Série N°2 en Fouille de données (Clustering)

Exercice n°1 : Soit les points A (2, 10) ; B (2,8) ; C (8,4) ; D (5,8) ; E (7,5) ; F (6,4) ; G (1,2) ; H (4,9)

1. Donner la répartition géométrique de ces points. Quels sont les clusters qu'on peut identifier visuellement ?
2. En prenant comme centroïdes initiaux les points A B et C, appliquer l'algorithme K-means pour regrouper les points en trois clusters (utiliser la distance de Manhattan).
3. Est-il possible de minimiser le nombre d'itérations par un autre choix des centroides initiaux ? Justifier la réponse.

Exercice n°2 : Soit la liste suivante des employés avec leur ancienneté en années et salaire par unité monétaire.

Employé	E1	E2	E3	E4	E5
Ancienneté	2	3	5	6	10
Salaire	2000	2100	3500	4100	10000

1. Regrouper les individus en utilisant la classification hiérarchique ascendante et en prenant comme mesure de similarité la distance euclidienne et comme linkage le linkage single.
2. Donner le tableau des valeurs standardisées des variables.
3. Regrouper à nouveau les employés avec les valeurs standardisées en utilisant la même technique en 1.
4. Comparer les résultats avant et après standardisation.

Exercice n°3 : Le conseil d'administration d'une entreprise a du mal à mettre d'accord ses membres sur chaque nouveau projet d'investissement. Chaque projet possède plusieurs points de discorde (p1 : emplacement, p2, durée, p3 : budget, p4 : nombre d'intervenants, etc.). Le président-directeur général soumet une proposition contenant **un choix** par point. Chaque membre peut être d'accord (OK) sur chaque point ou pas d'accord (NOK). L'objectif de l'entreprise est de minimiser les interactions entre les membres (deux à deux ou entre groupes) et d'accélérer la prise de décision. Elle met à leur disposition une fiche sur laquelle ils peuvent mentionner leur avis sur chaque point. Sur la base de cette fiche, on veut accélérer le processus de réunions selon le principe : plus on est d'accord sur beaucoup de points, plus on va vite dans la réunion.

1. Reformuler le problème en un problème de clustering en identifiant *les objets*, *les variables*, *l'objectif du clustering* et *la mesure de similarité*.

2. Quelle est la technique de clustering adéquate dans ce cas ? justifier la réponse.

3. Expliquer comment utiliser le résultat du clustering pour programmer les réunions.

4. Illustrer par un exemple pour 05 points du projet (p1,…,p5) et 05 membres du conseil d'administration (m1,…,m5) les itérations du clustering et l'ordre des réunions.