Universidade Federal de Ouro Preto – UFOP
Instituto de Ciências Exatas e Biológicas – ICEB
Departamento de Computação – DECOM
Estrutura de Dados II – BCC203
Prof. Guilherme Tavares de Assis – 1º/2016

Trabalho Prático III – Casamento Exato de Cadeias

Grupo de 4 alunos - Valor: 10,0 pontos - Data de entrega: 26/07/2016

O objetivo deste trabalho prático consiste em um estudo da complexidade de desempenho dos seguintes métodos de casamento exato de cadeias apresentados em sala de aula: (1) força bruta, (2) *Boyer-Moore* (BM) com a heurística *ocorrência* ou com a heurística *casamento* para determinação do tamanho do deslocamento, (3) *Boyer-Moore-Horspool* (BMH) ou *Boyer-Moore-Horspool-Sunday* (BMHS) e (4) *Shift-And* Exato. Como é de conhecimento, um método de casamento exato de cadeias tem como objetivo encontrar todas as ocorrências de um padrão **P** (cadeia de **m** caracteres) em um texto **T** (cadeia de **n** caracteres).

A 1ª fase deste trabalho corresponde à implementação em C dos métodos mencionados, considerando arquivos textos quaisquer (textos **T**) e memória interna disponível para armazenar as estruturas de dados auxiliares (pré-processamento do padrão **P**), quando for o caso.

A 2ª fase deste trabalho corresponde à análise experimental da complexidade de desempenho dos métodos mencionados, considerando as etapas de pré-processamento do padrão **P**, quando for o caso, e da própria busca das ocorrências do padrão **P** no texto **T**, por meio dos seguintes quesitos:

- número de comparações entre os caracteres do texto **T** e do padrão **P**, quando for o caso;
- número de deslocamentos realizados, quando for o caso, na varredura de todo o texto **T**;
- tempo de execução (tempo do término de execução menos o tempo do início de execução).

Para a 2ª fase, testes devem ser realizados considerando arquivos textos de tamanhos variados (textos **T**): pequeno ($10.000 \leq \mathbf{n} \leq 15.000$), médio ($250.000 \leq \mathbf{n} \leq 300.000$) e grande ($2.500.000 \leq \mathbf{n} \leq 3.000.000$). Para cada texto **T**, deve ocorrer a busca das ocorrências de 3 padrões **P** de tamanhos variados: pequeno ($5 \leq \mathbf{m} \leq 10$), médio ($40 \leq \mathbf{m} \leq 50$) e grande ($220 \leq \mathbf{m} \leq 250$). Assim, para cada experimento, considerando os mesmos parâmetros (método de casamento de cadeias, texto **T** e padrão **P**), é possível gerar o valor obtido por cada quesito a ser considerado no processo de análise experimental. Os valores de **m** e **n** devem ser escolhidos pelos componentes do grupo, dentro dos intervalos definidos.

Independente dos testes a serem efetuados, o programa deve ser implementado de tal forma que seja possível executá-lo, livremente, a partir da seguinte linha de comando no console:

casamento <método> <texto> <padrão> [-P]

onde:

- <método> representa o método de casamento exato de cadeias a ser executado, podendo ser um número inteiro de 1 a 4, de acordo com a ordem dos métodos mencionados;
- <texto> representa o nome físico do arquivo texto (texto **T**) a ser utilizado pelo método;
- <padrão> representa o padrão **P** a ser procurado no texto **T**;
- [-P] representa um argumento opcional que deve ser colocado quando se deseja que os valores obtidos referentes aos quesitos de análise (número de comparações, número de deslocamentos, tempo de execução) sejam apresentados na tela.

Uma execução do comando "casamento" deve retornar as ocorrências do padrão **P** no texto **T**, ou seja, as posições numéricas no texto **T** em que o padrão **P** aparece. Para tanto, é importante considerar que os caracteres do texto **T** encontram-se em posições consecutivas a partir da posição 0.

Para fins de avaliação deste trabalho, deve ser entregue:

1. Código-fonte do programa, bem indentado e comentado.
2. Relatório contendo os seguintes pontos:
   - Introdução: especificação do problema a ser resolvido; especificação dos métodos implementados para resolução do problema; descrição dos testes realizados.
   - Para cada método: análise dos quesitos especificados considerando os testes realizados.
   - Conclusão: análise comparativa de desempenho entre os métodos.

Observações Importantes:

- Toda mensagem de orientação e de erro deve ser devidamente tratada.
- O código-fonte do programa deve estar bem indentado e comentado.
- Trabalhos copiados terão suas notas divididas pelo número de cópias.
- Penalização por atraso: 5 pontos a cada aula.
- Para não ocorrer perda na pontuação do trabalho por atraso:
  - o código-fonte do programa deve ser encaminhado para o e-mail gtassis@gmail.com (arquivo "zipado") até às 10h do dia 26/07;
  - o código-fonte e o relatório solicitado devem ser entregues, impressos, na aula do dia 26/07;
  - o programa deve ser apresentado pelo grupo ao professor, diretamente, na aula do dia 26/07.