

Tarefa 1_A -- TEG

Entre no link de upload no Moodle (tarefa_1_A) e veja as instruções para entrega

Especificação da tarefa 1_A:

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

O grafo será representado na forma de uma matriz de adjacências primariamente carregado a partir de uma base de dados no formato CSV (valores separados por vírgulas). No caso, a base utilizada será uma versão reduzida da base *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:

- a) A estrutura de dados do grafo deve estar na forma de uma matriz de adjacências montada a partir da base de dados;
- b) O sistema tem que permitir a carga primária do grafo a partir da base Iris em formato CSV ignorando a primeira coluna da esquerda, a qual identifica o nome da espécie da flor, levando em conta as colunas: "Sepal.length" e "Sepal.width" (comprimento e largura das sépalas das flores, respectivamente), além de "Petal.length" e "Petal.width" (medidas em centímetros);
- c) Cada caso (observação) na base de dados será um vértice do grafo, portanto, $V=\{c1,c2,c3,...,c150\}$;
- d) Para todo vértice v_i do grafo, ocorrerá uma aresta $(v_i v_j)$ se e somente se 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 pares de vértices $DE(v_i v_j)$;

Veja conceitos sobre distâncias em:

<https://www.meuguru.com/blog/distancia-3-mais-usadas/>

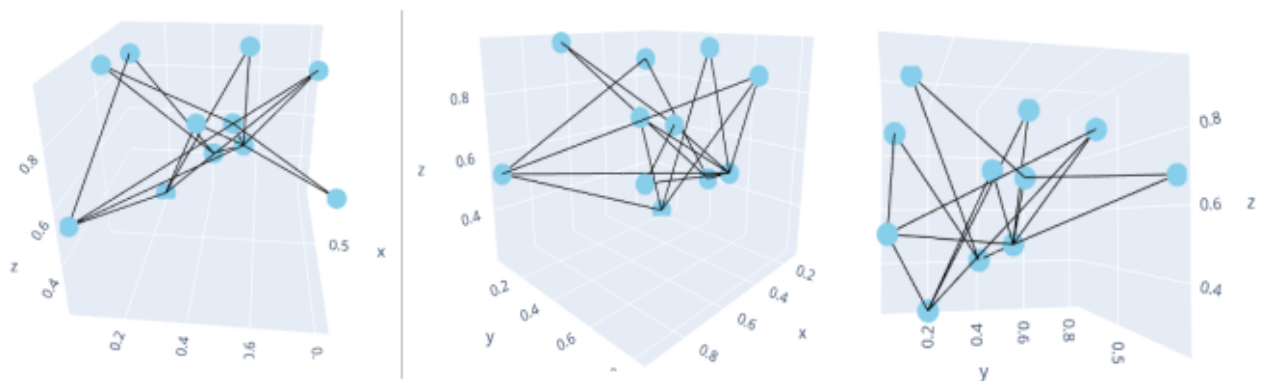
Sobre normalização em:

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

- e) O sistema deve permitir a persistência do grafo por meio de arquivo CSV, viabilizando a recarga do grafo sem a necessária realização da releitura e reprocessamento sobre a base original, bem como do recálculo da tabela de DENs e demais passos.

No cabeçalho deste arquivo CSV deve constar:

- a) O total de vértices lidos;
 - b) A maior DE e o par de vértices $(v_i v_j)$ que a determinaram;
 - c) A menor DE e o par de vértices $(v_i v_j)$ que a determinaram;
 - d) A maior DEN e o par de vértices $(v_i v_j)$ que a determinaram;;
 - e) A menor DEN e o par de vértices $(v_i v_j)$ que a determinaram;
- f) Não é exigida a integração do Python com o código em "C", mas é solicitada utilização de script Python (citando a fonte do script) para a visualização tridimensional do grafo obtido. Esse script tem que permitir a visualização por diferentes pontos de vista via utilização do mouse sobre a área gráfica (algo similar foi disponibilizado no Moodle):



Observações:

a) Resumo dos passos para implementação da carga primária do grafo:
base de dados csv → tabela com as distâncias euclidianas para todo par de vértices → tabela com as distâncias euclidianas normalizadas para todo par de vértices → aplicação de limiar 0,3 → matriz de adjacências

As arestas representam relações de agrupamento entre indivíduos de mesma espécie, portanto, não há sentido em arestas direcionadas, múltiplas ou laços (não é necessário avaliar o limiar na distância do espécimen para ele mesmo).

b) A $DEN(v_i, v_j)$ corresponde à normalização das distâncias euclidianas entre pares de vértices;

A normalização de dados é uma técnica de mudança de uma escala original (dos valores de uma variável) para a escala de 0 a 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 tabela DEN é calcular todas as possíveis distâncias no grafo, constituindo uma tabela, normalizar essa tabela e aplicá-la na criação das adjacências do grafo, conforme é requisitado no problema proposto.