COIS 4400H

Assignment 1 - Question 3

Winter 2025

Dikshith Reddy Macherla

Student Id : 0789055

Bachelor Of Computer Science, Trent University

COIS 4400H- Data Mining

Professor : Sabine McConnell

3rd February, 2025

Dataset Characteristics

The Gutenberg Collection in NLTK consists of various classical literary texts, including works by Shakespeare, Jane Austen, and Herman Melville. These texts are full-length books that represent different writing styles, vocabularies, and themes.

In processing this dataset, I removed stopwords (e.g., "the", "is", "and") and punctuation to focus only on meaningful terms. The 10 most frequent words per document were extracted, revealing insights into each book’s key themes. For instance, Moby Dick includes frequent terms like "whale" and "sea", while Hamlet emphasizes words like "king" and "thou".

A key observation is that each book’s most common words reflect its genre and author’s writing style. This dataset is particularly useful for natural language processing (NLP) tasks like text classification, sentiment analysis, and topic modeling. However, the dataset is large and unstructured, requiring careful preprocessing before analysis.