Universidade Federal de Uberlândia

Faculdade de Computação – Sistemas de Informação

GSI556 – Mineração de Dados / Monte Carmelo-MG

1ª Atividade Avaliativa

TEMA: Similaridade/Dissimilaridade e Pré-processamento
• Este exercício pode ser feito **Individual;** • Data de entrega: 29/02/2024 • Está atividade é composta por 2 (dois) exercícios distintos; • Forma de entrega: Arquivo contendo o código e um relatório. O relatório pode ser um arquivo de texto ou o próprio código contendo a descrição do que foi implementado (linha por linha). • Entrega: via Teams. • Apresentação: Presencial.

1) **Similaridade de enzimas entre diferentes organismos**

Os três arquivos anexados representam as sequências DNA de enzima topoisomerase 1 de três organismos: rato (rat.fasta), hamster chinês (hamster.fasta) e cavalo (horse.fasta). Usando sequências armazenadas nestes arquivos, implemente um algoritmo em Python que faça:

- uma comparação de proximidade usando verificação simples entre dois organismos;
- a contagem de ocorrência de cada aminoácido nas sequências construindo um vetor numérico de ocorrências e calcule as distâncias Manhattan, euclidiana, supremum, e a similaridade de cosseno entre dois organismos.

Observações:

1.1) Use a função abaixo para ler os arquivos fasta e retornar as sequências DNA como uma variável de texto.

```
def read_fasta(arq):
  seq = ''
  with open(arq) as f:
    f.readline()
    for line in f:
     seq += line.strip()
  return seq
```

1.2) A lista de letras que representam os aminoácidos no formato FASTA pode ser encontrada no site https://pt.wikipedia.org/wiki/Formato_FASTA

1.3) Anexa os programas em Python junto com arquivos de texto de saída que mostram os resultados.

2) **Analise de qualidade de vinho**

A tarefa é baseada no tutorial sobre pré-processamento de dados em Python do site Data Flair.

O arquivo anexado contem dados de analise de vinhos portugueses publicado no site UCI Machine Learning Repository. (winequality-red.csv)

O tutorial mostra como realizar normalização, padronização, transformação e binarização de dados usando pacotes de Python: Pandas e SKLearn. A segunda parte mostra como realizar a visualização de dados usando histogramas, gráficos de densidades, e outros usando pacotes: NumPy e Matplotlib.

O resultado deste exercício seria os programas demostrando as operações sobre dados e arquivos de gráficos como histogramas, etc.

Atenção: Precisa entregar os programas e histogramas de atributos para cada normalização.