ACH2023 - Algoritmos e Estruturas de Dados I
Exercício Programa - 2.o semestre de 2024
Indexador e Buscador de Palavras

O programa carrega o conteúdo de um arquivo texto, armazena-o em memória e também indexa suas palavras. Na sequência, o programa fica à disposição do(a) usuário(a) para que ele(a) possa realizar buscas. Caso a palavra procurada exista no texto, o programa deve exibir as linhas do texto nas quais a palavra ocorre.

Alunos

Victor Yodono - 13829040
Kevin Rodrigues Nunes - 15676030
João Pedro Nunes Aquino - 15463492

Visão geral

O programa oferece duas estratégias distintas para armazenar e buscar as palavras indexadas:

Árvore AVL - Estrutura balanceada com busca eficiente.
Listas - Construída a partir de uma lista ligada dinâmica, convertida posteriormente para uma lista sequencial ordenada com otimização para busca binária.

Estratégias de Indexação

Árvore AVL

A árvore AVL é uma estrutura de dados balanceada que mantém a eficiência das operações de busca e inserção. Sua principal característica é o balanceamento automático após cada inserção, garantindo que a profundidade da árvore permaneça próxima a log n.

Funcionamento

Durante a leitura do arquivo, cada palavra é buscada na árvore. Se a palavra já existir, suas ocorrências são atualizadas. Caso contrário, a palavra é inserida na árvore, criando uma nova lista de ocorrências.

Vantagens

Complexidade de busca e inserção: O(log n). Estrutura balanceada naturalmente, sem necessidade de otimizações adicionais.

Listas

A estratégia de lista é implementada em duas fases:

Fase Inicial - Lista Ligada:

Durante a leitura do arquivo, as palavras são armazenadas em uma lista ligada dinâmica. Como o tamanho total da lista não é conhecido antecipadamente, a lista ligada permite inserções flexíveis sem realocação de memória. A busca na lista ligada tem complexidade O(n) no pior caso, pois é necessário percorrer os elementos sequencialmente.

Fase de Otimização - Conversão para Lista Sequencial Ordenada:

Após a leitura completa do arquivo, a lista ligada é convertida em uma lista sequencial ordenada. A lista sequencial é armazenada em um vetor dinâmico e ordenada alfabeticamente. A busca na lista sequencial utiliza busca binária, com complexidade O(log n).

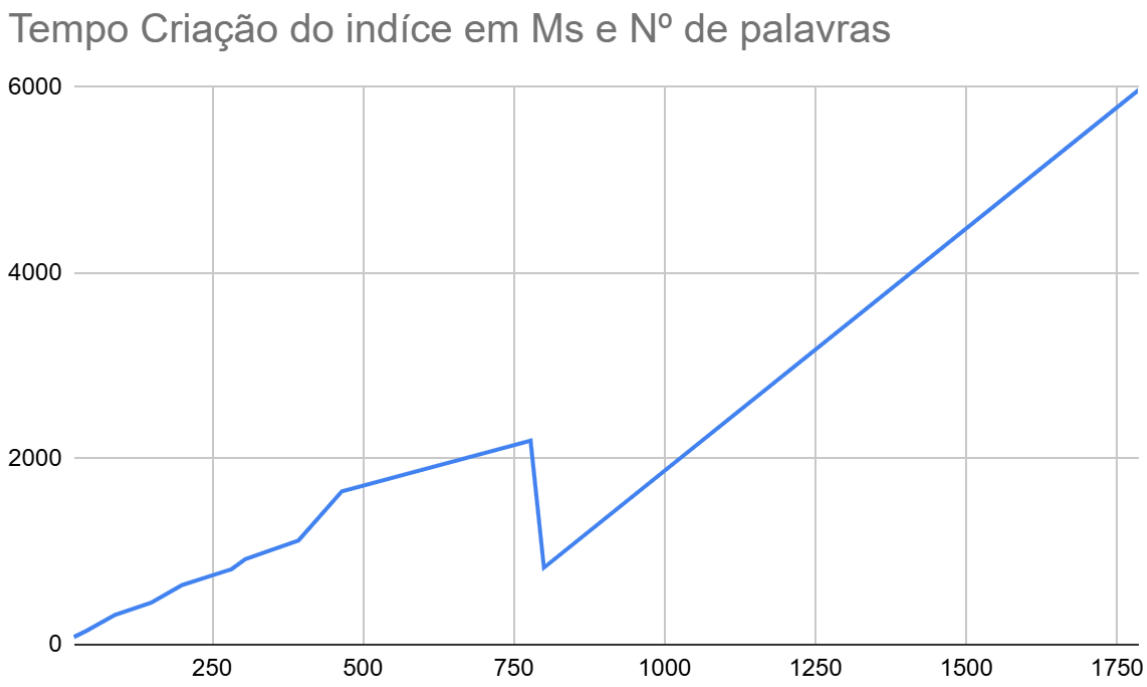
Vantagens

A lista ligada facilita a construção inicial com inserções dinâmicas. A conversão para lista sequencial ordenada otimiza o tempo de busca para O(log n).

