Lab 01 Assignment - EDA & Basics NLP with NLTK

Use the ID_NAME_Assignment 01.ipynb for solving the following tasks.

Tasks:

1. Download **Brown** corpus using NLTK and list all available text files in it. Choose any built-in corpus and use appropriate functions to display the list of text files.

2. Select a novel from **Gutenberg** corpus. Tokenize the text into words, remove the stop-words, generate and display a word cloud to visualize the most frequent words.

3. Select a novel from the gutenberg corpus. Tokenize the text into words, remove stopwords, and apply stemming. Identify the top 15 most frequent words before & after preprocessing, visualize the word frequency distribution using a bar chart.

4. Write a program that extracts and prints the 50 most common bigram (sequences of three consecutive words) from a text, without excluding trigrams that contain stopwords. Use the **Gutenberg** corpus as the text source.

5. Using the Hungarian corpus from NLTK's nltk.corpus.udhr dataset, extract words and identify their vowel sequences. Construct a bigram table representing the co-occurrence of vowel pairs. Write a Python program to process the text, extract vowel bigrams, and display the frequency distribution.