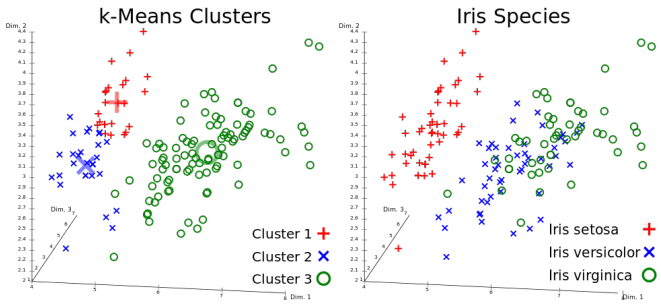
Big Data. TP Projet. Clustering.

Sergey KIRGIZOV

ESIREM



L'objectif est de se familiariser avec les techniques du clustering en utilisant Apache Spark (ou un autre logiciel), et les techniques de visualisation avec Matplotlib (ou un autre logiciel). Chaque groupe présentera ses réalisations lors de la dernière séance de TP. La présentation sera notée sur une échelle de 0 à 20.

☞ EXERCICE : **Sélectionner un jeu de données parmi les jeux suivants. Effectuer le traitement préliminaire des données (sélection des propriétés intéressantes, conversion des données pour le calcul de distances, etc). Effectuer une analyse des données en utilisant l'algorithme k-means. Tester différentes valeurs du paramètre k correspondant au nombre de clusters. Présenter graphiquement les résultats. Pourriez-vous interpréter (décrire chaque groupe en un mot ou une phrase) les clusters trouvés par l'algorithme k-means ?**

ASTUCE : **Si vous rencontrez des problèmes de traitement ou de visualisation de données multidimensionnelles, vous pouvez utiliser les techniques de la réduction dimensionnelle, par exemple PCA.** `https://spark.apache.org/docs/latest/mllib-dimensionality-reduction.html`

ASTUCE : **Si vous rencontrez des problèmes de visualisation de grand nombre de points, utilisez des techniques de binning rectangulaire ou hexagonale.**

Jeux de données

1. Iris, 150 instances `http://archive.ics.uci.edu/ml/datasets/Iris`
2. Vins Italiens, 178 instances `http://archive.ics.uci.edu/ml/datasets/Wine`
3. Qualité du vin, 4 898 instances `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`
4. Génomique, fréquences d'utilisation des codons dans ADN, 13 028 instances `http://archive.ics.uci.edu/ml/datasets/Codon+usage`
5. Chiffres manuscrits. Données optiques. 5620 instances `http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits`
6. Chiffres manuscrits. Tablette sensible à la pression. 10 992 instances `http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits`
7. Banque Marketing, 45 211 instances `http://archive.ics.uci.edu/ml/datasets/Bank+Marketing`
8. Recensement, 48 842 instances `http://archive.ics.uci.edu/ml/datasets/Adult`
9. HIGGS, 11 000 000 instances `http://archive.ics.uci.edu/ml/datasets/HIGGS`
10. Scrutins de l'assemblée nationale `http://data.assemblee-nationale.fr/travaux-parlementaires/votes` Exemple de clustering k-means : `https://www.data.gouv.fr/fr/reuses/clustering-k-means-des-deputes-par-leurs-votes/`

Vous pouvez utiliser également tout autre jeu de données légalement disponible sur le réseau Internet, par exemple les jeux de données provenant de

- UC Irvine Machine Learning Repository : `http://archive.ics.uci.edu/ml/datasets.php?format=&task=clu&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table`
- Plateforme ouverte des données publiques françaises : `https://www.data.gouv.fr/fr/`