Leitura de arquivo para contagem de palavras

Davi dos Santos Mattos
Daniel Li Vam Man
Pedro André Alves Chaves

Relatório Parcial
Programação Concorrente (ICP-361) - 2025/2

1. Descrição do problema geral

O problema escolhido é a contagem de palavras em um *arquivo.txt* . Uma tafera muito comum em aplicações de análise de dados, mineração de texto, recuperação da informação e processamento de linguagem neural.

Uma palavra consiste numa sequência de caracteres do alfabeto e a contagem ignora se a letra está em maiúsculo ou minúsculo. As pontuações e espaços indicam o começo e fim de uma palavra.

- **Entrada**: Arquivo de Texto (.txt) qualquer.
- **Saída**: Número de palavras encontradas no arquivo

O problema se beneficia da concorrência quando temos que lidar com arquivos de texto muito grandes, pois podemos dividir o arquivo em partes e cada thread executará sua tarefa de contagem de forma paralela, reduzindo dessa forma o tempo de processamento.

2. Projeto da solução concorrente

Dentre as estratégias que podemos usar temos

- Divisão do arquivo por linhas e usar várias threads para processar as linhas simultaneamente
- Divisão do arquivo por blocos de tamanhos fixos (bytes)
- Modelo produtor-consumidor, uma thread lê o arquivo e envia partes para uma fila compartilhada

Dentre as estratégias listadas acima optamos por seguir com o modelo produtor-consumidor, pois para arguivos muitos grande, se torna muito custoso carregar todo arquivo na memória, e como queremos reduzir o custo e aumentar o desempenho, o modelo se colocar como a melhor escolha.

3. Casos de teste de corretude e desempenho

Para podermos avaliar a solução vamos predefinir duas coisas, que são:

1. **Tamanho de arquivos**: Pequeno (KB), Médio (MB) e Grande (GB)
2. **Número de Threads**: {2, 4, 8, 16}

Para avaliar a **Corretude**:

- Testes com arquivos vazios, arquivos com uma palavra e arquivos somente com caracteres especiais e pontuação.
- Comparar o resultado da versão sequêncial com a versão concorrente para os casos pequeno(KB) e médio (MB)
- Comparar se a quantidade de palavras total coincide em ambos os casos

Para avaliar o **Desempenho**:

1. Rodar a solução para arquivos com tamanhos diferentes
2. Executar com diferentes números de threads
3. Medir o tempo de execução para cada execução

Vamos focar em utilizar somente arquivos de médio e grande porte, visto que arquivos pequenos podem não demonstrar ganhos significativos devido ao comportamento das threads.

4. Referências bibliográficas

References

[1] Rossetto, S. **Slides de Aula**.

[2] Maratona de programação paralela (Mackenzie). Disponível em: http://lspd.mackenzie.br/marathon/old.html.

[3] P. Pacheco, **An Introduction to Parallel Programming**, Morgan Kaufmann, 2011.

[4] Martin Porter's Stemming algorithm as a C library. Disponível em: https://github.com/wooorm/stmr.c