NATURAL LANGUAGE PROCESSING (MDS472C)

NAME: Harish G
REGISTER NUMBER: 2448023
CLASS: 4MDS
DATE: 20th JUNE, 2025
LAB NO.: 1

Program Question 1:

Installing NLTK, NLTK.book and Practice the NLP Environment using exercises 1 and 2.

Draft Plan:
a. Program Description:
Install NLTK and explore its core modules. Practice text analysis using built-in corpora and functions.

b. Program Logic:
Utilize the nltk library, nltk.book module, and explore basic functions.

Program:

```
import nltk
from nltk.book import *

12 / (4 + 1)
26**100
```

Test Cases

INPUT	OUTPUT
12 / (4 + 1)	2.4
26 ** 100	Very Large Number (e.g., 31429306...)

Program Question 2: Text Processing (Basics)

Tasks:
- Define a paragraph as a string.
- Count total and unique words.
- Find word frequency.
- Identify the most and least frequent words.
- Identify the longest word.

Draft Plan:
1. Define a paragraph
2. Tokenize it into words
3. Normalize to lowercase

4. Count total and unique words
5. Use Counter for frequency
6. Use max()/min() for extremes
7. Find the longest word

Program Description:
This program introduces basic NLP preprocessing: tokenizing, normalizing, and analyzing a paragraph of text.

Test Cases

Test Case	Input	Expected Output
Count total and unique words	"I am doing masters in data science at Christ..."	Total: 9, Unique: 9
Frequency, Most/Least	"Design and evaluate a multimodal machine..."	Most: 'machine' (2), Least: 'design' (1)
Longest/Shortest Word	From paragraph	Longest: 'multimodal', Shortest: 'a'

Program Question 3: Regular Expressions

Tasks: Solve Exercises 2.1 and 2.2 from Jurafsky & Martin

Draft Plan:
1. Understand the regex requirements
2. Use Python re module
3. Test patterns
4. Use lookaheads, lookbehinds, and boundaries

Program Description:
Code uses regex to match patterns: repeated words, word structures, and token features.

Program Logic:
- ^, $: string anchors
- \b, \s : word/space boundaries
- *, +, ? : quantifiers
- (), \1 : grouping, back-referencing
- (?=...), (?<=...) : lookahead/lookbehind

Test Cases

Regex Task	Input	Expected Match?/Output
Alphabet check	"Alphabet"	Yes
Ends with 'b'	"abcb"	Yes
'a' surrounded by 'b'	"babab"	Yes
Repeated word	"the the bug"	Yes
Int start, word end	"123 this is the endWord"	Yes
Sentence word starts	"He ran. The dog barked!"	Matches: "He", "The"

NATURAL LANGUAGE PROCESSING (MDS472C)

Exercise Questions 24-28 from Jurafsky

Tasks and Patterns:

Condition	Regex/Logic	Example Output
Words ending in 'ize'	r'ize$'	['finalize', 'maximize']
Words containing 'z'	r'z'	['zebra', 'zombie']
Words with 'pt'	r'pt'	['temptation', 'optical']
Titlecase words	r'[A-Z][a-z]+'	['King', 'Arthur']
Words starting with 'sh' in sent	r'^sh'	['she', 'shells', 'shore']
Words longer than 4 letters	len(word) > 4	['sells', 'shells', 'shore']

Functions:

```
def vocab_size(text):
    return len(set(text))

def percent(word, text):
    return 100 * text.count(word) / len(text)
```

Expected Output:

vocab_size(text1) ~19317
percent('the', text1) ~6.5%