

**Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação
Disciplina de Estrutura de Dados III (SCC0607)**

Docente

Profa. Dra. Cristina Dutra de Aguiar

cdac@icmc.usp.br

Monitores

Eduardo Souza Rocha

eduardos.rocha17@usp.br ou telegram: @edwolt

Beatriz Aimee Teixeira Furtado Braga

beatriztfb@usp.br ou telegram: @bia_aimee

Heitor Tanoue de Mello

heitortanoue@usp.br ou telegram: @heitortanoue

Trabalho Introdutório

Este trabalho tem como objetivo obter dados de um arquivo de entrada e gerar um arquivo binário com esses dados, bem como realizar operações de busca. Ele é um trabalho introdutório, de forma que será usado como base para o desenvolvimento de todos os demais trabalhos da disciplina.

O trabalho deve ser feito por 2 alunos. A solução deve ser proposta exclusivamente pelos alunos com base nos conhecimentos adquiridos nas aulas. Consulte as notas de aula e o livro texto quando necessário.

Fundamentos da disciplina de Bases de Dados

A disciplina de Estrutura de Dados III é uma disciplina fundamental para a disciplina de Bases de Dados. A definição dos trabalhos práticos é feita considerando esse aspecto, ou seja, os trabalhos são especificados em termos de várias funcionalidades, e essas funcionalidades são relacionadas tanto com desafios enfrentados no mercado de trabalho quanto com as funcionalidades da linguagem SQL (*Structured Query Language*), que é a linguagem utilizada por sistemas gerenciadores de banco de dados (SGBDs) relacionais. As características e o detalhamento de SQL serão aprendidos na disciplina de Bases de Dados. Porém, por meio do desenvolvimento deste trabalho prático, os alunos podem entrar em contato com alguns comandos da linguagem SQL e verificar como eles são implementados.

Os trabalhos práticos têm como objetivo armazenar e recuperar dados referentes aos ecossistemas tecnológicos e como as tecnologias estão relacionadas entre si. Esses dados encontram-se disponíveis no Stack Overflow e podem ser obtidos a partir da URL https://www.kaggle.com/datasets/stackoverflow/stack-overflow-tag-network?select=stack_network_nodes.csv.

Descrição do Arquivo de Dados

Um arquivo de dados possui um registro de cabeçalho e 0 ou mais registros de dados. A descrição do registro de cabeçalho é feita conforme a definição a seguir.

Registro de Cabeçalho. O registro de cabeçalho deve conter os seguintes campos:

- *status*: indica a consistência do arquivo de dados, devido à queda de energia, travamento do programa, etc. Pode assumir os valores ‘0’, para indicar que o arquivo de dados está inconsistente, ou ‘1’, para indicar que o arquivo de dados está consistente. Ao se abrir um arquivo para escrita, seu *status* deve ser ‘0’ e, ao finalizar o uso desse arquivo, seu *status* deve ser ‘1’ – tamanho: *string* de 1 byte.
- *proxRRN*: armazena o valor do próximo RRN (número relativo do registro) disponível. Deve ser iniciado com o valor ‘0’ e deve ser alterado sempre que necessário – tamanho: inteiro de 4 bytes.
- *nroTecnologias*: indica a quantidade de tecnologias diferentes que estão armazenadas no arquivo de dados. Note que, se duas ou mais tecnologias têm o mesmo nome, elas são consideradas a mesma tecnologia. Deve ser iniciado com o valor ‘0’ e deve ser alterado sempre que necessário – tamanho: *inteiro* de 4 bytes.
- *nroParesTecnologias*: indica a quantidade de pares (nomeTecOrigem, nomeTecDestino) diferentes que estão armazenados no arquivo de dados. Deve ser iniciado com o valor ‘0’ e deve ser alterado sempre que necessário – tamanho: *inteiro* de 4 bytes.

Representação Gráfica do Registro de Cabeçalho. O registro de cabeçalho deve ser representado da seguinte forma:

1 byte	4 bytes	4 bytes	4 bytes
<i>status</i>	<i>proxRRN</i>	<i>nroTecnologias</i>	<i>nroParesTecnologias</i>
0	1...4	5...8	9...12

Observações Importantes.

- O registro de cabeçalho deve seguir estritamente a ordem definida na sua representação gráfica.
- Os campos são de tamanho fixo. Portanto, os valores que forem armazenados não devem ser finalizados por '\0'.
- Neste projeto, o conceito de página de disco não está sendo considerado.

Registros de Dados. Os registros de dados são de tamanho fixo, com campos de tamanho fixo e campos de tamanho variável. Para os campos de tamanho variável, deve ser usado o método indicador de tamanho.

Os campos de tamanho fixo são definidos da seguinte forma:

- *grupo*: grupo ao qual a tecnologia de origem pertence – tamanho: *inteiro* de 4 bytes.
- *popularidade*: total de milhões de acessos que existe no Stack Overflow para aquela tecnologia de origem – tamanho: *inteiro* de 4 bytes.
- *peso*: frequência com que as *tags* de tecnologia no Stack Overflow aparecem juntas em relação à frequência com que aparecem separadamente. Uma tag representa a ideia de relacionamentos entre tecnologias – tamanho: *inteiro* de 4 bytes.

Os campos de tamanho variável são definidos da seguinte forma:

- *nomeTecnologiaOrigem*: nome da tecnologia de origem – *string*
- *nomeTecnologiaDestino*: nome da tecnologia de destino – *string*

Adicionalmente, o seguinte campo de tamanho fixo também compõe cada registro.