Universidade de Brasília

Faculdade UnB Gama

Trabalho Prático: Busca e Ordenação
Disciplina: FGA0147 - Estruturas de Dados e Algoritmos
Prof. Dr. Nilton Correia da Silva
Faculdade UnB Gama - FGA
Universidade de Brasília - UnB

Aplicação de Algoritmos de Busca e Ordenação em Análises de Atrasos de Voos de Companhias Aéreas Norte Americanas.

Fonte de dados: Airlines Dataset to predict a delay

Gama, Distrito Federal

Sumário

Orientações

Trabalho Prático para composição da média final da disciplina de Estruturas de Dados e Algoritmos. Este trabalho foi elaborado considerando que sua resolução deve ser distribuída para um grupo de alunos.

A solução deverá ser em linguagem C ou C++. No caso do grupo adotar C++, poderá se valer de uma solução orientada a objetos, contudo sua solução não deve usar classes e contêineres já prontos da linguagem C++ - por exemplo, algoritmos de ordenação, busca, objetos vector, list, matrix, etc.

1 Introdução

Algoritmos de busca e ordenação são repetidamente demandados em diferentes tipos de análises de dados. Análises estatísticas que visam quantificar ou qualificar uma determinada característica, por vezes, requerem que o algoritmo realize ordenações e buscas em um dataset tendo como chave uma ou mais de suas variáveis - geralmente, as variáveis vêm em forma de colunas do dataset.

Neste trabalho teremos a oportunidade de exercitar algoritmos de ordenação e busca em um caso concreto representado por um dataset que foi preparado para que engenheiros de IA (Inteligência Artificial) e outros profissionais pudessem apresentar soluções para um problema envolvendo o referido dataset. Nosso objetivo aqui, contudo, não será apresentar uma solução final para o problema apresentado no link da fonte de dados (veja a referência abaixo do título do trabalho), mas sim, apresentar uma solução mais simples que será detalhada nas seções posteriores.

Caso o dataset disponibilizado tenha uma quantidade de registros maior do que os seus recursos computacionais conseguem processar, você pode excluir registros ou colunas (desnecessárias à solução), deixando apenas a quantidade máxima de dados que seus recursos computacionais conseguem processar.

2 Dataset

O dataset escolhido para este trabalho refere-se ao registro de voos de diferentes companhias aéreas norte americanas. Ademais aos detalhes do voo, o dataset informa ainda se o voo atrasou ou não (variável booleana).

Desafio para o Engenheiro de IA: Gerar um modelo de IA que seja capaz de prever se um voo irá atrasar ou não.

2.1 Volumetria

Instâncias: 539.383 linhas
Atributos: 18 colunas
Tamanho: Compactado: 6,5MB, Descompactado: 18,51 MB
Formato: csv (valores separados por vírgula)

2.2 Variáveis de Interesse

Sua solução não demandará todas as variáveis do dataset. Veja abaixo o detalhamento das variáveis (colunas do dataset) que você precisará processar.

2.2.1 ID do Voo

Esta variável está na coluna **id**. Esta coluna apresenta um número sequencial que identifica unicamente um voo.

Trata-se de um valor inteiro.

2.2.2 Companhia Aérea

Esta variável está na coluna **Airline**. Esta coluna apresenta a sigla da companhia aérea.

Trata-se de uma string com 2 caracteres (letras e números).

2.2.3 Voo Atrasado ou Não

Esta variável está na coluna **Delay**. Esta coluna apresenta 0, caso voo não teve atraso ou 1, quando se tratar de voo com atraso.

Trata-se de um inteiro [0 | 1].

3 Objetivo

A solução deste trabalho consiste em responder à seguinte questão:

Qual a quantidade média de vezes que as companhias aéreas norte americanas atrasam?

Para responder a esta pergunta, você deve apresentar um gráfico com valores médios de atraso das companhias aéreas (Figura 1).

