UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL - UFRGS
INSTITUTO DE INFORMÁTICA - DEPTO INFORMÁTICA TEÓRICA
BIOLOGIA COMPUTACIONAL- 2018

LISTA DE EXERCÍCIO VIII

Instruções:

- A resolução do exercício deve ser feita **individualmente**. Cópias evidentes entre trabalhos não serão aceitas.
- A entrega deve ser online via Moodle (exclusivamente), somente até a data especificada. Não serão aceitos trabalhos atrasados.
- Para cada uma das tarefas deve-se entregar o com código fonte. O nome do aquivo deve identificar a tarefa, exemplo "e8-1a.py" referente ao item "1a" da tarefa. Arquivos corrompidos serão desconsiderados.
- Além do código fonte deve-se entregar um um único arquivo PDF apresentando o pseudocódigo do algoritmo desenvolvido e os resultados encontrados.
- Data de entrega: 20.11.2018 (terça-feira) até as 13:00 via Moodle (`https://moodle.ufrgs.br/login/index.php`).

NOME: ..... CARTÃO: .....

Objetivos: Implementação do Algoritmo de Agrupamento K-means, Dados de Microarranjo.

1. Relize a implementação do algoritmo K-means apresentado em sala de aula. Um guia para implementação do algoritmo em python pode ser encontrado em `https://mubaris.com/posts/kmeans-clustering`
2. Baixe o arquivo de treinamento disponível no link: `https://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv`. Esta base de treinamento é formada por perfis de expressão gênica de 72 amostras de medula óssea de pacientes com leucemia aguda, cada perfil consiste da expressão de 7129 genes. Os exemplos de treinamento estão rotulados como `ALL` (*acute lymphoid leukemia*) e `AML` (*acute myeloid leukemia*), dois tipos distintos de leucemia.
3. Utilizando o algoritmo k-means implementado, Agrupe os dados da base de treinamento por amostra (paciente) usando como parâmetro k=2 (dois clusteres) e k=3 (três clusteres). Entregar relatório de no mínimo 2 paginas descrevendo o experimento e os resultados obtidos. Para cada um dos respectivos grupos (2 e 3) diga quantas amostras de pacientes `ALL` e `AML` aparecem em cada grupo. Apresente graficamente estes resultados.

→ Bibliotecas prontas não devem ser utilizados. Cada um deve desenvolver o seu próprio algoritmo de agrupamento.