Mini-Projet Spark

Saison 2022-2023

Composante : ESIR

Spécialité : Technologies de l'information option informatique 3ème année

Module : Data management for big data

Chargé de TP : Maria Massri, maria.massri@irisa.fr

1 Description

Ce mini-projet est à effectuer seul ou en binôme. Le compte-rendu ainsi que le code Spark sont à rendre pour le 23/01. Le rendu se fait par mail à l'adresse maria.massri@irisa.fr.

Le jeu de données, dont dépendra les questions que vous vous poserez, est laissé libre. (N.B. Vous pouvez choisir un jeu de données orienté graphe et l'interpréter avec la librairie GraphX).

Des exemples sont cependant fournis en fin de sujet. Nous conseillons de poser trois questions préalables dont la difficulté à y répondre est variable (une facile, moyenne, et difficile par exemple). Vous pouvez aussi rajouter des informations intermédiaires auxquelles vous aurez facilement accès lors de vos analyses, comme le résultat intermédiaire nécessaire pour répondre à une question par exemple.

Le livrable demandé est composé d'un compte-rendu et du code Spark. Si vous avez des graphiques, ceux-ci doivent êtres inclus dans le compte-rendu ou à part dans un des formats suivants : JPG, PNG, SVG, EPS, ou PDF. Le code Spark peut être fournis sous forme de lien vers un dépôt git, une archive au format ZIP contenant les scripts Scala. Le code doit pouvoir être exécuté tel quel sur les données fraîchement téléchargées, si vous avez des phases de pré-traitement (nettoyage des données erronées, etc.), pensez à inclure ces scripts.

2 Quelques conseils

- Pensez à d'abord travailler sur un échantillon, avant de passer sur les données complètes.
- Faites attention aux données erronées ou incomplètes.
- Pensez à noter en commentaire l'état de vos RDDs : que contiennent-ils à un moment donné ? Ça vous évitera bien des erreurs et vous facilitera les corrections si besoin.
- Avant de vous lancer dans du code, vous pouvez travailler quelques minutes sur papier pour identifier les différentes étapes et traitements à effectuer.

3 Exemples de jeu de données

Des exemples de sites recensant des jeux de données :

- Kaggle (compte requis)

— Data.gouv
— SNAP
— Awesome Public Datasets

Des exemples de jeux de données, ainsi que de questions auxquelles vous pouvez répondre à partir de celles-ci :

— **Accidents Corporels en France :** Des statistiques sur les endroits, les types de véhicules impliqués, etc. Evolution au fil du temps des accidents corporels ? (baisse ou augmentation ? Qu'en est-il par type d'accident ? Par catégorie de route ?) PUBG second dataset : chaque élimination de joueur. Quelles sont les armes utilisées par les meilleurs / pires joueurs ? Quelles sont les zones les plus meurtrières de chaque carte ? Quelle est la distance effective de chaque arme ? (corps à corps, courte, moyenne, longue distance)
— **Recherches AOL :** fichier de journal du défunt moteur de recherche. Quels sont les mots les plus recherchés ? Etant donné un mot, quels sont les autres mots auxquels celui-ci est le plus couramment associé ?
— **MovieLens :** notation de films par des utilisateurs. Quels sont les films les plus appréciés ? Et les moins appréciés ? Qu'en est-il par genre ou catégorie de film ? Quels sont les genres les plus appréciés de chaque utilisateur ? Quels films recommander à un utilisateur ?
— **Adult Dataset :** données sur une population d'adulte et de leurs revenus. Quels sont les facteurs qui influencent le plus le salaire ? Quelles sont les catégories qui gagnent le plus (métier, éducation, etc.) ?