Les RDD Exercices

Exercice 1 : Apprendre à manipuler les RDD's

Programmation des RDD's avancées ¼

- Map
  - Explication /objectif de la fonction
  - Application
    - Créer un fichier .txt
    - Ecrire une phrase en minuscule
    - Appliquer via la fonction map une fonction upper sur le fichier afin de mettre le texte en majuscule.
- FlatMap
  - Explication /objectif de la fonction
    - Quelle est la différence entre la fonction Map et la fonction flatMap ?
  - Application
    - Comment appliquer le flatMap sur la liste suivante pour obtenir une liste de chaque mot.
      - ["mot1 mot2 ", "mot3   mot4 "," mot5   mot6 "]
      - Appliquer un map puis un flatMap pour effectuer un comparatif des fonctions.
- Filter
  - Explication /objectif de la fonction
  - Application :
    - Chercher les prénoms composés dans cette liste à l'aide de la méthode filter:
      - ["Pierre", "Alpha", "Mehdi", "Jean-Pierre", "Thierry"]
- Distinct
  - Explication /objectif de la fonction
  - Application
      - Supprimer les doublons de la liste suivante : ["Pierre", "Alpha", "Mehdi", "Jean-Pierre", "Thierry", "Pierre"]

- GroupBy
  - Explication /objectif de la fonction
  - Application
    - Regrouper sur chaque premier élément de chaque caractère de la liste suivante :
      ["Pierre", "Alpha", "Mehdi", "Jean-Pierre", "Thierry", "Pascal"]
- Sample
  - Explication /objectif de la fonction
  - Application
    - Prendre un échantillon de 50% sur une tranche de 1 à 999999
    - Afficher le ombre d'échantillon

Programmation des Rdds avancées 2/4

- Union
  - Explication /objectif de la fonction
  - Application
    - Créer 2 RDDs composés de 2 listes de nombre entiers
      - liste1 = [1, 2, 3]
      - liste2 = [3, 4, 5]
    - Faire l'union de ces 2 RDDs

- Intersection
  - Explication /objectif de la fonction
  - Application
    - Créer 2 RDDs composés de 2 listes de nombre entiers
      - liste1 = [1, 2, 3]
      - liste2 = [3, 4, 5]
    - Trouver l'intersection de ces 2 RDDs

- Substract
  - Explication /objectif de la fonction
  - Application
    - Créer 2 RDDs composés de 2 listes de nombre entiers
      - liste1 = [1, 2, 3]
      - liste2 = [3, 4, 5]
    - Faire le substract de ces 2 RDDs

- Cartesian
  - Explication /objectif de la fonction
  - Application
    - Créer 2 RDDs composés de 2 listes de nombre entiers
      - liste1 = [1, 2, 3]
      - liste2 = [3, 4, 5]
    - Faire le cartesian de ces 2 RDDs

- Reduce
  - Explication /objectif de la fonction
  - Application
    - Créer 1 RDDs composés d'1 liste issue d'un range de 1 à 6
    - Appliquer un reduce pour additionner x + y.
    - Quelle est la particularité de la fonction reduce ?

Programmation des Rddds avancées 3/4

- Création d'un pair rdd : map() et keyBy()
    - Explication /objectif de la fonction
    - Application :
        - Soit la liste suivante :
            - ["cle1 valeur1", "cle2 valeur2", "cle3 valeur3"]
            - Extraire la clé à l'aide de la fonction map
            - Extraire la clé à l'aide de la fonction keyBy

- groupByKey
    - Explication /objectif de la fonction
    - Application
        - Soit la liste suivante :
            - ["0, 11", "1, 11", "0, 4", "2, 8", "1, 1", "9, 8"]
            - Effectuer un regroupement par clé à l'aide de la fonction map et groupByKey.

- reduceByKey
    - Explication /objectif de la fonction
    - Application
        - Soit la liste suivante :
            - ["0, 11", "1, 11", "0, 4", "2, 8", "1, 1", "9, 8"]
            - Faire la somme des clés qui sont identiques à l'aide des fonctions map et reduceByKey

- mapValues
    - Explication /objectif de la fonction
    - Application
        - Soit la liste suivante :
            - ["0, 11", "1, 11", "0, 4", "2, 8", "1, 1", "9, 8"]
            - Appliquer la valeurs au carré à l'aide des fonctions map et mapValues

- Keys/values
    - Explication /objectif de la fonction
    - Application
    - A partir des résultats derniers, récupérer les clés du RDD.
    - A partir des résultats derniers, récupérer les valeurs du RDD

- sortByKey
    - Explication /objectif de la fonction

- Application
  - Soit la liste suivante :
    - ["0, 11", "1, 11", "0, 4", "2, 8", "1, 1", "9, 8"]
    - Trier la liste par clé

Programmation des Rdds avancés 4/4

- join
  - Explication /objectif de la fonction
  - Application
    - Créer un RDD à partir de la liste suivante :
      - ["a, 1", "b, 10", "c, 4"]
      - Tout en créant un tuple clé, valeur.
    - Créer un second RDD à partir de la liste suivante :
      - ["d, 6", "e, 1", "a, 9"]
      - Tout en créant un tuple clé, valeur.
    - Récupérant les éléments de la clé commune

- rightOuterJoin
  - Explication /objectif de la fonction
  - Application
    - Créer un RDD à partir de la liste suivante :
      - ["a, 1", "b, 10", "c, 4"]
      - Tout en créant un tuple clé, valeur.
    - Créer un second RDD à partir de la liste suivante :
      - ["d, 6", "e, 1", "a, 9"]
      - Tout en créant un tuple clé, valeur.
    - Récupérer les éléments commun de l'ensemble des clés du RDD 2

- leftOuterJoin
  - Explication /objectif de la fonction
  - Application
    - Créer un RDD à partir de la liste suivante :
      - ["a, 1", "b, 10", "c, 4"]
      - Tout en créant un tuple clé, valeur.
    - Créer un second RDD à partir de la liste suivante :
      - ["d, 6", "e, 1", "a, 9"]
      - Tout en créant un tuple clé, valeur.
    - Récupérant les éléments commun de l'ensemble des clés du RDD 1

- fullOuterJoin
    - Explication /objectif de la fonction
    - Application
        - Créer un RDD à partir de la liste suivante :
            - ["a, 1", "b, 10", "c, 4"]
            - Tout en créant un tuple clé, valeur.
        - Créer un second RDD à partir de la liste suivante :
            - ["d, 6", "e, 1", "a, 9"]
            - Tout en créant un tuple clé, valeur.

        - Récupérant les éléments commun de l'ensemble des clés du RDD 1 et du RDD2

Exercice 2 :

Fichier à traiter : fakefriends.csv

A partir du fichier fakefriends.csv qui correspond à des données issues d'un réseau social fictif, déterminer le nombre moyen d'amis ventilé par âge des personnes dans ce réseau social.

Nous allons utiliser des paires clé/Valeur dans le RDD pour ce faire.

Exercice 3 :

Fichier à traiter : 1800.csv

Traitement à faire :

- Récupérer la température MAX avec son identifiant de station
- Récupérer la température MIN avec son identifiant de station
- Récupérer l'identifiant de station avec les 2 température MIN et MAX

Exercice 4 :

Fichier à traiter : Lorem_Ipsum.txt

A partir du fichier Lorem_Ipsum.txt. Compter les occurrences de mots à l'aide de flatMap.