Compressão de Tweets

A cada dia novos usuários se juntam ao Twitter fazendo o número diário de posts aumentar continuamente. Com isso, o Twitter precisa agora diminuir o espaço de armazenamento ocupado por seus twets. Para isso, seu grupo deverá implementar uma solução capaz de comprimir os tweets que foram publicados. A fim de obter a melhor solução, seu grupo deverá analisar o desempenho e a taxa de compressão dos algoritmos de compressão vistos em aula considerando as seguintes métricas: taxa de compressão, armazenamento físico (tamanho do arquivo em disco) e tempo gasto para compressão (tempo de processamento e não o tempo de relógio).

Você deverá utilizar um conjunto real de *tweets* públicos para os experimentos. No link https://archive.org/details/twitter_cikm_2010, há mais de 5 milhões de *tweets* coletados entre Setembro de 2009 a Janeiro de 2010 e considerar a mesma estrutura de dados dos Trabalho 1 e 2:

- USERID (i.e., identificador do usuário que postou o *tweet* – tipo inteiro)
- TWEETID (i.e., identificador do *tweet* – tipo inteiro e **chave para ordenação**)
- TWEET (i.e., texto do *tweet* – tipo char, máximo de 140 caracteres)
- DATE (i.e., texto contendo a data do tweet – tipo char, formato AAAA-MM-DD HH:MM:SS)

1 – Análise dos Algoritmos de Huffman, LZ77, LZ78 e LZW para compressão

Você deverá avaliar o desempenho e a taxa de compressão ao comprimir um tweet utilizando os algoritmos de Huffman, LZ77, LZ78 e LZW

Você ainda deverá implementar funções/métodos para importar os conjuntos de elementos aleatórios. Estes métodos/funções devem ser chamados uma vez para cada um dos N elementos a serem ordenados.

Análise:

Os algoritmos deverão ser aplicados a entradas com diferentes tamanhos (parâmetro N). Para cada valor de N, você deve gerar 5 (cinco) conjuntos de elementos diferentes, utilizando sementes diferentes para o gerador de números aleatórios. Você pode gerar um número aleatório com valores entre 1 e o número de linhas do seu arquivo de dados e importar o dado correspondente ao número da linha gerado. Experimente, no mínimo, com valores de N = 1000, 5000, 10000, 50000, 100000, 500000 e 1000000. Os algoritmos serão avaliados comparando os valores médios das 5 execuções para cada valor de N testado.

O seu programa principal deve ser receber um arquivo de entrada (`entrada.txt`) com o seguinte formato:

7 → número de valores de N que se seguem, um por linha
1000

5000
10000
50000
100000
500000
1000000

Para cada valor de N, lido do arquivo de entrada `entrada.txt`:

- Gera cada um dos conjuntos de elementos e salva apenas esse conjunto em disco.
- Comprime os tweets e contabiliza estatísticas de desempenho para o algoritmo analisado.
- Salva o conjunto de tweets comprimidos em disco.
- Armazena estatísticas de desempenho em arquivo de saída (`saida.txt`)

Ao final, basta processar os arquivos de saída referentes a cada uma das sementes, calculando as médias de cada estatística, para cada valor de N e para cada algoritmo considerado.

Resultados:

Apresente gráficos e tabelas para as três métricas pedidas, taxa de compressão, armazenamento em disco e tempo de execução (tempo de processamento), comparando o desempenho dos algoritmos e diferentes valores de N. Discuta seus resultados. Quais são os compromissos de desempenho observados?

Considerações

1) Todo código fonte deve ser documentado. A documentação inclui, dentre outros, a documentação de procedimentos, de funções, de variáveis, de partes do código fonte que realizam tarefas específicas. Ou seja, o código fonte deve ser documentado tanto em nível de rotinas quanto em nível de variáveis e blocos funcionais.
2) A interface pode ser feita em modo texto (terminal) ou modo gráfico e deve ser funcional.
3) A implementação deve ser realizada usando a linguagem de programação C, C++ ou Java.

Entrega

O grupo deverá ser formado por 4 alunos e as responsabilidades de cada aluno deve ser documentada e registrada. O prazo final para entrega é dia **04/12.** Deverá ser agendada uma data para entrega e apresentação do trabalho para a professora.

Deve ser entregue os códigos implementados e um relatório com os seguintes itens:
1) Descrição das atividades realizadas por cada membro do grupo
2) Análises da Parte 1

Critérios de avaliação

Você não fechará o trabalho só tendo um "sistema que funciona". O sistema deve funcionar bem e o quão bem ele funcionar será refletido na sua nota. A nota poderá ser comparativa, então se esforce para ter uma solução melhor que a dos outros colegas. O objetivo do trabalho é testar a sua capacidade de fazer boas escolhas (e boas adaptações) de estruturas para resolver problemas. **Então usar classes prontas ou métodos prontos não são permitidos aqui.** Você poderá, se quiser, comparar sua solução com outras prontas. Mas deve perseguir o seu melhor sem usar recursos de terceiros.

Os membros da equipe serão avaliados pelo produto final do trabalho e pelos resultados individuais alcançados. Assim, numa mesma equipe, um membro pode ficar com nota 90 e outro com nota 50, por exemplo. Dentre os pontos que serão avaliados, estão:

- Execução do programa (caso o programa não funcione, a nota será zero)
- Código documentado e boa prática de programação (o mínimo necessário de variáveis globais, variáveis e funções com nomes de fácil compreensão, soluções elegantes de programação, código bem modularizado, etc)
- Testes: procure fazer testes relevantes como, por exemplo, aqueles que verificam casos extremos e casos de exceções
- Relatório bem redigido

Note que o grande desafio deste trabalho está na avaliação dos vários algoritmos nos diferentes cenários, e não na implementação de código. Logo, na divisão de pontos, a documentação receberá, no mínimo, 50% dos pontos totais.
Uma boa documentação deverá apresentar não somente resultados brutos mas também uma discussão dos mesmos, levando a conclusões sobre a superioridade de um ou outro algoritmo, para cada métrica avaliada.