M2 TIDE, 2024-2025

Big Data Analytics

Projet comptant pour l'évaluation de la session 1

Exercice 1 :

On souhaite afficher pour les utilisateurs d'un réseau social le nombre d'amis en commun avec un autre utilisateur quand il visite la page de ce dernier.

1. Écrivez sous python, un programme de type MapReduce qui calcule le nombre d'amis en communs pour chaque paire d'utilisateurs sachant que le fichier amis.txt contient les identifiants des utilisateurs suivis des identifiants de leurs amis.

2. Ré-ecrirez votre programme MapReduce en utilisant Pyspark.

NB : Un programme de type MapReduce comprend deux phases : une phase de map où vous auriez en sortie des données sous forme clé-valeur, et une phase d'aggrégation. Vous pouvez vous inspirer des indications en fin de chapitre 2.

Exercice 2 :

Ecrire un programme qui prend en entrée un fichier contenant une liste de liens (url) de pages web et un nombre entier k, puis, collecte des informations/données et renvoie les k mots les plus fréquents sur chaque page web.

Pour illustrer, vous supposerez que les données recherchées sont les adresses email valident de chaque page. Une fois collectées, votre programme stockera ces données ainsi que les k mots les plus fréquents sur chaque page web, dans un fichier se trouvant sur le disque dur.

Lorsque cela est possible, l'utilisation du MapReduce sur Spark ou une procédure du type MapRe-duce sur Python est vivement recommandée.

Exercice 3 :

Le département du crédit à la consommation d'une banque souhaite automatiser le processus de prise de décision pour l'approbation des lignes de crédit hypothécaires. Pour ce faire, ils suivront les recommandations de l'Equal Credit Opportunity Act afin de créer un modèle de score de crédit empiriquement dérivé et statistiquement solide. Ce modèle sera basé sur les données collectées auprès des candidats récents ayant obtenu un crédit par le biais du processus actuel d'octroi de prêts. L'objectif est de construire un modèle suffisamment interprétable pour fournir une justification en cas de décisions défavorables (rejets).

Le jeu de données Home Equity (HMEQ) contient des informations de base et de performance des prêts pour 5 960 prêts hypothécaires récents. La variable cible (BAD) est binaire et indique si un candidat a finalement fait défaut ou a été gravement en retard dans ses paiements. Ce résultat négatif s'est produit dans 1 189 cas (20%). Pour chaque candidat, 12 variables d'entrée ont été enregistrées.

1. Traitez les valeurs manquantes de cette table de données.

2. Proposer à la banque un algorithme de prédiction. Vous pourrez mettre en oeuvre 3 modèles et sélectionner le meilleur sur votre base de test.

Pour cet exercice, l'utilisation de Sagemaker est très recommandé (obligatoire).

Consignes

— Votre rendu sera un dossier compressé (Zip) contenant un notebook (jupyter ou vs code) avec obligatoirement les sorties des cellules où il y'a du code et un fichier pdf décrivant votre approche/méthodologie, étapes etc.

— La qualité du code sera prise en compte dans l'évaluation de votre projet. Les commentaires devrait accompagner votre code.

— Pour toute question sur la réalisation du projet, veuillez m'écrire à l'adresse kamilakare@gmail.com.

— Votre projet sera rendu au plus tard le **vendredi 2 mai 2024** à l'adresse sus-indiquée.