Conceptos y aplicaciones en Big Data

Trabajo Práctico 3 – Modalidad de cursada presencial

Spark streaming – Cálculo del índice TF-IDF

Pautas generales

- La entrega consiste en la implementación de un script con Spark streaming, resolviendo todas las consignas presentes en este enunciado. Se deberá entregar el código fuente implementado.
- Los alumnos pueden conformar grupo de no más de dos integrantes y hacer una única entrega grupal.
- La entrega se realiza por la mensajería del curso en IDEAS.
- La fecha límite de entrega es el 27 de diciembre de 2023.
- La calificación obtenida en este TP será tenida en cuenta en la nota final de la materia.

Enunciado

El dataset MovieDataBase contiene 12500 revisiones de películas. Las revisiones (en inglés) son hechas por usuarios del sitio imdb.com. Cada revisión está en su propio archivo.

Basado en el cálculo TF-IDF presentado en el TP2 y el dataset de revisiones de películas publicado con la actividad, implemente una solución con Spark Streaming que calcule el TF-IDF histórico para cada palabra contenida en las revisiones.

El flujo de datos lo puede simular al ir copiando periódicamente cada una de las revisiones en una carpeta previamente configurada como entrada del flujo.