Codebase Overview

Code and report by Benjamin Smith

Memory

This codebase uses a modified Trie design. All data is held in arrays, and relationships between parent and child are calculated, not saved, like a traditional tree. This allows us to compare children of the same layer more efficiently and also saves space by not keeping the relationship in memory. This does mean calculating the relationship is more expensive once the trie is fully made; however, in the document description, the relationships between the grams are not emphasized. That document also outlines that memory and computational speed are key. This approach to the Trie also requires substantial continuous memory in the heap. Below is a table outlining the general memory requirements of the Trie.

Name	Data type	Quantity	Total memory
gram_1	uint64_t	26	208 B
gram_2	uint64_t	676 (26^2)	5.408 KB
gram_3	uint64_t	17576 (26^3)	140.608 KB
gram_4	uint64_t	456976 (26^4)	3.655808 MB

This Trie is stored in the heap because of its size. It uses uint64_t as counters because a 100GB file could cause a count of one hundred billion 1 grams; uint64_t is the smallest standard count that can hold this.

The main code has two memory parts: the window and the buffer. The buffer is 1 MB and is used to hold large chunks of the text file, reducing system calls. The window is a string that holds up to four letters that are used in the gram calculations. The buffer reads large chunks from the file, and the window extracts letters from the buffer. Only after letters are moved into the window are they processed by the Trie. Important note: just because the letters are in the window does not mean they are processed; only the first letter in the updateWindow method is updated.

Time Complexity

Name	Time Complexity	Estimated Cycles
updateWindow()	O(1)	5
AddWindowGrams()	O(1)	42
Main Read Loop	O(N)	48 per byte
Trie Value Lookup	O(1)	1-7 based on layer
Generation Comparisons	O(N)	7 per Array index
Gram Key From Index: Gram 1	O(1)	32
Gram Key From Index: Gram 2	O(1)	62
Gram Key From Index: Gram 3	O(1)	124
Gram Key From Index: Gram 4	O(1)	189

Cycles were estimated by ChatGPT for an ARM Cortex-M4/M7

This program excels in reading and lookup efficiency, but sacrifices gram key calculations. This program also allows for comparing grams of the same generation very efficiently.