ESIREM
ÉCOLE SUPÉRIEURE D'INGÉNIEURS NUMÉRIQUE ET MATÉRIAUX

Contrôle final : Big Data. 11 janvier 2021

Durée 2h, tous documents non électroniques autorisés.

Le barème est donné à titre indicatif.

---

1. (2 points) Donnez la définition formelle de la notation asymptotique "grand-O".
2. (2 points) Un algorithme de complexité $O(n!)$ peut-il fonctionner plus rapidement qu'un autre algorithme de complexité $O(n)$? Si oui, dans quel cas ? Sinon, expliquez pourquoi.
3. (1 point) Qu'est-ce qui n'est pas un système de traitement de grandes quantités de données ?
   A. Apache Hadoop
   B. Map Reduce chez Google
   C. Apache Spark
   D. LibreOffice Calc
   E. Apache HTTP Server
4. (5 points) *Bibliothèque à sens unique* est une structure de fichiers permettant d'effectuer seulement deux opérations :
   - `add (F)` : enregistrer le fichier F ;
   - `inside? (F)` : c'est une fonction booléenne répondant correctement, avec une haute probabilité, à la question "Le fichier F, a-t-il déjà été enregistré précédemment ?"

   Supposons que la taille moyenne d'un fichier est 4Mo, et que vous avez une fonction de hachage produisant des hashs de longueur 512 bit. Quelle méthode suggérez-vous pour la construction d'une bibliothèque à sens unique afin qu'elle puisse enregistrer le nombre maximal de fichiers ? Combien de fichiers peuvent être enregistrés en utilisant cette bibliothèque sur un disque dur de 64Mo ?

   Rappel :

   > 1Ko c'est 1024 octets.
   > 1Mo c'est 1024 Ko.
   > 1Go c'est 1024 Mo

5. (5 points) **K-means.** Un jeu de données contient 4 points. La matrice de données est la suivante

$$\begin{pmatrix} 1 & 3 \\ 5 & 7 \\ 7 & 5 \\ 3 & 1 \end{pmatrix}.$$

   6.1. Calculez les coordonnées des centres optimaux, dessinez les clusters pour les valeurs de k suivantes :
   - $k = 1$,
   - $k = 2$.

   6.2. Illustrez les itérations de l'algorithme "k-means" lorsque les centres initiaux ont les coordonnées : $(2, 6)$ et $(6, 3)$.

6. (5 points) **PCA.** Un jeu de données contient 5 points. La matrice de données est la suivante $\begin{pmatrix} 1 & 2 & 10 \\ 2 & 4 & 10 \\ 312 & 624 & 10 \\ 11 & 22 & 10 \\ 12 & 24 & 10 \end{pmatrix}$.

   Illustrez les projections de données, 3D $\rightarrow$ 1D, sur :
   - la première composante principale ;
   - la deuxième composante principale.