Cours : Compléments d'algorithmique

Mini-Projet en DataMining

Découverte de schéma pour des sources de données NoSql

OBJECTIF GENERAL :

Les bases de données de données NoSql représentent une grande source d'informations qui ouvrent l'horizon sur de nouvelles applications dans différents domaines.
Une des particularités de ces sources est que les données ne sont pas nécessairement décrites par un schéma.
Contrairement aux bases de données relationnelles, les instances d'une même classe ne sont pas forcément décrites par les mêmes propriétés.
Par conséquent, l'exploitation des sources de données NoSql reste difficile.

Le but de ce mini-projet est d'implémenter un algorithme de clustering afin de calculer les groupes d'instances similaires et qui appartiennent à la même classe.

Étapes à suivre :

1. Charger le fichier NoSql.txt : chaque ligne du fichier représente une instance. Une instance est représentée par un ensemble de propriétés séparées par des espaces.

2. Implémenter un algorithme de clustering (DBSCAN ou K-means).

3. Exécuter le clustering sur les instances chargées. La distance entre les instances est calculée en utilisant Jaccard. Plus des entités partagent des propriétés communes, plus elles sont proche

   $$Jaccard(a,b) = 1 - \frac{|a \cap b|}{|a \cup b|}$$

   a et b représente l'ensemble des propriétés des instances a et b.

4. Interpréter les résultats obtenus : que représente chaque clusters produit.

Cours : Compléments d'algorithmique

Mini-Projet en DataMining

Travail à rendre (Date limite le 06/02/2022) :

1. Le code source de votre programme en commentant les lignes principales
2. Un rapport de 2 pages maximum :
    a. Expliquant votre objectif et justifiant le choix de l'algorithme de clustering
    b. Justifiant les paramètres que vous avez fixé pour votre algorithme
    c. Une interprétation des résultats obtenus : que représentent les clusters que vous avez découvert