Pre-processing research

# Introduction

Pre-processing is a common method to optimize text searching. However, it is an expensive process. Thus, how to decrease the cost is a huge issues to be resolved. In this paper, it provides a method to minimize the cost of text pre-processing.

# Method

One solution is that the pre-processing results are saved as a document, and reload to searching data structures directly.

We save the pre-processing results generated when we first read the file into a pre-processed file and then save it to the local hard disk. Record the result of this operation to a log file. This log file can use a binary tree or map to facilitate the structure of the search. When the same file is opened next time, the search is first performed in the log file. If it is found that there is a record of pre-processed file, the pre-processed file is directly called locally and loaded into the data structure for searching. By this way, when the text file is opened again, its content does not need to be re-read and processed so that it saves pre-processing time.

Another method is to use BufferReader instead of Scanner when reading the text file in Java, because BufferedReader is a bit faster as compared to scanner. Scanner does parsing of input data and BufferedReader simply reads sequence of characters[1].

# Comparison

Compared to other methods, such as using a faster read file method, such as buffer reader instead of scanner, or other pre-processing methods. Reload pre-processing does not re-read the text file, but searches in the log file and calls it. This process expenses the fastest time O(1) as expected. Re-read and process the file, even if it uses more efficient methods, the time required is at least O(n). At the same time, the log file's memory usage is much smaller than the pre-processing process.

# Conclusion

In summary, the method of saving the pre-processing file and calling it again is better than pre-processing at each time. It is not only faster, but also saves memory usage.

# Reference

[1] geeksforgeeks, “Difference between Scanner and BufferReader Class in Java

”, geeksforgeeks, 2017. [Online]. Available: https://www.geeksforgeeks.org/difference-between-scanner-and-bufferreader-class-in-java/. [Access October 14, 2018].