Comparação das estratégias

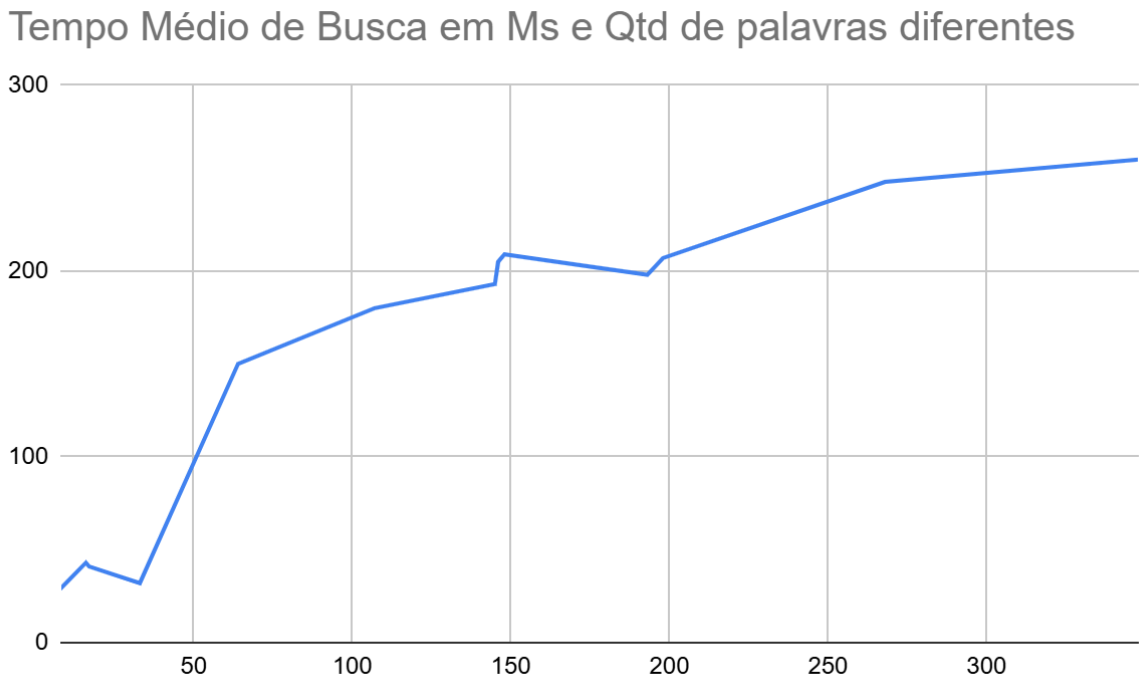
Critério	Árvore AVL	Lista
Inserção	O(log n)	O(n) (durante a lista ligada)
Busca Inicial	O(log n)	O(n) (lista ligada)
Busca após otimização	Não se aplica (já eficiente)	O(log n) (busca binária)
Uso de Memória	Dinâmico, balanceado	Dinâmico → Vetor ordenado
Flexibilidade	Balanceamento automático	Precisa de otimização manual

