INF 112 – Programação II

Trabalho prático II: Agregação por chave!

Seu objetivo neste trabalho será mostrar que você entendeu bem os conceitos de ordenação, arquivos e ordenação externa.

Uma tarefa muito comum em banco de dados é a agregação por chave. Considere, por exemplo, a tabela de valores abaixo:

categoria	salario	idade	vontade
A	10	30	muita
B	20	32	pouca
C	30	33	pouca
D	10	23	muita
A	5	34	pouca
A	40	54	muita
B	50	43	muita

Uma pergunta que pode surgir nesse contexto é: qual é a média dos valores dos salários de acordo com cada categoria? Nesse caso, o resultado é apresentado abaixo.

categoria	salario
A	18.333333
B	35.000000
C	30.000000
D	10.000000

No caso geral, a essa tarefa é dada o nome de **agregação**, a qual pode ser resolvida de forma simples quando a tabela em questão é pequena. Infelizmente, nem sempre a tabela é pequena e é com esse caso que vamos trabalhar.

Sua entrada será um arquivo de texto, onde cada linha do arquivo representa uma linha da tabela. Esse arquivo poderá ter várias colunas, sendo que em cada linha, o valor de cada coluna será separado dos demais por vírgulas. A primeira linha do arquivo indica o nome de cada coluna. O texto abaixo representa o arquivo da tabela usada no exemplo acima.

```
categoria,salario,idade,vontade
A,10,30,muita
B,20,32,pouca
C,30,33,pouca
D,10,23,muita
A,5,34,pouca
A,40,54,muita
B,50,43,muita
```

Suponha agora que:

- O arquivo possua N linhas, mas a memória do seu computador só tenha capacidade para M linhas (N é muito maior que M);
- Você tem interesse em calcular a média da coluna A, de acordo com as categorias da coluna C.

Para resolver o problema, você deve executar os seguintes passos:

1. Ordenar o arquivo de acordo com a coluna C (categoria). Como o arquivo pode ser muito grande, você precisará usar ordenação externa.
2. Iterar sobre as chaves do arquivo ordenado e calcular as médias dos valores da coluna A para todas as linhas com mesmos valores na coluna C. Veja que agora essa é uma tarefa fácil, pois todas as linhas da tabela com mesmo valor da coluna C serão contíguas no arquivo ordenado e você não precisará colocar o arquivo todo em memória.

ATENÇÃO: há outras formas de resolver o problema em questão. No entanto, você deverá usar o algoritmo descrito acima.

O que deve ser entregue

Você deve entregar apenas um arquivo, `agg.cpp` Esse arquivo deve ser comprimido no formato `.tar.gz` usando o comando `tar -czvf trab2.tar.gz agg.cpp` no Linux.

Entrada

A entrada será feita pela linha de comando. Os seguintes argumentos deverão ser informados:

1. nome do arquivo;
2. número de linhas do arquivo que cabem em memória do computador;
3. nome da coluna contendo a chave de agregação, C;
4. nome da coluna contento o campo que deseja-se calcular a média, A.

Saída

A saída será feita por meio da saída padrão.

Exemplo

A seguir, um exemplo de possível execução no terminal. Suponha que o arquivo `teste.txt` contenha as informações descritas no exemplo anterior.

```
g++ -o agg agg.cpp
./agg teste.txt 3 categoria salario
categoria, salario
A, 18.333333
B, 35.000000
C, 30.000000
D, 10.000000
```

No exemplo acima, a entrada está no arquivo `teste.txt`, a memória do computador pode armazenar no máximo 3 linhas do arquivo, a chave de agregação é a coluna `categoria` e a coluna que terá a média calculada é `salario`.

```
g++ -o agg agg.cpp
./agg teste.txt 3 vontade idade
vontade, idade
muita, 37.5
pouca, 33.0
```

No exemplo acima, a entrada está no arquivo `teste.txt`, a memória do computador pode armazenar no máximo 3 linhas do arquivo, a chave de agregação é a coluna `vontade` e a coluna que terá a média calculada é `idade`.

ATENÇÃO: o formato de saída deve seguir o padrão acima.

ATENÇÃO: cada linha do arquivo pode ter um comprimento arbitrário e várias colunas. Além disso, uma coluna qualquer pode ser usada como chave de agregação e uma coluna qualquer pode ser ter a média calculada. A quantidade de valores únicos na coluna de chave de agregação não necessariamente será pequena.

Critérios de correção

Os seguintes critérios serão considerados:

- Se usar variáveis globais, nota zero;
- Se usar `goto`, nota zero;
- Se seu trabalho não compilar, nota zero. Se por algum motivo seu código compilar (ou funcionar) no seu computador, mas não no computador do professor, o critério de "desempate" será se o seu trabalho compila (funciona) nos computadores do laboratório CCE 416, considerando o Sistema Operacional Ubuntu. Muita atenção aos usuários de Windows!
- Fração de respostas corretas;
- Organização e modularização do código;
- Legibilidade, i.e., código comentado e nomes intuitivos para as variáveis;

- Presença de vazamento de memória e acesso a posições inválidas de memória. Use `Valgrind` desde os primeiros testes! Em casos extremos, a não observância desse quesito implicará em nota zero;
- Eficiência! Haverá bônus para os trabalhos com as implementações mais eficientes.

Como deve ser entregue

Cada **trio** deve enviar apenas um trabalho para o e-mail `gcom.tp.sub@gmail.com`, até às 23:59 do dia 22 de outubro de 2019. O trabalho deve ser enviado por um e-mail de domínio `ufv.br` (e-mails de qualquer outro domínio não serão considerados).

O título (subject) do e-mail deve conter o número de matrícula dos integrantes do grupo (sem o ES), separados por uma vírgula. Se você optar for fazer o trabalho sozinho, esse campo terá apenas seu número de matrícula.

Se o mesmo grupo enviar mais de um trabalho, apenas o e-mail mais recente será considerado. O e-mail deve ter em anexo apenas um arquivo comprimido, no formato `.tar.gz`, com nome `trab2.tar.gz`, contendo seus arquivos fontes.

Após baixar o arquivo, o professor digitará os comandos:

```
tar -xzvf trab2.tar.gz
gcc -o agg agg.cpp
```

ATENÇÃO: a conformidade com os critérios aqui estabelecidos faz parte da avaliação. Se você não entregar o trabalho no prazo, ou se o executável não for gerado da forma indicada, você receberá nota zero.

Algumas outras regras

- O trabalho deve ser implementado em C++ (compilado com `g++`);
- O trabalho pode ser feito em trio. Pessoas do mesmo grupo no trabalho 1 não poderão estar no mesmo grupo no trabalho 2;
- Plágio não será tolerado. O regimento acadêmico será seguido à risca em caso de suspeita.