Práctica 2: HADOOP Y MAPREDUCE

ENUNCIADO

En esta práctica se ha proporcionado un conjunto de libros organizados en carpetas por autores (véase libros.zip). Cada uno de estos libros se puede procesar mediante un algoritmo map-reduce que sea capaz de calcular la longitud media de las palabras que contiene: es decir, tendríamos una característica propia de cada libro (longitud media de palabras y su desviación estándar).

Si partimos de la asunción que cada escritor tiene un gusto particular por el uso en sus textos de un tipo de palabras (más largas o más cortas), podríamos intentar clasificar cualquier libro utilizando estas características.

Se propone como tarea que se procese toda la colección de libros proporcionados para obtener de cada uno de ellos la longitud media y su desviación. Se proporciona para ellos dos ficheros que contienen el código fuente de la fase map (mapper.py en Python) de la fase reduce (reducer.awk en el lenguaje awk).

Un ejemplo sencillo de ejecución de la aplicación MapReduce en hadoop se muestra a continuación:

```
hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar -input
practica2/libros/Bailen.txt -output practica2/Bailenstats -file ./mapper.py
-file ./statsreducer.awk -mapper ./mapper.py -reducer ./statsreducer.awk
```

Se debe crear una tabla resumen que contenga, para cada libro, los siguientes datos: autor, titulo, longitud media y desviación estándar.

Indica qué algoritmo de clasificación podrías utilizar para corroborar, o no, la hipótesis expuesta: *"es posible clasificar el estilo de un autor a través de estos valores: longitud media de palabras y su desviación estándar".*