CS for Big Data
Christine Largouët
UP Informatique

Projet Big Data

Le but de ce projet est de mettre en œuvre la démarche d'un "data scientist" qui doit analyser un volume conséquent de données hétérogènes à l'aide des techniques abordées dans ce module pour en extraire une information pertinente.

1- Objectifs

Le projet se décompose en cinq étapes

1. Choisir un jeu de données et formuler une problématique
2. Dans un application "Python"
   2.1. Constituer une base de données Nosql à partir des données
   2.2. Traiter les données pour répondre en partie à la question formulée à l'aide de requêtes sur la base de données
   2.3. Effectuer un traitement plus complexe sur les données à l'aide d'un algorithme de type Map/Reduce avec Spark (un plus fortement récompensé...)
3. Réaliser une interprétation graphique des résultats (matplotlib ou autre en python....)

2- Délivabres

Différents él éments doivent être remis

1. Un document écrit (en format pdf), envoyé par mail à largouet@agrocampus-ouest.fr le 19 novembre 2021 au plus tard, décrivant :
   - La problématique étudiée
   - Le choix des données
   - Les outils utilisés en argumentant sur leur utilisation
   - L'analyse des données effectuée pour répondre à la question
2. L'accès au code permettant de constituer la base de données et d'effectuer la première partie de l'analyse à l'aide de la base Nosql
3. L'accès au code permettant la résolution d'un sous-problème en Spark

4- Sources de données

De nombreuses données sont maintenant disponibles sur internet. Voici une liste de liens non exhaustive

- https://www.kaggle.com. (des batailles dans Game of Thrones aux caddies de supermarché)
- https://archive.ics.uci.edu/ml/datasets.php (classés par type de problème)
- site beta : https://archive-beta.ics.uci.edu/
- https://data.worldbank.org (informations démographiques, économiques, etc.)
- https://www.kdnuggets.com/datasets/index.html
- https://www.quora.com/What-kinds-of-large-datasets-open-to-the-public-do-you-analyze-the-mostly
- https://www.imf.org/en/Data (données internationales sur la finance)

5- Evaluation

Les critères d'évaluation de ce projet sont les suivants:

- La pertinence de la question étudiée
- La mise en forme des données
- La complexité des données
- Le choix de la base de données Nosql en fonction du type de données et de leur utilisation
- La justification et la maitrise des outils utilisés
- La beauté du code
- La présentation graphique des résultats