CSE440: Natural Language Processing II

Lab Assignment 1

1. Write a Python program using NLTK to explore two corpora. First, load the Gutenberg corpus and display all of its available file IDs using appropriate NLTK functions. Then, load the Reuters corpus and display the list of all available categories.

2. Select two text files from the Inaugural corpus. Tokenize both texts into words and generate separate word clouds for each to compare the most frequent words used in both files.

3. Select a text from the Movie Reviews corpus that belongs to the 'neg' (negative) category. Preprocess it by tokenizing into words, removing stopwords, and applying lemmatization. Identify the 30 most frequent words in the cleaned text and visualize their frequencies using a bar plot.

4. Choose a text file from the State of the Union corpus in NLTK. Preprocess the text by converting it to lowercase, removing punctuation, and filtering out stopwords. Identify the top 10 most frequent words from the cleaned text. Then, use trigrams (3-grams) to extract all consecutive word triples. Build a co-occurrence matrix showing how often each pair of the top 10 words appears together within the same trigram in the text.

5. Write a Python program that processes the French text from the udhr corpus in NLTK. For each word in the text, extract the sequence of vowels it contains (ignore consonants). From these vowel sequences, generate all possible bigrams of consecutive vowels. Build a frequency distribution of these vowel bigrams across the entire text and display the result as a sorted list showing each bigram and its count.