Practical No-7

Date of Conduction : **Date of Checking:**

Text Analytics

1. Extract Sample document and apply following document preprocessing methods:

Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.

2. Create representation of document by calculating Term Frequency and Inverse Document Frequency.

Theory:

Tokenization:

Definition: Tokenization is the process of breaking a text into individual units, known as tokens. Tokens can be words, phrases, symbols, or other meaningful elements.
Example: For the sentence "Natural language processing is interesting," tokenization would result in the tokens: ["Natural", "language", "processing", "is", "interesting"].

POS Tagging (Part-of-Speech Tagging):

Definition: POS tagging involves assigning a part-of-speech category (such as noun, verb, adjective, etc.) to each word in a sentence.
Example: For the sentence "She is reading a book," POS tagging would identify the parts of speech: [("She", "PRP"), ("is", "VBZ"), ("reading", "VBG"), ("a", "DT"), ("book", "NN")].

Stop Words Removal:

Definition: Stop words are common words (e.g., "and," "the," "is") that are often removed from text during preprocessing. These words don't contribute much to the meaning of the text and are often excluded to focus on more meaningful content.
Example: For the sentence "The quick brown fox jumps over the lazy dog," stop words removal might result in: ["quick", "brown", "fox", "jumps", "lazy", "dog"].

Stemming:

Definition: Stemming is the process of reducing words to their root or base form. It involves removing suffixes from words to obtain a common base form.
Example: For the words "running," "runner," and "ran," stemming might produce: ["run", "run", "run"].

Lemmatization:

Definition: Lemmatization is similar to stemming, but it involves reducing words to their base or dictionary form (lemma). It considers the meaning of the word and applies a morphological analysis to obtain the base form.

Example: For the words "running," "runner," and "ran," lemmatization might produce: ["run", "runner", "run"].

These preprocessing techniques are essential in natural language processing and text analytics to standardize and simplify textual data for further analysis. The choice of which methods to apply depends on the specific requirements of the task and the nature of the text data being processed.

Python Code:

```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import PorterStemmer, WordNetLemmatizer
from nltk.probability import FreqDist
from nltk.tag import pos_tag
from nltk.tokenize import sent_tokenize
from sklearn.feature_extraction.text import TfidfVectorizer

nltk.download('punkt')
nltk.download('stopwords')
nltk.download('averaged_perceptron_tagger')
nltk.download('wordnet')

# Sample document
sample_document = """Natural language processing (NLP) is a subfield of
artificial intelligence (AI) that focuses on the interaction between
computers and humans using natural language. It enables computers to
understand, interpret, and generate human-like text. NLP involves various
tasks, such as tokenization, part-of-speech tagging, stop words removal,
stemming, and lemmatization."""

# Tokenization
tokens = word_tokenize(sample_document)

# POS Tagging
pos_tags = pos_tag(tokens)

# Stop Words Removal
stop_words = set(stopwords.words('english'))
filtered_tokens = [word for word in tokens if word.lower() not in
stop_words]

# Stemming
porter_stemmer = PorterStemmer()
stemmed_tokens = [porter_stemmer.stem(word) for word in filtered_tokens]

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

# TF-IDF Representation
documents = [sample_document]
tfidf_vectorizer = TfidfVectorizer()
tfidf_matrix = tfidf_vectorizer.fit_transform(documents)
feature_names = tfidf_vectorizer.get_feature_names_out()
```

```
# Display Results
print("Original Document:\n", sample_document, "\n")
print("Tokenization:\n", tokens, "\n")
print("POS Tagging:\n", pos_tags, "\n")
print("Stop Words Removal:\n", filtered_tokens, "\n")
print("Stemming:\n", stemmed_tokens, "\n")
print("Lemmatization:\n", lemmatized_tokens, "\n")
print("TF-IDF Representation:\n", tfidf_matrix.toarray(), "\n")
print("Feature Names:\n", feature_names)
```

OUTPUT:

"C:\Users\Ram Kumar Solanki\PycharmProjects\pythonProject\venv\Scripts\python.exe" "C:\Users\Ram Kumar Solanki\PycharmProjects\MBA_BFS\main.py"
[nltk_data] Downloading package punkt to C:\Users\Ram Kumar
[nltk_data] Solanki\AppData\Roaming\nltk_data...
[nltk_data] Unzipping tokenizers\punkt.zip.
[nltk_data] Downloading package stopwords to C:\Users\Ram Kumar
[nltk_data] Solanki\AppData\Roaming\nltk_data...
[nltk_data] Unzipping corpora\stopwords.zip.
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data] C:\Users\Ram Kumar
[nltk_data] Solanki\AppData\Roaming\nltk_data...
[nltk_data] Unzipping taggers\averaged_perceptron_tagger.zip.
[nltk_data] Downloading package wordnet to C:\Users\Ram Kumar
[nltk_data] Solanki\AppData\Roaming\nltk_data...
Original Document:
Natural language processing (NLP) is a subfield of artificial intelligence (AI) that focuses on the interaction between computers and humans using natural language. It enables computers to understand, interpret, and generate human-like text. NLP involves various tasks, such as tokenization, part-of-speech tagging, stop words removal, stemming, and lemmatization.

