REPORT ON ASSIGNMENT 7

For Part 2 of the assignment, we focused on the corpus of M.K. Gandhi. We implemented the A6 dictionary approach, where, upon inserting each line, we manage a dictionary named "gd." This dictionary essentially records every word in every book along with its corresponding count. Additionally, we maintain a vector of dictionaries, denoted as "vd."

During the insertion of each line, we capture the paragraph number. If there is a change in the paragraph number, we create a new dictionary to store the words and their counts for that specific paragraph. Subsequently, we add this paragraph dictionary to the vector of dictionaries.

For scoring words and paragraphs, we follow a similar approach as in Part 1. However, in this case, I exclude the scoring of common words in the query, such as "I," "me," "my," "myself," "we," "our," "ours," "ourselves," "you," "your," "yours," "yourself," "yourselves," "he," "him," "his," "himself," "she," "her," "hers," "herself," "it," "its," etc.

This exclusion is impactful, especially in queries like "What were the views of Mahatma Gandhi on partition?" After removing common words, the relevant words left in the query are "views," "Mahatma Gandhi," and "partition."

We continue to rank paragraphs as in Part 1, returning the top 5 paragraphs. While considering alternative methods for scoring words, we found that the Part 1 method remains effective. For a slight improvement, we have also eliminated these common words from the CSV file. We have determined the value of k as 5 based on extensive query analysis