Mini-projet de DMB

Ce mini-projet est à effectuer seul, en binôme ou en trinôme. Un total de **4 heures** de TP sont bloquées pour sa réalisation, et le compte-rendu ainsi que le code Spark sont à rendre pour le **17/02**. Le rendu se fait par mail à l'adresse tompoariniaina.andriamilanto@irisa.fr, en mettant zoltan.miklos@irisa.fr en copie.

Le **jeu de données**, dont dépendra les questions que vous vous poserez, est laissé libre. 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**. Le compte-rendu peut prendre la forme d'un notebook Jupyter, d'un rapport au format PDF, ou d'un simple fichier texte. 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 Python, ou un notebook. Le code doit pouvoir être **exécuté tel quel** sur les données fraichement téléchargées, si vous avez des phases de pré-traitement (nettoyage des données erronées, etc.), pensez à inclure ces scripts.

Quelques conseils

- Un **découpage suggéré des 4h de TPs** : 45 minutes pour récupérer le jeu de données et préparer les questions, 2h30 pour développer le code Spark et récupérer les résultats, et 45 minutes pour préparer le compte-rendu.
- Prenez un jeu de données au **format textuel et simple à manipuler**, afin d'éviter de passer trop de temps sur des phases de pré-traitement.
- N'hésitez pas à **utiliser la persistance**, mais de manière raisonnée (ne persistez pas 10Go de données brutes, si vous ne possédez que 4Go de mémoire et que vous n'avez besoin que d'une part négligeable de celles-ci).
- 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 ? Ca vous évitera bien des erreurs et vous facilitera les corrections si besoin est.

- Avant de vous lancer dans du code, vous pouvez travailler quelques minutes sur papier pour identifier les différentes étapes et traitements à effectuer.
- Si vous ne connaissez aucune bibliothèque de génération de graphiques, voici quelques-unes en Python : matplotlib, plotly, et seaborn.

Barème de notation

- **Jeu de données** (cohérence, taille, complexité) : 2 points
- **Code Spark** : 8 points
  - Fonctionnel : 3 points
  - Utilisation correcte de Spark : 3 points
  - Chargement et sauvegarde des données : 1 point
  - Performance et optimisations : 1 point
- **Analyse** (qualité des questions, résultats fournis, interprétation) : 8 points
  - Qualité des questions : 2 points
  - Résultats fournis (répondent aux questions, complexité, varité) : 4 points
  - Interprétation des résultats : 2 points
- **Compte-rendu** (forme, lisibilité des résultats) : 2 points

Exemples de jeu de données

Des exemples de sites recensant des jeux de données :

- Kaggle (compte requis)
- Data.gouv
- Awesome Public Datasets
- Cool 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.) ?
- **Mots de passe** (demander au chargé de TP) : différents jeux de données de mots de passe.
    - Quels sont les mots de passe les plus utilisés ?
    - Qu'en est-il des suites de lettres ou de chiffres ?
    - Quel est le format de mot de passe le plus courant ?
    - Générer un dictionnaire d'attaque composé des milles mots de passe les plus courants, avec leur hashé dans un algorithme choisis.
- SpamAssassin : mails classés en spam ou non.
    - Quels sont les indicateurs d'un mail de spam ? (pistes : mots, urls, métadonnées)

Andriamilanto Tompoariniaina (tompoariniaina.andriamilanto@irisa.fr) 2019 - 2020