Tecnológico de Monterrey

Reporte Final: Implementación de Suffix Array, BWT y FM-Index

Estefania Antonio Villaseca A01736897

28 de Septiembre de 2024

Modelación de sistemas multiagentes con gráficas computacionales

Prof. Luciano García Bañuelos

Durante esta actividad, tuve la oportunidad de implementar y explorar diferentes estructuras de datos avanzadas, como el Suffix Array, la Burrows-Wheeler Transform (BWT) y el FM-Index, utilizando la tecnología de Grandes Modelos de Lenguaje (LLMs), representada en este caso por Chat GPT. El objetivo de la actividad era aplicar los conceptos vistos en clase para la búsqueda eficiente de patrones dentro de cadenas de texto y documentar todo el proceso en un repositorio Git.

Implementación del Suffix Array

El primer paso fue la implementación del Suffix Array. Esta estructura de datos permite ordenar todos los sufijos de una cadena de texto de manera lexicográfica, lo cual es esencial para la construcción de la BWT. Al analizar la complejidad temporal de este algoritmo, se determinó que la ordenación de los sufijos tiene un costo de O(n log n), siendo n el tamaño de la cadena. Especialmente, el Suffix Array requiere O(n) espacio, ya que se almacenan los índices de todos los sufijos.

Generación de la Burrows-Wheeler Transform

Una vez construido el Suffix Array, generé la Burrows-Wheeler Transform (BWT), que es crucial para comprimir la cadena de texto y permitir búsquedas eficientes. La BWT fue construida en O(n) tiempo, manteniendo un costo espacial también de O(n).

Construcción de los Arreglos C y Occur

Para implementar el FM-Index, construí los arreglos C y Occur. El arreglo C mapea cada carácter a su posición inicial en la BWT ordenada, y la tabla Occur cuenta las frecuencias acumuladas de cada carácter en la BWT. Estos pasos son fundamentales para la búsqueda de patrones utilizando el método backward search.

Análisis de una Cadena de Genoma

Como parte del proyecto, realicé un análisis sobre una cadena de genoma, donde busqué la subcadena “TCGA” para verificar su presencia en la secuencia genómica. Este análisis es particularmente relevante en el campo de la bioinformática, ya que permite identificar secuencias específicas dentro de datos biológicos, lo que puede ser crucial para investigaciones genéticas y estudios sobre enfermedades.

Implementación de la Búsqueda con FM-Index

La búsqueda de patrones en la cadena comprimida mediante el FM-Index fue uno de los aspectos más interesantes del proyecto. La implementación de la función backwardSearch me permitió realizar búsquedas eficientes sin necesidad de descomprimir la cadena, lo cual optimiza el proceso de búsqueda considerablemente.

Reflexión Personal

Este proyecto me permitió profundizar en algoritmos avanzados y aplicar estructuras de datos complejas de manera práctica. El uso de Chat GPT fue enriquecedor, ya que facilitó la iteración rápida y la resolución eficiente de problemas. Sin embargo, es esencial complementar esta tecnología con un sólido conocimiento teórico para garantizar la precisión de las soluciones.

El análisis de una cadena de genoma y la búsqueda de la subcadena “TCGA” demostraron la versatilidad de estos algoritmos en bioinformática, ampliando mi perspectiva sobre sus aplicaciones reales en la ciencia. La actividad fue un desafío que me permitió aplicar conocimientos y explorar nuevas tecnologías. Estoy satisfecha con el resultado y motivada a seguir investigando el potencial de la inteligencia artificial en la biología y la salud.