Conceptos y aplicaciones en Big Data

Trabajo Práctico 2 – Modalidad de cursada presencial

Spark – Cálculo del índice TF-IDF

Pautas generales

- La entrega consiste en la implementación de un script con Spark, resolviendo todas las consignas presentes en este enunciado. Se deberá entregar el código fuente implementado y un documento con la comparación solicitada.
- 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 30 de noviembre de 2023.
- La calificación obtenida en este TP será tenida en cuenta en la nota final de la materia.

Enunciado

Basado en el cálculo TF-IDF presentado en el TP1 y el dataset de recetas también del TP1, implemente una solución con Spark que calcule el TF-IDF para cada palabra contenida SOLO en los textos de preparaciones de las recetas de cocina.

Incluya en la entrega un documento explicando las diferencias, en cuanto a complejidad de programación, entre la solución planteada en este TP y la correspondiente al TP1.

Aclaración

Suponga que la cantidad de documentos es Big Data.