Logo

Description automatically generated

UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE CIENCIA DE LA COMPUTACIÓN

Curso: CC332, Programación paralela

Práctica calificada 05

Procesamiento de datos a gran escala

20210279F Calagua Mallqui Jairo Andre

20210035J Pisfil Puicon Angello Jamir

20210164D Salcedo Alvarez Guillermo Ronie

Setiembre 28, 2023

**Resumen**

El trabajo se centra en la Programación Paralela para abordar el Procesamiento de Datos a Gran Escala utilizando un dataset bidimensional. Se emplean medidas como la distancia Hamming, distancia euclidiana, correlación de Pearson y correlación de Spearman para analizar patrones y tendencias en los datos. A medida que las dimensiones del dataset aumentan, se destaca la complejidad computacional, abordada eficientemente mediante la implementación en Java con Threads para aprovechar la computación paralela y mejorar la eficiencia en el manejo de grandes conjuntos de datos.

**Procesamiento de datos a gran escala**

1. **Objetivos**

Se presenta los objetivos para esta práctica:

* Análisis de los algoritmos a utilizar para mostrar análisis de tendencias.
* Implementar las versiones secuencial y paralela para estos algoritmos.
* Análisis de los resultados a través de comparaciones de tiempos.

1. **Análisis teórico de los algoritmos**
   1. **Hamming**

La distancia de Hamming es un coeficiente que mide la diferencia entre dos cadenas de igual longitud contando el número de posiciones en las que los bits o elementos correspondientes difieren.

La fórmula para calcular dicha distancia (H) entre dos cadenas A y B de longitud N es la siguiente:

En donde la función es aquella que es igual a cero cuando los elementos en la posición son iguales, y es uno cuando son iguales.

* 1. **Distancia Euclidiana**

La distancia euclidiana es un coeficiente que mide la longitud del segmento que conecta dos puntos en un espacio multidimensional, tratando los elementos como coordenadas.

La fórmula para calcular dicha distancia (D) entre dos puntos y se halla como:

En donde  y . Luego, este mismo caso se puede generalizar para más dimensiones.

* 1. **Correlación de Pearson**

La correlación de Pearson, vista como el coseno del ángulo, es una herramienta poderosa para cuantificar la relación lineal entre variables continuas. Su aplicación se extiende a diversas áreas, siendo esencial en análisis de regresión, validación de modelos y exploración de la estructura de datos.

La fórmulam que representa dicha correlación es la siguiente:

En donde e son nuestras columnas de los vectores dados.

* 1. **Spearman**

Wa