TP1 Apache Spark

Lire un fichier de données non structurées

Charger le fichier README.md

```
scala> val textFile = spark.read.textFile("path/to/README.md")
```

Afficher dans le terminal les cinq premières lignes du fichier

```
scala> textFile.take(5).foreach(println)
```

Faire un WordCount sur le README.md du dossier spark et afficher les résultats

```
scala> val counts = textFile.flatMap(line => line.split(" "))
                 .map(word => (word, 1))
                 .reduceByKey(_ + _)
scala> counts.saveAsTextFile("hdfs://…")

scala> val output = counts.collect()
scala> println(output)
```

Afficher les résultats du WordCount sous forme du table, avec les mots les plus fréquents d'abord

```
scala> val df = counts.toDF()
scala> val df_renamed = df.toDF("Word", "Count")
scala> df_renamed = df_renamed.sort(desc("Count"))
scala> df_renamed.show()

+---------+-----+
|     Word|Count|
+---------+-----+
|          |   67|
|       the|   21|
|        to|   14|
|     Spark|   13|
|       for|   11|
|       and|   10|
|         a|    8|
|        ##|    8|
|       run|    7|
|       can|    6|
|        is|    6|
|        on|    5|
|        in|    5|
|        of|    5|
|        if|    4|
|      also|    4|
|       you|    4|
|   Hadoop|    3|
|including|    3|
|        an|    3|
+---------+-----+
only showing top 20 rows
```

Un même mot peut être présent dans la table précédente avec des majuscules et des minuscules. Mettre les mots en minuscule.

```
scala> df_renamed = df_renamed.select(lower($"Word"), $"Count")
scala> df_renamed.show()

+-----------+-----+
|lower(Word)|Count|
+-----------+-----+
|            |   67|
|         the|   21|
|          to|   14|
|       spark|   13|
|         for|   11|
|         and|   10|
|           a|    8|
|          ##|    8|
|         run|    7|
|         can|    6|
|          is|    6|
|          on|    5|
|          in|    5|
|          of|    5|
|          if|    4|
|        also|    4|
|         you|    4|
|      hadoop|    3|
|  including|    3|
|          an|    3|
+-----------+-----+
only showing top 20 rows
```

Sommer la valeur count pour un même mot mis en minuscule, et afficher les résultats sous forme de table.

```
scala> var df_toLowerCase = df_renamed.groupBy("lower(Word)").sum("Count")
scala> df_toLowerCase = df_toLowerCase.sort(desc("sum(Count)"))
scala> df_toLowerCase.show()

+-------------+----------+
|  lower(Word)|sum(Count)|
+-------------+----------+
|              |        67|
|           the|        22|
|            to|        16|
|           for|        13|
|         spark|        13|
|           and|        11|
|             a|         9|
|            ##|         8|
|           you|         7|
|           run|         7|
|           can|         6|
|            is|         6|
|            in|         5|
|            of|         5|
|            on|         5|
|            if|         4|
|documentation|         4|
|       example|         4|
|          also|         4|
|          with|         3|
+-------------+----------+
only showing top 20 rows
```