UNIVERSITÉ DE SHERBROOKE
DÉPARTEMENT D'INFORMATIQUE

**IFT 870 BIN 710 - Forage de données**

# TP#4 : Fonctions descriptives

Hiver 2020

---

Le but de ce devoir est de pratiquer la comparaison et l'évaluation de méthodes de clustering.

**Ce devoir est à faire individuellement. Il devra être complété avant le jeudi 9 avril 2020 à 23h59. Vous devez remettre, sur `opus.dinf.usherbrooke.ca`, un fichier Ipython notebook (nommé `tp4.ipynb`) contenant votre rapport et vos scripts Python pour ce devoir.**

---

**Description des tâches à réaliser : Ensemble de données d'images de visages de personnages connus**

On vous fournit un fichier de départ `tp4_debut.ipynb` dans lequel on récupère un ensemble de données d'images de visages de personnages connus. Les détails de cet ensemble de données sont disponibles à l'adresse `http://vis-www.cs.umass.edu/lfw/`.

Vous devez appliquer des méthodes d'évaluation extrinsèques et intrinsèques pour comparer deux algorithmes de clustering sur ce jeu de données : K-Means et DBSCAN.

1. Pour commencer, les données ne sont pas équilibrées car certains personnages sont beaucoup plus représentés que d'autres. Pour pallier à cela, filter les données pour ne conserver que 40 visages au maximum par personne.
2. Ensuite, appliquer une réduction de la dimension à 100 composantes et une normalisation en utilisant le modèle `PCA()` de `sklearn` avec les options `whiten=True` et `random_state=0`.
3. Analyse avec K-Means
   (a) Implémenter la méthode du coude (Elbow method) pour essayer de déterminer un nombre de clusters optimals dans l'ensemble suivant [40, 45, 50, 55, 60, ..., 80] sans utiliser les données réelles (noms associés aux images). La mesure de score à utiliser pour tout nombre de clusters k est la suivante : moyenne des distances euclidiennes des données à leur plus proche centre de cluster pour le modèle à k clusters. Analyser le résultat et donner vos conclusions.
   (b) Appliquer une approche de validation croisée en divisant les données en 10 parties et en utilisant les données réelles et le score `Adjusted_Rand_Index (ARI)` pour déterminer un nombre de clusters optimal dans l'ensemble [40, 45, 50, 55, 60, ..., 80]. Analyser le résultat et donner vos conclusions.

4. Analyse avec DBSCAN
   (a) Utiliser le coéfficient de silhouette pour déterminer les meilleurs valeurs de paramètres (nombre minimum d'éléments dans un cluster `min_samples`, et rayon du voisinage autour de chaque donnée `eps`) pour la méthode DBSCAN avec `min_samples` dans l'intervalle [1, ..., 10] et `eps` dans l'intervalle [5, ..., 15] ;
   (b) En fixant le paramètre `min_samples = 3`, appliquer DBSCAN en faisant varier le paramètre `eps` dans l'intervalle [5, ..., 15]. Observer des échantillons d'images des clusters pour chaque rayon dans l'intervalle [5, ..., 15], et tenter de déterminer la signification sémantique des clusterings estimés. Elle peut correspondre à un clustering suivant les personnages, ou suivant d'autres caractéristiques commune comme l'orientation du visage, l'arrière plan, le port de lunette, etc. Lister vos conclusions pour chaque valeur de `eps`.

**Remise du travail**

Pour soumettre votre travail, connectez-vous, dans un fureteur, au serveur `http://opus.dinf.usherbrooke.ca` en utilisant votre CIP, puis choisissez le cours `IFT870 (BIN710)` et le projet `TP4`. Chargez votre fichier `tp4.ipynb` et soumettez-le. Le nom de votre fichier de remise doit être exactement `tp4.ipynb`. Indiquez bien votre nom dans le fichier. Ne remettez pas d'autre fichier.