Tower Research - Take Home Assignment

Pavithran Chelliahpillai

March 2025

1 Overview of the Program

The program is a 4-stage program as outlined below. In each section, the specific motivation of each step will be explained.

File Processing → Summarization → Analysis → Final Writing

2 File Processing

All books and files are uploaded into the "books" directory. The program utilizes a variety of existing Python libraries to sweep through the books and generate ".txt" files with the same content under the "converted_books" directory. This allows uniform processing in the future and removes clutter from the more complicated analysis stages of the program.

3 Summarization

The summarization makes use of a technique inspired by the Lagrange Interpolation Method. Essentially, in a good summary, we want to minimize the amount of text while retaining the maximum amount of key information. We can translate this problem to; what is the function of the least degree that fits through all key points - a very fundamental calculus problem.

However, instead of working with derivatives which would hardly make contextual sense, we employ Newton's method of constant differences.

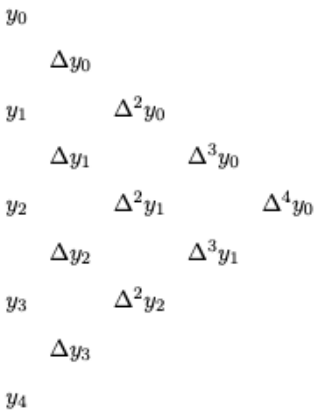

Figure 1: Newton's Interpolation Method

The idea is to initially section of the text off arbitrarily into paragraphs, recursively summarizing (the equivalent in this mathematical method of finding the "difference") until the differences converge to a single point; the number of steps being the minimum degree.

This method also prevents hallucinations having a large impact as they are easily filtered out through the process, and ensure key ideas are kept.

When sectioning off the initial text, the program also populates a database filled with over 100 quotes per text. Each book has its own ".pkl" file populated with the relevant strings for quick loading and storage of raw text quotes. The main search function uses a ".faiss" file that stores vector embeddings of quotes generated by BERT. By applying simple linear algebra techniques, the FAISS indices allow fast similarity searches by calculating projections over a 768-dimensional embedding space.

The key features of this database are: semantic searching (ability to find relevant quotes), book isolation (no hallucinations across books), efficient query (O(1) through FAISS), and preprocessing of data.

3.1 Analysis and Final Writing

Since the heavy-loading of computation and organization of data is done, the final 2 steps are relatively simple. The analysis stage takes the summary and connects it to the theme of the essay individually for each book. The program enters the final writing stage, in which BERT is utilized to find 3 relevant quotes, and Gemini is prompted to write the essay. The prompts are intelligently crafted with special weightings and penalties to ensure a high standard of writing.