EXPERIMENT – 3:: Pre-processing of text (Tokenization, Filtration, Script Validation, Stop Word Removal, Lower case conversion, Stemming).

INTRODUCTION ::

Tokenization: Splitting the sentence into words.

Filtration: It the process of removing stop words or any unnecessary data from the sentence

Script Validation :: At the stage of training a model in a data science project, and after training a model, you will want to know the performance of your model on unseen data. We can make changes to some parameters the model used in learning, this is called **hyper-parameter Tuning.**

Stop Word Removal :: Stop word removal is one of the most commonly used preprocessing steps across different NLP applications.The idea is simply removing the words that occur commonly across all the documents in the corpus.

Lower case conversion :: Converting all your data to lowercase helps in the process of preprocessing and in later stages in the NLP application, when you are doing parsing.So, converting the text to its lowercase format is quite easy.

Stemming :: Stemming, in Natural Language Processing (NLP), refers to the process of reducing a word to its word stem that affixes to suffixes and prefixes or the roots.

PROGRAM ::

```
# module -3  ----------   Tokenization   ----------

from nltk.tokenize import sent_tokenize

text="Text preprocessing. is an important step in Natural Language Processing NLP. It involves cleaning and transforming raw data suitable format"

print("--------------------SENTENCE TOKENIZATION-----------------------")

print(sent_tokenize(text))

print("\n")

# ---------- pre-processing (script validation)  ----------

import nltk

from nltk.corpus import stopwords

from nltk.tokenize import word_tokenize

from nltk.stem import PorterStemmer

nltk.download('stopwords')

nltk.download('punkt')

def preprocess_text(text):

    text = text.lower()

    tokens = word_tokenize(text)

    words = set(stopwords.words('english'))

    tokens = [word for word in tokens if word.isalnum() and word not in words]

    stemmer = PorterStemmer()

    tokens = [stemmer.stem(word) for word in tokens]
```

```
    perprocessed_text = ' '.join(tokens)
    return perprocessed_text
if __name__ == '__main__':
    input_text = "Text preprocessing is an important step in Natural Language Processing NLP.It involves cleaning and transforming raw text data into a format suitable for analysis or modeling."
    perprocessed_text = preprocess_text(input_text)
    print("----------------------(script validation) ----------------------------")
    print(perprocessed_text)


# --------------------  Stop Word Removal and Fliteration ------------------
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
text="Text preprocessing is an important step in Natural Language Processing NLP.It involves cleaning and transforming raw text data into a format suitable for analysis or modeling."
stop_words=set(stopwords.words('english'))
word_tokens=word_tokenize(text)
filtered_scentence=[w for w in word_tokens if not w in stop_words]
print("\n")
print("------------------------------STOP WORD REMOVAL & FILTERATION------------------------------")
print(filtered_scentence)
```

```
# ---------- Lower Case Converstion ----------

text="REGISTERED students ARE required to ATTEND THE INTERVIEW WITHOUt FAIL"

text=text.lower()

print("\n")

print("----------LOWER CASE CONVERSTION----------")

print(text)



# ----------  ****Stemming****  ----------

print("\n")

print("----------STEMMING----------")

from nltk.stem import PorterStemmer

ps=PorterStemmer()

text="final year is going to complete within 6 months"

for word in text.split():

   print(ps.stem(word))
```

OUTPUT ::

```
----------SENTENCE TOKENIZATION----------
['Text preprocessing.', 'is an important step in Natural Language
Processing NLP.', 'It involves cleaning and transforming raw data suitable
format']


----------(script validation) ----------
text preprocess import step natur languag process involv clean transform
raw text data format suitabl analysi model


----------STOP WORD REMOVAL &
FILTERATION----------
['Text', 'preprocessing', 'important', 'step', 'Natural', 'Language',
'Processing', 'NLP.It', 'involves', 'cleaning', 'transforming', 'raw',
'text', 'data', 'format', 'suitable', 'analysis', 'modeling', '.']
```

----------LOWER CASE CONVERSTION----------
registered students are required to attend the interview without fail

----------STEMMING----------
final
year
is
go
to
complet
within
6
month