfield', 'artificial', 'i
ntelligence', '(', 'AI', ')', 'focuses', 'interaction', 'computers', 'humans', 'usin
g', 'natural', 'language', '.', 'involves', 'analysis', ',', 'understanding', ',',
'generation', 'human', 'language', ',', 'enabling', 'machines', 'process', 'comprehe
nd', 'text', 'meaningful', 'way', '.', 'NLP', 'techniques', 'widely', 'used', 'vario
us', 'applications', 'sentiment', 'analysis', ',', 'machine', 'translation', ',', 'c
hatbots', ',', 'information', 'retrieval', '.', 'Preprocessing', 'essential', 'ste
p', 'NLP', ',', 'involves', 'tokenization', ',', 'part-of-speech', 'tagging', ',',
'stop', 'words', 'removal', ',', 'stemming', ',', 'lemmatization', '.']
```

```
Stemmed Tokens:
 ['natur', 'languag', 'process', '(', 'nlp', ')', 'subfield', 'artifici', 'intelli
g', '(', 'ai', ')', 'focus', 'interact', 'comput', 'human', 'use', 'natur', 'langua
g', '.', 'involv', 'analysi', ',', 'understand', ',', 'gener', 'human', 'languag',
',', 'enabl', 'machin', 'process', 'comprehend', 'text', 'meaning', 'way', '.', 'nl
p', 'techniqu', 'wide', 'use', 'variou', 'applic', 'sentiment', 'analysi', ',', 'mac
hin', 'translat', ',', 'chatbot', ',', 'inform', 'retriev', '.', 'preprocess', 'esse
nti', 'step', 'nlp', ',', 'involv', 'token', ',', 'part-of-speech', 'tag', ',', 'sto
p', 'word', 'remov', ',', 'stem', ',', 'lemmat', '.']

Lemmatized Tokens:
 ['Natural', 'language', 'processing', '(', 'NLP', ')', 'subfield', 'artificial', 'i
ntelligence', '(', 'AI', ')', 'focus', 'interaction', 'computer', 'human', 'using',
'natural', 'language', '.', 'involves', 'analysis', ',', 'understanding', ',', 'gene
ration', 'human', 'language', ',', 'enabling', 'machine', 'process', 'comprehend',
'text', 'meaningful', 'way', '.', 'NLP', 'technique', 'widely', 'used', 'various',
'application', 'sentiment', 'analysis', ',', 'machine', 'translation', ',', 'chatbot
s', ',', 'information', 'retrieval', '.', 'Preprocessing', 'essential', 'step', 'NL
P', ',', 'involves', 'tokenization', ',', 'part-of-speech', 'tagging', ',', 'stop',
'word', 'removal', ',', 'stemming', ',', 'lemmatization', '.']
```

Part B

In [27]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [28]:
```
# List of documents
documents = [
    "Natural language processing is a subfield of artificial intelligence.",
    "It focuses on the interaction between computers and humans using natural langu
    "NLP techniques are widely used in various applications such as sentiment analy
    "Preprocessing is an essential step in NLP.",
]
```

In [29]:
```
# Create an instance of TfidfVectorizer
vectorizer = TfidfVectorizer()
```

In [30]:
```
# Fit and transform the documents
tfidf_matrix = vectorizer.fit_transform(documents)
```

In [31]:
```
# Get the feature names (terms)
feature_names = vectorizer.get_feature_names_out()
```

In [32]:
```
# Print the TF-IDF representation
for i, doc in enumerate(documents):
    print(f"Document {i+1}:")
    for j, term in enumerate(feature_names):
        tfidf_value = tfidf_matrix[i, j]
        if tfidf_value > 0:
            print(f"{term}: {tfidf_value:.4f}")
    print()
```

```
Document 1:
artificial: 0.3817
intelligence: 0.3817
is: 0.3009
language: 0.3009
natural: 0.3009
of: 0.3817
processing: 0.3817
subfield: 0.3817

Document 2:
and: 0.2392
between: 0.3034
computers: 0.3034
focuses: 0.3034
humans: 0.3034
interaction: 0.3034
it: 0.3034
language: 0.2392
natural: 0.2392
on: 0.3034
the: 0.3034
using: 0.3034

Document 3:
analysis: 0.2686
and: 0.2117
applications: 0.2686
are: 0.2686
as: 0.2686
in: 0.2117
machine: 0.2686
nlp: 0.2117
sentiment: 0.2686
such: 0.2686
techniques: 0.2686
translation: 0.2686
used: 0.2686
various: 0.2686
widely: 0.2686

Document 4:
an: 0.4129
essential: 0.4129
in: 0.3256
is: 0.3256
nlp: 0.3256
preprocessing: 0.4129
step: 0.4129
```