Tarea 3

INF 564 - DISEÑO AVANZADO DE ALGORITMOS

Formato de entrada

Se entrega un dataset en csv, donde cada fila contiene la siguiente información:

ID_SOURCE, ID_TARGET, RATING

donde:

- ID_SOURCE: ID de usuario origen
- ID_TARGET: ID del usuario destino
- RATING: calificación del trust de ID_SOURCE sobre ID_TARGET, en {-10,10}

Algoritmo k – means

- Defina la función parse_vector para que lea el dataset y cree una RDD o DataFrame en Spark de vectores de usuarios. La representación vectorial de los usuarios corresponde a vectores donde cada componente es el rating. Por ejemplo, para el vector u del usuario i, la componente j corresponde a RATING(i,j). Su RDD de vectores debe tener n-1 slices, donde n es el número de cores de su máquina.
- Defina la función similitud de coseno en los primeros dos cuadrantes del plano cartesiano para que pueda medir la similitud entre dos vectores cuyas componentes tienen valores en el intervalo discreto {-10, 10}.
- Implemente el algoritmo k-means en Spark para que corra sobre el dataset enregado. Puede reusar el código comentado en clases. Extiéndalo para que pueda trabajar sobre los datos de esta tarea.

Experimentación.

- Compute el histograma de similitudes entre los vectores Sim(ui,u.j). Considere 20 bins para su histograma. Usando el histograma, defina un valor adecuado para convDist. Justifique su decisión.
- Corra su algoritmo con los siguientes valores de k: {10, 20, 30, 40, 50}. Corra cada configuración 5 veces, registrando el númer de iteraciones y el tiempo. Para cada valor de k, haga un box plot del tiempo de ejecución y otro del número de iteraciones. Comente sus resultados.

Entrega:

- Escriba un jupyter notebook donde en las primeras celdas pueda invocar a su implementación de k-means. Esto puede hacerlo tanto en pySpark como en scala nativo. Para este último caso, invoque a su programa desde la celda, no es necesario que lo implemente dentro del jupyter. Muestre un gráfico de hilos con el monitor de sistema u otra herramienta que permita visualizar cuantos hilos levantó en su máquina.
- Luego de su live demo, inserte en su jupyter los resultados de la experimentación, incluyendo gráficos y comentarios.
- Coloque su nombre en el jupyter notebook.

Fecha de entrega:

- Subir a moodle el Lunes 22 de Julio, hasta las 9:45 hrs. AM
- Presentar la tarea en clases ese mismo dia, corriendo su jupyter y comentando sus resultados.