Gráficos do tempo de execução



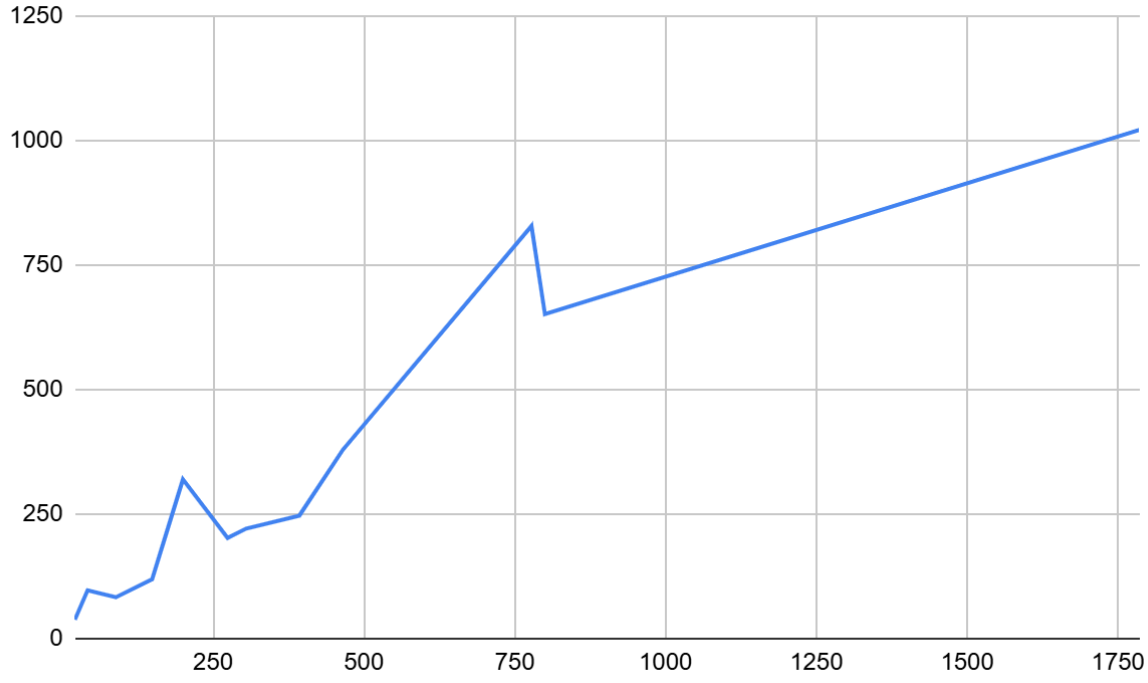
Com o eixo X sendo o número de palavras no arquivo original, temos um gráfico quase linear, considerando que os valores de n são relativamente baixos e então uma parte considerável do tempo pode ser ocupada pela inicialização do programa, e Y representa os milissegundos do processo de criação da lista sequencial

ordenada.

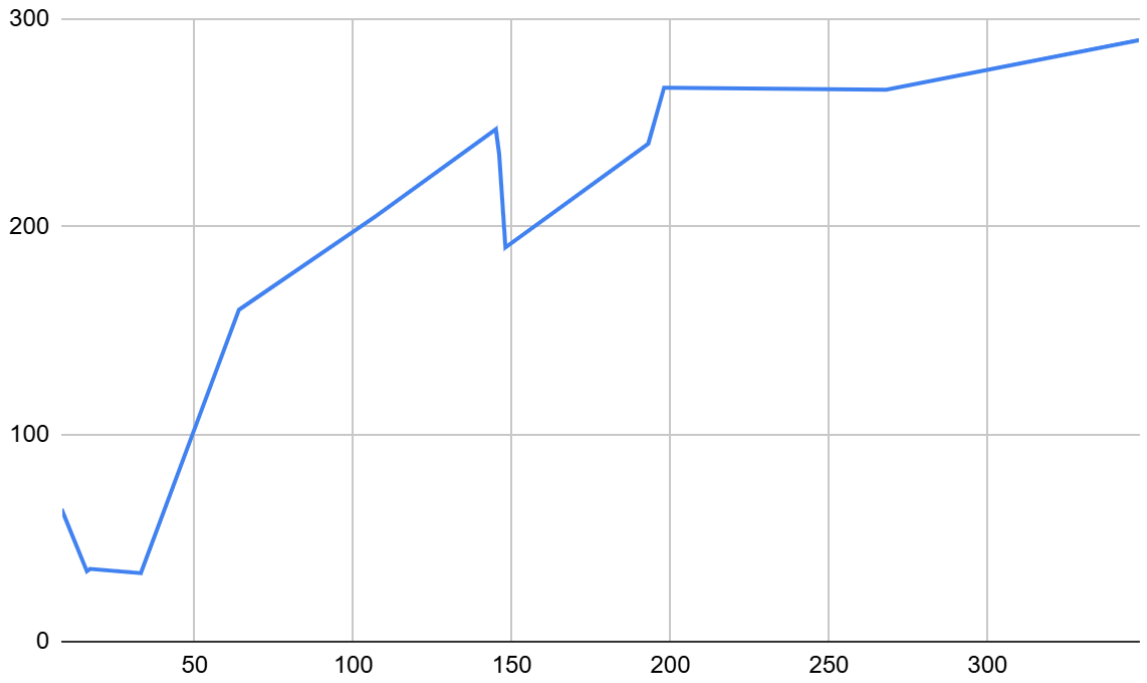


Dessa vez o eixo X representa a quantidade/variedade de palavras diferentes no texto, e o eixo Y novamente é o tempo, só que dessa vez de busca e é possível ver que o crescimento do tempo vai se achatando, não é tão alto.

Árvore AVL



Com o eixo X sendo o número de palavras no arquivo original, temos um gráfico quase linear, considerando que os valores de n são relativamente baixos, e assim existem desvios na inicialização e Y representa os milissegundos do processo de criação da lista sequencial ordenada



Dessa vez o eixo X representa a quantidade/variedade de palavras diferentes no texto, e o eixo Y novamente é o tempo, só que dessa vez de busca.