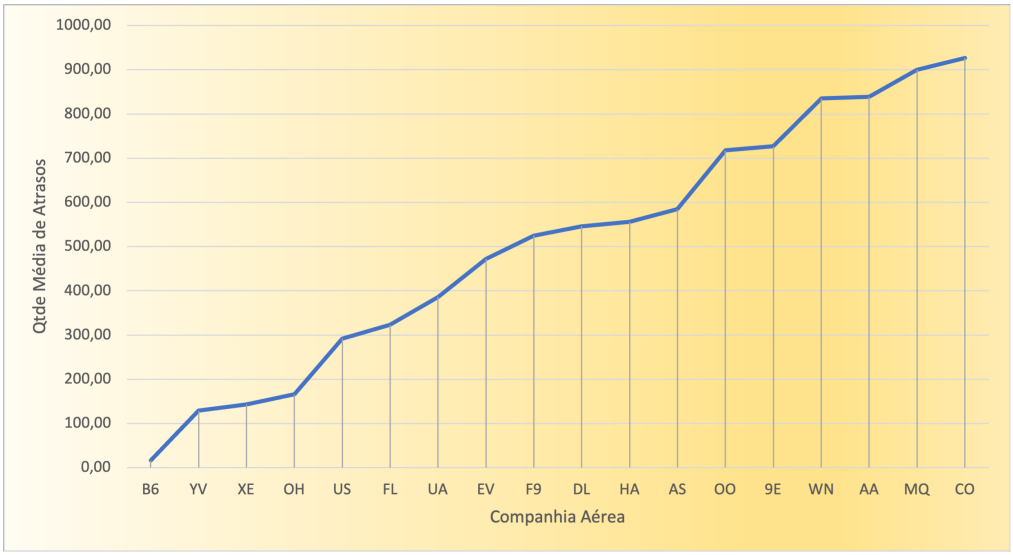

Figura 1: Formato do Gráfico da solução encontrada.

O gráfico da Figura 1 poderá ser gerado por qualquer aplicativo (MS Excel, Google Sheet, etc) desde que o mesmo seja gerado importando um arquivo csv que a sua solução (seu programa) deverá gerar. Segue abaixo o detalhamento do arquivo de saída que sua solução deverá gerar.

3.1 Variáveis do Arquivo de Saída

Companhia:. Esta coluna deve trazer a sigla das companhias aéreas encontradas no arquivo de entrada.

Media:. Esta coluna deve trazer a quantidade média de vezes que uma linha aérea atrasa. É um campo que deve ser calculado dividindo-se a quantidade de vezes que a companhia aérea atrasou pela quantidade total de voos realizados por ela.

3.2 Escolha do Método de Ordenação

A ordenação é um ponto importante deste trabalho. Um dos métodos abaixo deverá ser adotado para a solução:

1. Insert Sort
2. Bubble Sort

3. Selection Sort

4. Quick Sort

Veja: Principais Algoritmos de Ordenação

4 Condições de Contorno da Solução

Sua solução deve atender às seguintes condições:

1. O arquivo de entrada (dataset) deve ser lido por seu programa em sua apresentação original - não deve ser ordenado antes por outro programa.

2. As linhas do arquivo de saída devem estar em ordem crescente pela coluna **Media**.

5 Entregáveis e Notas

A avaliação deste trabalho se dará mediante a apresentação dos itens abaixo.

5.1 Código-fonte da Solução - 5 pontos

Mostrar execução do programa. Se for demorado executar para o dataset inteiro, prepare um dataset reduzido para esta etapa.

5.2 Gráfico conforme Figura 1 - 2 pontos

Mostrar a importação do arquivo de saída gerado em um aplicativo e gerar o gráfico.

5.3 Relatório Técnico - 3 pontos

5.3.1 Tempo de Processamento para o Cálculo das médias da Figura 1 - 1 ponto

5.3.2 Complexidade do Algoritmo de Ordenação escolhido - 2 pontos

1. O() - 1 ponto

2. Tempo de Processamento - 1 ponto