CSE440: Natural Language Processing II

Lab Assignment 1

1. Download Reuters 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, generate and display a word cloud to visualize the most frequent words.

3. Select a novel from the Reuters corpus in NLTK and tokenize the text into words, remove stopwords, and apply stemming. Identify the top 17 most frequent words before and after preprocessing. Visualize the word frequency distribution using a histogram.

4. Write a program to extract and print the 70 most common bigrams (pairs of consecutive words) from a text, ensuring that stop words are included. Use the Reuters corpus as the text source.

5. Using the Spanish corpus from NLTK's nltk.corpus.udhr dataset, extract words and identify their consonant sequences. Construct a trigram table representing the co-occurrence of consonant triplets. Write a Python program to process the text, extract consonant trigrams, and display the frequency distribution.