Curso de Análisis de Datos Masivos

PCIC, UNAM

Mini-proyecto de la unidad 2: Modelo de mapeo y reducción

Fecha de entrega: Miércoles 09 de octubre.

Formato: Libreta de Jupyter con código documentado y presentación en clase. Usar databricks / Zepellin y Spark.

Forma de entrega: Enviar libreta o liga al correo.

Descripción general

El objetivo de este mini-proyecto es procesar y analizar una colección de documentos usando las funciones de mapeo y reducción en Spark. El estudiante es libre de decir el conjunto de datos que utilizará en el desarrollo del mini-proyecto. El proyecto debe incluir:

- Contar el número de ocurrencias totales de cada palabra, bigrama y trigrama en la colección
- Contar el número de documentos en los que ocurre cada palabra, bigrama y trigrama
- Ordenar las palabras, bigramas y trigramas por su número de ocurrencias totales
- Filtrar las palabras, bigramas y trigramas que ocurren en menos del 5 % de la colección
- Calcular el pesado *tf-idf* para cada palabra, bigrama y trigrama
- Calcular el histograma de ocurrencias por documento de las 10 palabras con mayor ocurrencia total.
- Entrena un modelo de aprendizaje de máquinas usando los datos extraídos (punto extra).

Despliega y visualiza los resultados que obtuviste de estas operaciones y analiza el tiempo de ejecución de las mismas.