IEEE
Computational
Intelligence
Society

Universidade de Brasília
IEEE Student Branch

Treinamento CIS
2º Período
Clusterização

Descrição do Tema

“Agrupamento (Clustering) é uma técnica de aprendizado não supervisionado que organiza um conjunto de objetos de forma que os objetos dentro do mesmo grupo (ou *cluster*) sejam mais semelhantes entre si do que em relação aos objetos de outros grupos.

O agrupamento auxilia na compreensão da estrutura dos dados e na revelação de padrões ocultos. Ele também pode ser utilizado em conjunto com algoritmos de aprendizado supervisionado quando os dados rotulados são escassos ou caros de se obter, em um processo conhecido como **aprendizado semi-supervisionado**.” - Roi Yehoshua, Assistant Professor at Northeastern University

Conteúdos

1. Introdução a clusterização;
    a. Conceitos fundamentais;
    b. Tipos de algoritmos de clusterização;
    c. O algoritmo K-Means;
2. Tópicos extras de desenvolvimento;
    a. Visualização de dados em Python;
    b. Escrita de código em Python;
    c. Estudos;
    d. Engenharia de Software;

Materiais:

1. Introdução à Clusterização;
    a. Livro base - Data Science do Zero - Capítulo 19;
    b. Playlist teórica em inglês - Victor Lavrenko - Clustering Algorithms; (Vídeos 1 ao 8)
    c. Vídeo teórico em português - Prof. Paulo Santos - Algoritmos de Clusterização;
    d. Vídeo resumo em inglês - 4 Basic Types of Clustering Used in Data Analytics;

2. Tópicos extras de desenvolvimento
    a. Bibliotecas para visualização de dados em Python - Vídeo;
    b. 5 Dicas para escrever um bom código em Python - Vídeo;
    c. Estudar muito não significa aprender - Vídeo;
    d. O colapso da engenharia de software - Vídeo;

Tarefas

Utilizando como base o dataset Student Habits vs Academic Performance:

1. Apenas com a observação dos dados por meio de tabelas e dataframe, apresente suas hipóteses;
2. Realize uma EDA da forma que julgar mais adequada;
3. Com base na EDA realizada, revise as suas hipóteses a respeito do dados;
4. Construa um algoritmo de K-Means from scratch utilizando a linguagem Python;
5. Apresente a justificativa de forma discursiva e por meio de cálculos, o valor de K utilizado;
6. Revise as hipóteses levantadas no exercício 1 e 2 e com base no algoritmo desenvolvido, apresente análise crítica a respeito das informações extraídas dos dados caso o algoritmo refute ou confirme suas hipóteses;

Tarefas Extra:

7. Exercite sua habilidade de pesquisa e curadoria de informações e a partir de fontes próprias, escreva, com suas palavras, de maneira teórica e com o aprofundamento que achar adequado os algoritmos DBSCAM e Hierarquical Clustering;
8. Exercite sua habilidade de pesquisa e curadoria de informações e a partir de fontes próprias, escreva, com suas palavras, de maneira teórica e com o aprofundamento que achar adequado qual o algoritmo state-of-the-art para Clustering. Apresente as referências utilizadas.

Recomenda-se que para neste segundo período, o estudante realize as tarefas em ordem de apresentação e se utilize da criatividade, habilidade essencial para um cientista de dados e lembre-se: “Se você torturar os dados por tempo suficiente, eles confessarão”.