Institut Supérieur d'Informatique et de Mathématiques de Monastir ISIMM

Examen -Session Principale– S1 – 2023/2024

Filière : *Ing2_Info*	Matière : **Bases de données NoSQL et Big Data**		Enseignante : **Asma KERKENI**
Date : 12 / 01 / 2024	Nbr de Crédits : **3**	Coefficient : **3**	**Documents autorisés** **Calculatrice autorisée**
Durée de l'examen : 1h30	Régime d'évaluation : **Mixte** EX (45%) + DS (22%) + TP (33%)		**Nombre de pages : 4**

Exercice 1 : (10 points)

Cet exercice vise à appliquer Hadoop MapReduce pour détecter les co-occurrences de mots dans un corpus de texte en anglais, en utilisant l'approche "***Stripes***". Les co-occurrences de mots représentent la fréquence à laquelle certains mots apparaissent simultanément dans un même document.

Contrairement à d'autres approches qui utilisent des matrices, ***"Stripes"*** représente les relations entre les mots en stockant les occurrences conjointes de chaque mot avec ses voisins. L'objectif est de générer des associations de mots fréquemment rencontrés pour mieux appréhender les relations entre ces mots.

Exemple :

Considérons le texte suivant comme exemple :

```
Hadoop is used for big data processing.
Big data requires efficient processing techniques and frameworks.
Hadoop and Spark are big data frameworks.
```

Le résultat souhaité est le suivant :

```
big {'data': 3, 'frameworks': 1, 'processing': 1, 'requires': 1, 'efficient': 1}
data {'frameworks': 1, 'processing': 2, 'requires': 1, 'efficient': 1}
efficient {'processing': 1, 'techniques': 1, 'frameworks': 1}
hadoop {'spark': 1, 'big': 2, 'data': 2, 'used': 1}
processing {'techniques': 1, 'frameworks': 1}
 requires {'efficient': 1, 'processing': 1, 'techniques': 1}
```

```
spark {'big': 1, 'data': 1, 'frameworks': 1}
techniques {'frameworks': 1}
used {'big': 1, 'data': 1, 'processing': 1}
```

Questions :

1. Implémentez le code du Mapper pour résoudre ce problème. Vous pouvez utiliser la liste suivante pour les mots vides (à ignorer) :

   "the", "and", "are", "for", "is", "in", "to", "of", "a", "an", "by", "with"
2. Donnez le code du Reducer permettant d'afficher les associations de mots les plus fréquentes. Chaque mot doit être associé à d'autres mots avec lesquels il co-occurre fréquemment, avec une fréquence supérieure ou égale à 1.
3. Écrivez la (ou les) commande(s) nécessaire(s) pour exécuter ce job sur le fichier ***"texte.txt"*** situé à l'emplacement ***"/home/utilisateur/documents/texte.txt"*** sur le système de fichiers local, sur un cluster Hadoop en utilisant le fichier JAR ***"hadoop-streaming.jar"*** situé dans le répertoire ***$HADOOP_HOME***. Veillez à inclure les options appropriées pour spécifier les fichiers d'entrée, les fichiers de sortie, le mapper et le reducer.

Exercice 2 : (10 points)

La figure en annexe est un extrait d'un collection nommée **Paris**. Elle contient des lieux de Paris qui ont été agrégés sur le site tourPedia. Vous y trouverez différents catégories, et informations pour les lieux :

- Des POI (points d'intérêts) ;
- Des restaurants ;
- Des logements (accommodation), avec les services associés ;
- Des attractions ;
- A chaque lieu sera associé des commentaires sur internet (Facebook, Foursquare), et des notes utilisateurs.

Ecrivez les requêtes sur la collection Paris permettant de :

1. Donner les noms et notes des lieux ayant au moins une note (reviews.rating) supérieur à 4.
2. Donner la langue et note des commentaires (reviews) de lieux, contenant au moins un commentaire écrit en anglais ("language" :"en") avec une note supérieure à 3.

3. Donner les noms et contacts "Foursquare" et "Website" des lieux ayant des URLs contacts renseignées de type "Foursquare" et "GooglePlaces".
4. Donner les adresses des lieux dont le nom contient le mot "hotel" (en minuscule ou majuscule).
5. Ajouter aux lieux un tableau contenant la liste de toutes les langues utilisées dans les commentaires (clé : Langues).
6. Pour chaque nom de lieu de catégorie "poi", donner le nombre de commentaires dont la source (reviews.source) est "Facebook" et les trier par ordre décroissant de ce nombre.
7. Pour chaque langue utilisée dans les commentaires, trouver le nombre total de commentaires donné par cette langue et trier les résultats par ordre décroissant de ce nombre.
8. Pour chaque catégorie de lieux, donner le nombre de commentaires par langue.
9. Donner le nombre de langues de commentaires différents par type de source.
10. Donner le nombre total de commentaires par mois dans l'année 2012.

Annexe


```
{
    "_id" : 83419,
    "contact" : {
        "GooglePlaces" : null,
        "Foursquare" : "https://foursquare.com/v/pe%C3%B1a-festayre-paris-%C
    },
    "name" : "Peña Festayre",
    "location" : {
        "city" : "Paris",
        "coord" : {
            "coordinates" : [
                2.3860357589657,
                48.896621743257
            ],
            "type" : "Point"
        },
        "address" : "80 Boulevard Macdonald"
    },
    "category" : "restaurant",
    "description" : ""
    "services" :[ "jardin", "terrasse", "journaux"]
    "reviews" : [
        {
            "wordsCount" : 6,
            "rating" : 0,
            "language" : "ca",
            "details" : "http://tour-pedia.org/api/getReviewDetails?id=52a74
            "source" : "Foursquare",
            "text" : "Entree + Plat (grillade a volonte) + dessert 19€",
            "time" : "2011-01-22",
            "polarity" : 0
        },
        {
            "wordsCount" : 20,
            "rating" : 0,
            "language" : "fr",
            "details" : "http://tour-pedia.org/api/getReviewDetails?id=52a74a85ae9eef5a50671b09",
            "source" : "Foursquare",
            "text" : "Tous les mercredis jusqu'en juin : Soirées Salsa... concert live puis dj ;
            "time" : "2012-01-11",
            "polarity" : 5
        }
    ],
    "likes" : {},
    "nbReviews" : 2
}
```


Bon courage !