# Conceptos y aplicaciones en Big Data

## Trabajo Práctico 1 – Modalidad de cursada presencial

### MapReduce – Cálculo del índice TF-IDF

**Pautas generales**

- La entrega consiste en la implementación de Jobs MapReduce, resolviendo todas las consignas presentes en este enunciado. Se deberá entregar el código fuente implementado y un documento con el DAG del proceso completo.
- 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 31 de octubre de 2023.
- La calificación obtenida en este TP será tenida en cuenta en la nota final de la materia.

**Enunciado**

El índice TF-IDF es un indicador de la relevancia de un término (por ejemplo: una palabra) en un conjunto de documentos. Su cálculo es el siguiente:

TF-IDF(t,d,D) = TF(t,d) . IDF(t,D)

Donde:

TF(t,d) = ft(d) / #d

Donde ft(d) es la cantidad de veces que aparece el término t en el documento d y

#d es la cantidad de palabras total del documento d.

IDF (t, D) = log (#D / D(t) )

Donde #D es la cantidad de documentos totales y

D(t) es la cantidad de documentos que contienen al menos una vez el término t.

Cómo el cálculo de TF-IDF es para un término y para un documento, el cálculo del índice para un término se puede calcular como la suma de todos los TF-IDF calculados para dicho término:

TF-IDF(t, D) = $\sum_{i=1}^{n} TF - IDF(t, d_i(t), D)$

Donde $d_i(t)$ es el i-ésimo documento que contiene el término t.

El archivo Recetas.txt contiene la lista de ingredientes y la preparación de recetas de comida. El archivo tiene dos campos: el UID de la receta y un texto que puede ser, o bien la lista de ingredientes, o bien parte de la preparación.

Para identificar que contiene el texto: Si el mismo empieza con la palabra “Ingredientes:” entonces esa línea solo contiene ingredientes. En cualquier otro caso es parte de la preparación de la receta.

Implemente una solución en MapReduce que calcule el TF-IDF para cada palabra contenida SOLO en los textos de preparaciones de las recetas de cocina.

**Aclaración**

Suponga que la cantidad de documentos es Big Data.

En cada job planteado en la solución piense si una función combiner contribuye o no para la optimización del job. En caso de contribuir implemente dicha función.