Experiment – 1: Preprocessing text document using NLTK of Python:

a. Stopword elimination

In [1]:

```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
```

In [2]:

```
def Stopword(sentence):
    stopword =set(stopwords.words('english'))
    word_tokens = word_tokenize(sentence)
    filtered_sentence = [w for w in word_tokens ifnot w.lower() in stopword]
    filtered_sentence = []
    for w in word_tokens:
        if w notin stopword:
            filtered_sentence.append(w)
    #print(word_tokens)
    return(filtered_sentence)
```

In [3]:

```
sentence =input()
print()
print()
print(Stopword(sentence))
```

```
    This is a sample sentence showing off the stop words filtration.
```

OUTPUT

```
['This', 'sample', 'sentence', 'showing', 'stop', 'words', 'filtration', '.']
```

b. Stemming

```
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
ss = PorterStemmer()#(language = 'english')
sentence =input()
words = word_tokenize(sentence)
print()
print()
for w in words:
   print(w,":",ss.stem(w))
likes liked likely liking
```

OUTPUT:

likes : like
liked : like
likely : like
liking : like

c. Lemmatization

```
# import these modules
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
import nltk
lemmatizer = WordNetLemmatizer()
sentence =input()
words = word_tokenize(sentence)
for w in words:
        print(w+":"+lemmatizer.lemmatize(w))
```

OUTPUT

```
Rocks books
Rocks:Rocks
books:book
```

d. POS tagging

```
import re
def custom_word_tokenize(text):
    tokens = re.findall(r"\b\w+\b|\S", text)
    return tokens
text = "Laughed my heart out after ages!! Great acting by both the actors rest of the cast was also great! Anushka was so pretty and loved the storyline dialogues and timing! I would go watch the movie second time in theaters which is very rare for me! I loved it as much as Jaatiratnalu Naveen polishetty you gotta do more movies!! Amazing acting!"
tokens = custom_word_tokenize(text)
print(tokens)
```

#OUTPUT:

['Laughed', 'my', 'heart', 'out', 'after', 'ages', '!', '!', 'Great', 'acting', 'by', 'both', 'the', 'actors', 'rest', 'of', 'the', 'cast', 'was', 'also', 'great', '!', 'Anushka', 'was', 'so', 'pretty', 'and', 'loved', 'the', 'storyline', 'dialogues', 'and', 'timing', '!', 'I', 'would', 'go', 'watch', 'the', 'movie', 'second', 'time', 'in', 'theaters', 'which', 'is', 'very', 'rare', 'for', 'me', '!', 'I', 'loved', 'it', 'as', 'much', 'as', 'Jaatiratnalu', ' ', ' ', 'Naveen', 'polishetty', 'you', 'gotta', 'do', 'more', 'movies', '!', '!', 'Amazing', 'acting', '!']

```
import re
def custom_pos_tag(text):
    sentences = re.split(r'(?<=[.!?])\s+', text)
    pos_tags = []
articles=['a','an','the','A','An','The']pronoun=['he','him','his','She','her','This','this','That','that','Them','them','It','it','I','i','me','you','yours','yourself','Your','my','mine','myself']
    adjective
    for sentence in sentences:
        words = sentence.split()
        for word in words:
            if word in articles:
                pos_tags.append((word,"Determiner"))
            elif word in pronoun:
                pos_tags.append((word,'Pronoun'))
            elif word[0].isupper():
```

```
            pos_tags.append((word, "Noun"))
        elif word.endswith("ed"):
            pos_tags.append((word, "Past Tense Verb"))
        elif word.endswith("ing"):
            pos_tags.append((word, "Adjective"))
        else:
            pos_tags.append((word, "Noun"))
    return pos_tags
text = "Laughed my heart out after ages!! Great acting by both the actors rest of the cast was also great! Anushka was so pretty and loved the storyline dialogues and timing! I would go watch the movie second time in theaters which is very rare for me! I loved it as much as Jaatiratnalu Naveen polishetty you gotta do more movies!! Amazing acting!"
tags = custom_pos_tag(text)
for word, tag in tags:
    print(f"{word}: {tag}")
#OUTPUT
        Laughed: Noun
        my: Pronoun
        heart: Noun
        out: Noun
        after: Noun
        ages!!: Noun
        Great: Noun
        acting: Adjective
        by: Noun
        both: Noun
        the: Determiner
        actors: Noun
        rest: Noun
        of: Noun
        the: Determiner
        cast: Noun
        was: Noun
        also: Noun
        great!: Noun
```

Anushka: Noun
was: Noun
so: Noun
pretty: Noun
and: Noun
loved: Past Tense Verb
the: Determiner
storyline: Noun
dialogues: Noun
and: Noun
timing!: Noun
I: Pronoun
would: Noun
go: Noun
watch: Noun
the: Determiner
movie: Noun
second: Noun
time: Noun
in: Noun
theaters: Noun
which: Noun
is: Noun
very: Noun
rare: Noun
for: Noun
me!: Noun