

Tarefa 1_A ----- TEG

Especificação da tarefa 1_A:

Implementação de um grafo com carga primária a partir de um arquivo CSV.

O grafo será representado na forma de uma lista de adjacências primariamente carregado a partir de uma base de dados no formato CSV (valores separados por vírgulas). No caso, a base utilizada será a *Iris dataset*.

A Iris é um gênero de flor que apresenta três espécies distintas: virgínica, setosa e versicolor, as quais podem ser caracterizadas por meio dos comprimentos de pétalas e sépalas (partes que sustentam o botão floral). Partindo dessa observação, foi construída uma base de dados contendo características morfológicas da flor Iris.

A base Iris é muito aplicada em exemplos de casos de uso para análise de dados, se constituindo de 150 observações (50 para cada espécie de Iris), sendo que cada observação corresponde a 4 medições das seguintes variáveis: “Sepal.length” e “Sepal.width” (comprimento e largura das sépalas das flores, respectivamente), além de “Petal.length” e “Petal.width” (comprimento e largura das pétalas das flores, respectivamente).

Referencias sobre a base Iris podem ser encontradas em diferentes fontes, por exemplo:

<https://archive.ics.uci.edu/ml/datasets/iris>

https://pt.wikipedia.org/wiki/Conjunto_de_dados_flor_Iris



A tarefa ora em especificação demanda a implementação de um grafo na forma de um tipo abstrato de dados, cuja interface contempla algumas operações descritas nos requisitos listados a seguir.

Requisitos funcionais:

- Implementação em linguagem C;
- A estrutura de dados do grafo deve estar na forma de uma lista de adjacências montada a partir da base de dados Iris;
- O sistema tem que permitir a carga primária do grafo a partir da base Iris em formato CSV;

d) Cada caso (observação) na base de dados Iris será um vértice do grafo, portanto, $V=\{c1,c2,c3,...,c150\}$;

e) Para todo vértice v_i do grafo, ocorrerá uma aresta $(v_i v_j)$ apenas se a distância euclidiana normalizada $DEN(v_i v_j) \leq 0,3$. A distância euclidiana normalizada $DEN(v_i v_j)$ corresponde à normalização da distância euclidiana entre o par de vértices $(v_i v_j)$;

f) O sistema deve permitir a persistência do grafo por meio de arquivo TXT, viabilizando a recarga do grafo sem a necessária realização da releitura da base Iris CSV, bem como do recálculo da tabela DEN e demais passos.

Requisitos não funcionais:

a) Resumo dos passos para implementação da carga primária do grafo:

IRIS csv \rightarrow tabela com as distâncias euclidianas para todo par de vértices \rightarrow normalização \rightarrow aplicação de limiar 0,3 \rightarrow lista de adjacências

b) A distância euclidiana normalizada $DEN(v_i v_j)$ corresponde à normalização da distância euclidiana entre o par de vértices;

A normalização de dados é uma técnica de mudança de uma escala original (dos valores de uma variável) para a escala 0 – 1. Você pode encontrar uma discussão sobre isso no link abaixo:

<https://towardsdatascience.com/normalization-vs-standardization-explained-209e84d0f81e>

Uma forma simples de obter a distância euclidiana normalizada (DEN) é calcular todas as possíveis distâncias euclidianas no grafo, constituindo uma tabela, normalizar essa tabela e aplicá-la na criação das adjacências do grafo, conforme é requisitado no problema proposto.

c) Formato do arquivo TXT para a persistência desejada e respectivo grafo:

5 <número de vértices>

1, 2

2,3

3,4

1,4

