Virtual Thread Text Simplifier

Github - Eric Murray G0042390

Description

This is a Java-based program designed to simplify input text files by replacing words with their closest semantic match. It utilizes embeddings to calculate similarities and uses virtual threads to enhance I/O performance. The program ensures efficiency and scalability while adhering to Java features like structured concurrency and multithreading.

Features

- User-Friendly Menu – Intuitive CLI interface with options to specify file paths.
- Virtual Threads – Concurrent loading of embeddings & google words using virutal threads.
- Text Simplification – Replaces non-google words with the most semantically similar word in the Google list.
- Error Handling – Validates file paths and handles exceptions for robustness
- Cosine Similarity – Calculates the cosine similarity between words for accurate replacements.
- Progress Indicator – Real-time progress indication for feedback during execution.

Functionality

- File Handling – Specifying paths for embeddings file, Google-1000 file, input & output file locations
- Processes the input file line-by-line, handling punctuation using regex to split.
- Outputs a simplified text file with replacements based on semantic similarities

Conclusion

The program demonstrates the use of Virtual Threads in Java, efficient file handling and semantic text processing, offering a modern and practical solution for large-scale text simplification tasks.

References

- https://GeeksForGeeks.org - For Java Virtual Thread & Concurrency
- https://docs.oracle.com/en/java - For Java documentation
- https://chatgpt.com - For help debugging and Big-O notation
- https://baeldung.com - For multi-threading & CompletableFuture