Tokenization:
['Natural', 'language', 'processing', '(', 'NLP', ')', 'is', 'a', 'subfield', 'of', 'artificial', 'intelligence', '(', 'AI', ')', 'that', 'focuses', 'on', 'the', 'interaction', 'between', 'computers', 'and', 'humans', 'using', 'natural', 'language', '.', 'It', 'enables', 'computers', 'to', 'understand', ',', 'interpret', ',', 'and', 'generate', 'human-like', 'text', '.', 'NLP', 'involves', 'various', 'tasks', ',', 'such', 'as', 'tokenization', ',', 'part-of-speech', 'tagging', ',', 'stop', 'words', 'removal', ',', 'stemming', ',', 'and', 'lemmatization', '.']

POS Tagging:
[('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', 'NN'), ('between', 'IN'), ('computers', 'NNS'), ('and', 'CC'), ('humans', 'NNS'), ('using', 'VBG'), ('natural', 'JJ'), ('language', 'NN'), ('.', '.'), ('It', 'PRP'), ('enables', 'VBZ'), ('computers', 'NNS'), ('to', 'TO'), ('understand', 'VB'), (',', ','), ('interpret', 'VB'), (',', ','), ('and', 'CC'), ('generate', 'VB'), ('human-like', 'JJ'), ('text', 'NN'), ('.', '.'), ('NLP', 'NNP'), ('involves', 'VBZ'), ('various', 'JJ'), ('tasks', 'NNS'), (',', ','), ('such', 'JJ'), ('as', 'IN'), ('tokenization', 'NN'), (',', ','), ('part-of-speech',

'JJ'), ('tagging', 'NN'), (',', ','), ('stop', 'VB'), ('words', 'NNS'), ('removal', 'JJ'), (',', ','), ('stemming', 'VBG'), (',', ','), ('and', 'CC'), ('lemmatization', 'NN'), ('.', '.')]

Stop Words Removal:
['Natural', 'language', 'processing', '(', 'NLP', ')', 'subfield', 'artificial', 'intelligence', '(', 'AI', ')', 'focuses', 'interaction', 'computers', 'humans', 'using', 'natural', 'language', '.', 'enables', 'computers', 'understand', ',', 'interpret', ',', 'generate', 'human-like', 'text', '.', 'NLP', 'involves', 'various', 'tasks', ',', 'tokenization', ',', 'part-of-speech', 'tagging', ',', 'stop', 'words', 'removal', ',', 'stemming', ',', 'lemmatization', '.']

Stemming:
['natur', 'languag', 'process', '(', 'nlp', ')', 'subfield', 'artifici', 'intellig', '(', 'ai', ')', 'focus', 'interact', 'comput', 'human', 'use', 'natur', 'languag', '.', 'enabl', 'comput', 'understand', ',', 'interpret', ',', 'gener', 'human-lik', 'text', '.', 'nlp', 'involv', 'variou', 'task', ',', 'token', ',', 'part-of-speech', 'tag', ',', 'stop', 'word', 'remov', ',', 'stem', ',', 'lemmat', '.']

Lemmatization:
['Natural', 'language', 'processing', '(', 'NLP', ')', 'subfield', 'artificial', 'intelligence', '(', 'AI', ')', 'focus', 'interaction', 'computer', 'human', 'using', 'natural', 'language', '.', 'enables', 'computer', 'understand', ',', 'interpret', ',', 'generate', 'human-like', 'text', '.', 'NLP', 'involves', 'various', 'task', ',', 'tokenization', ',', 'part-of-speech', 'tagging', ',', 'stop', 'word', 'removal', ',', 'stemming', ',', 'lemmatization', '.']

TF-IDF Representation:
[[0.12309149 0.36927447 0.12309149 0.12309149 0.12309149 0.24618298
0.12309149 0.12309149 0.12309149 0.12309149 0.12309149 0.12309149
0.12309149 0.12309149 0.12309149 0.12309149 0.12309149 0.24618298
0.12309149 0.12309149 0.24618298 0.24618298 0.24618298 0.12309149
0.12309149 0.12309149 0.12309149 0.12309149 0.12309149 0.12309149
0.12309149 0.12309149 0.12309149 0.12309149 0.12309149 0.12309149
0.12309149 0.12309149 0.12309149 0.12309149 0.12309149 0.12309149
0.12309149]]

Feature Names:
['ai' 'and' 'artificial' 'as' 'between' 'computers' 'enables' 'focuses'
'generate' 'human' 'humans' 'intelligence' 'interaction' 'interpret'
'involves' 'is' 'it' 'language' 'lemmatization' 'like' 'natural' 'nlp'
'of' 'on' 'part' 'processing' 'removal' 'speech' 'stemming' 'stop'
'subfield' 'such' 'tagging' 'tasks' 'text' 'that' 'the' 'to'
'tokenization' 'understand' 'using' 'various' 'words']

Process finished with exit code 0