MCO2: Corpus Analysis in Multiple Programming Languages

Deadline: November 22, 2024

Goal: Develop a program or script to extract insights from a collection of text files (tweets) using different programming languages (Go, Kotlin, R, and Ruby).

Tasks:

1. Load Data File:
   - Implement methods to load the data file into memory.
   - Consider using libraries or built-in functions for text processing.
   - A dummy data is provided for testing purposes. THe actual data to be processed will be provided later.

2. Corpus Analysis. Across all the input files, perform the compute for the following descriptive statistics:

   - Word count: Calculate the total number of words.
   - Vocabulary size: Calculate the number of unique words.
   - Word Frequency: Calculate the number of times each word appears in the corpus, then sort them based on their frequency.
   - Character Frequency: Calculate the number of times each character appears in the corpus (including letters, digits, and symbols), and then sort them based on frequency.
   - Frequency analysis. Identify the top 20 most frequent words and their counts.
   - Stop Word Identification: Identify 10 common words (e.g., "the," "and," "a") that don't provide significant meaning.

3. Data Visualization. Generate the following figures/images in either (a) file format or (b) a image in a window/browser tab.
   - A word cloud of the top 20 most frequent words
   - A bar chart / histogram showing the total number of posts per month
   - A pie chart showing the distribution of the different symbols found in the corpus. A symbol is any character that is neither a letter nor a digit.

Rubric:

Task 1: Loads Text File successfully	2
Task 2: Corpus Analysis	

Word Count: Accurately calculates the total number of words	3
Vocabulary Size: Accurately calculates the number of unique words	3
Word Frequency: Correctly calculates word frequencies and sorts them	3
Character Frequency: Correctly calculates character frequencies and sorts them	3
Task 3: Data Visualization	
Generates an informative word cloud	2
Generates an informative Histogram on Monthly Posts	2
Generates an informative Pie Chart on Symbol Frequency	2