Projet

Consignes

- Projet à réaliser en Python / PySpark
- 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 4 dernières années. 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

Vous devrez présenter votre code source et expliquer vos choix d'implémentation. Le code sera exécuté avec un **sleep** afin de pouvoir naviguer à travers la spark UI et expliquer les différents éléments.