Projet

Consignes

- Projet à réaliser en Python / PySpark 3.1
- Votre code devra tourner le plus rapidement possible.
- Vous devrez justifier des choix d'implémentation et démontrer en quoi votre code a été optimisé.

Rendu

- Votre code source (repo Git ou archive).
- La commande permettant de lancer votre code.

Récupérez le dataset **full.csv** du projet GitHub Commit Messages sur Kaggle.

Votre application Spark devra effectuer les actions suivantes sur ce dataset :

1. Afficher dans la console les 10 projets Github pour lesquels il y a eu le plus de commit.
2. Afficher dans la console le plus gros contributeur (la personne qui a fait le plus de commit) du projet apache/spark.
3. Afficher dans la console les plus gros contributeurs du projet apache/spark sur les 6 derniers mois. Le code doit être générique, si on le relance dans 6 mois il devra donner les plus gros contributeurs des 6 prochains mois, pas de date en dur dans le code 😉. Pour la conversion vous pouvez vous référer à cette documentation.
4. Afficher dans la console les 10 mots qui reviennent le plus dans les messages de commit sur l'ensemble des projets. Vous prendrez soin d'éliminer de la liste les stopwords pour ne pas les prendre en compte. Vous êtes libre d'utiliser votre propre liste de stopwords, vous pouvez sinon trouver des listes ici.

Soutenance

Lors de la soutenance le code sera lancé avec un sleep pour que vous puissiez accéder à la Spark UI.

Vous devrez présenter

1. Votre code source et expliquer vos choix d'implémentation.
2. La Spark UI et expliquer les métriques qui sont affichées concernant votre application.