Algoritmos e Estruturas de Dados II
Trabalho I

Especificação do Trabalho I:
Inicialmente cada equipe deverá criar o seu projeto no Git Hub e fazer as postagens de implementação de código e demais arquivos necessários.

Organização: em duplas (ou individual) - definir seu grupo para a dupla no AVA, se for trabalhar em dupla.

Explicação geral:
Atualmente, as redes sociais possuem uma grande variedade de informação e de dados disponíveis. A busca por termos (conhecidos por *hashtags*) é amplamente utilizada através de mecanismos de busca ou na linha do tempo. Com esta quantidade de dados e informações disponíveis, é possível fazer a extração de dados visando responder a uma pergunta específica ou para fazer análise de dados sobre um contexto específico. **Por exemplo**, se a seguinte pergunta fosse relevante: *Qual é a medida preventiva para a COVID-19 que é mais comentada no Brasil na rede social Twitter?* Para responder a esta pergunta, pode-se criar uma base de dados a partir de extração dos *tweets* pesquisando pelo usuário correspondente @COVID19 ou por palavras chaves, tais como, #covid19prevencao, #covid_19brasil, bem como, pesquisar nas páginas do Ministério da Saúde @minsaude , Secretaria da Saúde do RS @SES_RS, entre outras.

A partir deste contexto, o próximo passo é extrair as informações e montar uma base de dados. Para definir a estrutura da base de dados, é necessário definir quais as informações serão relevantes. Para este contexto apresentado como exemplo (COVID-19) poderia se considerar relevantes *o nome do usuário que postou a mensagem*, *a mensagem em si*, *o local de postagem*, *as palavras de busca, etc*.

Após a base de dados ser construída, ainda considerando o exemplo apresentado, pode-se fazer uma pesquisa pelo tipo de prevenção, se é indicação de remédio, de atitudes, de uso de máscara, de distanciamento social, etc.

Atividades a realizar

1) **Definição do contexto a ser explorado:** escolha um dos contextos a seguir:
   - eleições municipais 2020
   - queimadas no Brasil

2) **Montagem do arquivo de dados**

A primeira atividade do trabalho envolve a construção do arquivo de dados. Para tanto, **cada equipe deve escolher o contexto que será considerado**, e para esse contexto

definir uma ou mais hipóteses (perguntas) a serem testadas em uma rede social. Assim, as seguintes tarefas deverão ser realizadas:

- Definir uma ou mais hipóteses (serão as consultas que serão realizadas nos dados);
- Escolher a rede social a ser utilizada, e pesquisar como é possível extrair os dados dela (e se há restrições para a extração, como por exemplo se há uma quantidade máxima de dados que podem ser extraídos por dia);
- Extrair os dados, fazer uma limpeza e preparar a carga dos dados, definindo a estrutura da base de dados para, no mínimo, 10 mil linhas com conteúdo válido.

Para a extração dos dados de uma rede social, cada equipe poderá desenvolver o seu próprio código ou poderá utilizar um código disponível na internet para a extração de dados da rede social. Sugere-se o uso da rede social Twitter, pela facilidade de uso das APIs, mas pode ser outra.

2.1) Registros do Arquivo de Dados:

Twitter: a estrutura do registro a ser utilizado para os dados extraídos da rede social Twitter é composta pelos campos *id_twitter* (inteiro, chave primária), *usuário* (caractere de 20 posições), *mensagem* (caractere de 280 posições), *data* (caractere de 8 posições), *país* (caractere de 20 posições) e *hashtags* (caractere de 200 posições).

Para a implementação dessa funcionalidade deve-se criar um **registro de tamanho fixo** com o caractere '\n' de final de linha (opcional) e **sem separador entre os campos** (obrigatório) em uma linguagem de programação (C, C#, C++, Python, PHP, Java ...) que possua o comando *seek.*

Se outra rede social for a escolhida, estes campos podem ser adequados aos dados extraídos, assim como as demais orientações apresentadas nas demais explicações do trabalho.

Organização de Arquivo de Dados:

- Criar o registro de dados descrito acima para a organização de arquivo do tipo sequencial usando registro de tamanho fixo com o caractere '\n' de final de linha (opcional) e sem separador entre os campos (obrigatório).
- Implementar:
  - 1) um procedimento para inserir as, no mínimo, 10 mil linhas de dados,
  - 2) um procedimento para mostrar os dados,
  - 3) um procedimento para realizar a pesquisa binária e
  - 4) um procedimento para consultar dados a partir da pesquisa binária.

Os dados devem ser coletados ao longo do mês de outubro de 2020 (podem ser coletados dados anteriores se for possível).

2.2) Índices em arquivo:

- Implemente **um arquivo de índice** para o campo ***id_twitter*** de acordo com a descrição do índice de arquivo da organização sequencial-indexado. **Implemente um procedimento de consulta a partir deste índice** usando a **pesquisa binária** para pesquisar no arquivo de índice e, depois o comando *seek* para pesquisar no arquivo de dados.

- Implemente **um arquivo de índice** para o campo ***hashtags*** de acordo com a descrição do índice de arquivo da organização sequencial-indexado. **Implemente um procedimento de consulta a partir deste índice** usando a **pesquisa binária** para pesquisar no arquivo de índice e, depois o comando *seek* para pesquisar no arquivo de dados.

Índices em memória:

- Implemente uma estrutura de ***hash em memória*** para o campo ***data****.* Implemente um procedimento de consulta a partir deste índice e, depois o comando *seek* para pesquisar no arquivo de dados.

- Implemente uma estrutura de **árvore binária** para o campo ***hashtags***. Implemente um procedimento de consulta a partir deste índice e, depois o comando *seek* para pesquisar no arquivo de dados.

3) Resposta à(s) hipótese(s):

Implemente um procedimento para responder a hipótese (ou as hipóteses) definida no início do trabalho.

4) Postar no AVA:

- Código fonte
- Arquivos de dados
- Link para o projeto no GiT Hub

Avaliação:

- O trabalho vale 10 pontos e será avaliado conforme o cumprimento das atividades propostas e a utilização de boas práticas de programação.

- Não é permitido o uso da memória RAM para armazenar todos (ou grande parte) dos registros do arquivo para efetuar as buscas. Todas as operações solicitadas devem ser executadas no arquivo de dados armazenado em memória secundária (disco rígido e similares).