CURSO: CMP 0575 - TÓPICOS 2 (DATA MINING)
COLEGIO: POLITÉCNICO
Semestre: Primer Semestre 2019/2020

Proyecto 1: Ejercicio usando la técnica ***MapReduce***

Problema:

1. Se desea implementar un programa que permita la lectura de un fichero (Word, texto) y devuelva un índice de palabras populares empleadas en el documento (la estructura del índice tiene que ser en el formato <*palabra, frecuencia*>).

Esquema de trabajo:

- Equipo de 2 estudiantes.

Requisitos:

- Es obligatorio el uso de la filosofía ***MapReduce*** optimizada, en este caso sobre una arquitectura simple, pero, paralelizable.
- Cargar al D2L los códigos implementados (fichero compactado) dentro del plazo de entrega de las distintas fases de evaluación.

Requisitos funcionales de la técnica:

- El documento debe ser dividido en ficheros de 25 líneas (secuencial).
- La cantidad de nodos puede ser aleatoria, pero, con un mínimo de 6 nodos ***map*** y 2 ***reduce***.

Evaluación:

- Fase 1:
  - Aplicar correctamente el ***MapReduce*** sobre un fichero de entrada y obtener la salida que da solución a la problemática planteada. (40% de la nota final de la tarea)
- Fase 2:
  - Aplicar elementos de paralelización a nivel de tareas ***Map*** y ***Reduce***, y obtener la salida que da solución a la tarea asignada. (30% de la nota final de la tarea)
- Fase 3:
  - Aplicar y corregir fallos a los distintos nodos: ***coordinator***, ***map, combiner, sort-shuffle*** y ***reduce*** (30% de la nota final de la tarea)
- Trazabilidad obligatoria:
  - En todas las fases se debe imprimir el resultado
- +1
  - Usar la técnica distribuida

Nota: En cada fase de evaluación el profesor aplicará puntos de chequeo sobre el código implementado y basado en la trazabilidad.