Unidad 2: Modelo de mapeo y reducción

Algoritmos

Blanca Vázquez

20 de marzo de 2025

Modelo de programación MapReduce

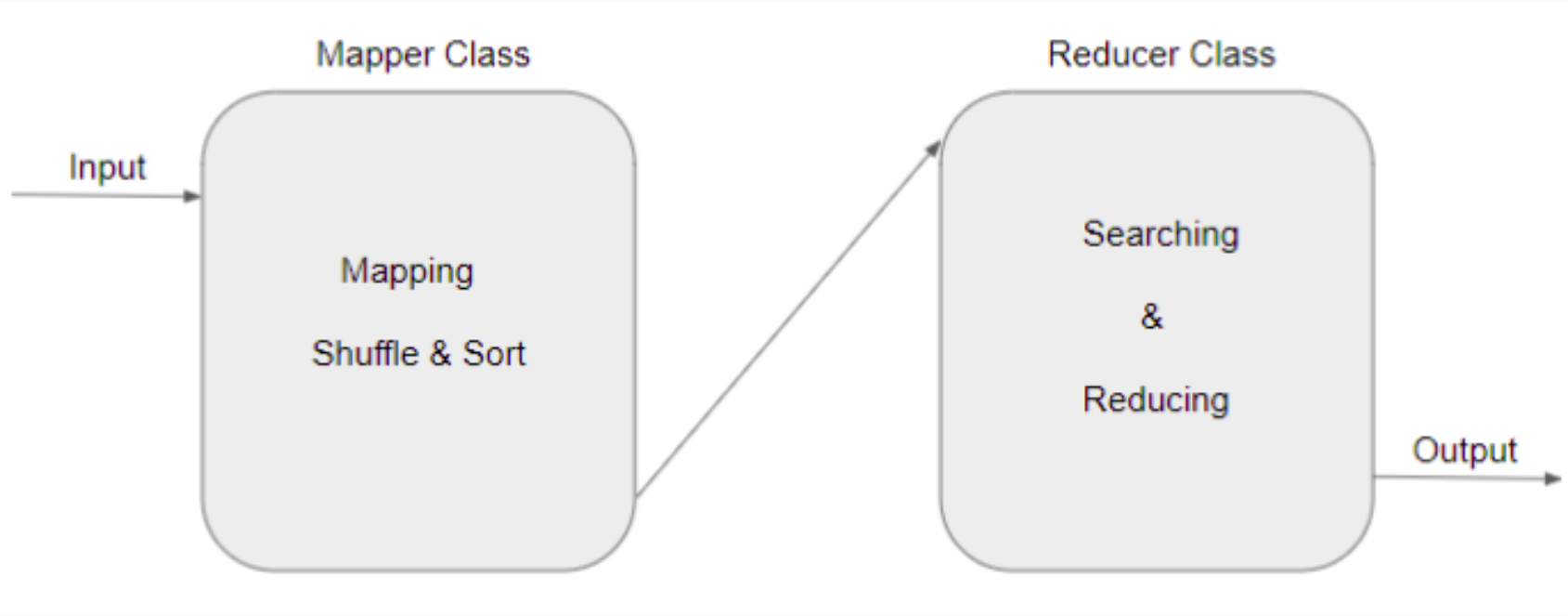


Imagen tomada de MS Minu Sanjudharan

Algoritmos del modelo MapReduce

El modelo de programación implementa diversos algoritmos matemáticos para dividir una tarea en pequeñas partes y para asignarlas a múltiples nodos.

- Ordenamiento
- Búsqueda
- Índice invertido
- TF-IDF (Term Frequency – Inverse Document Frequency)

Ordernamiento - sorting

Este algoritmo organiza datos a partir un criterio de ordenamiento

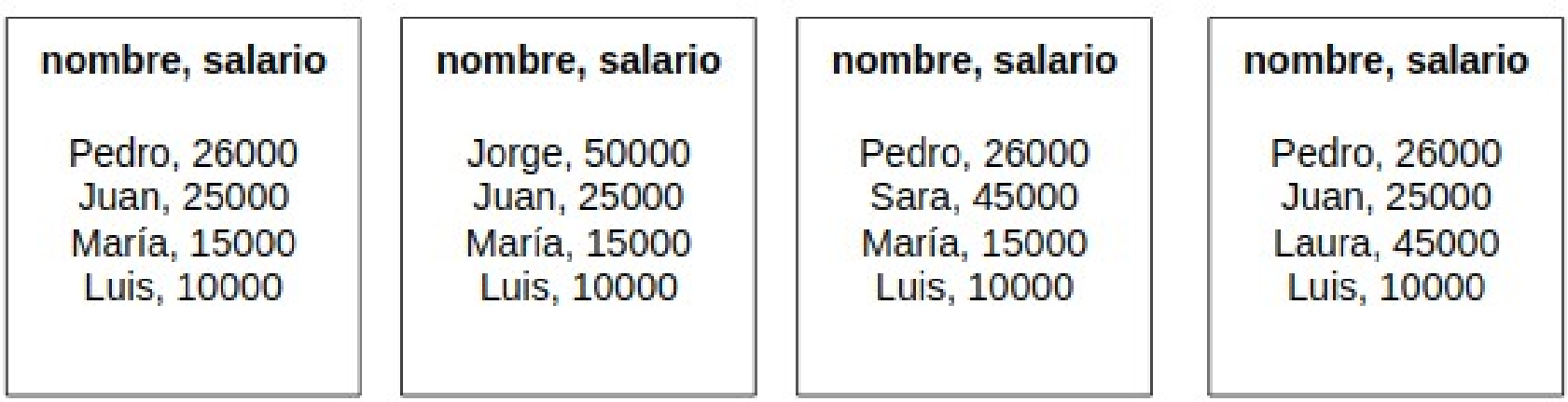
- Este es el algoritmo básico de MapReduce para ordenar los pares llave-valor a partir de sus llaves
- Una vez que los valores han sido tokenizados (<k,v>), la tarea de mapeo usa la clase **RawComparator** para ordenar los pares por llave.

Ordenamiento pares llave - valor

Búsqueda - searching

Este algoritmo está diseñado para localizar un elemento dentro de una estructura de datos.

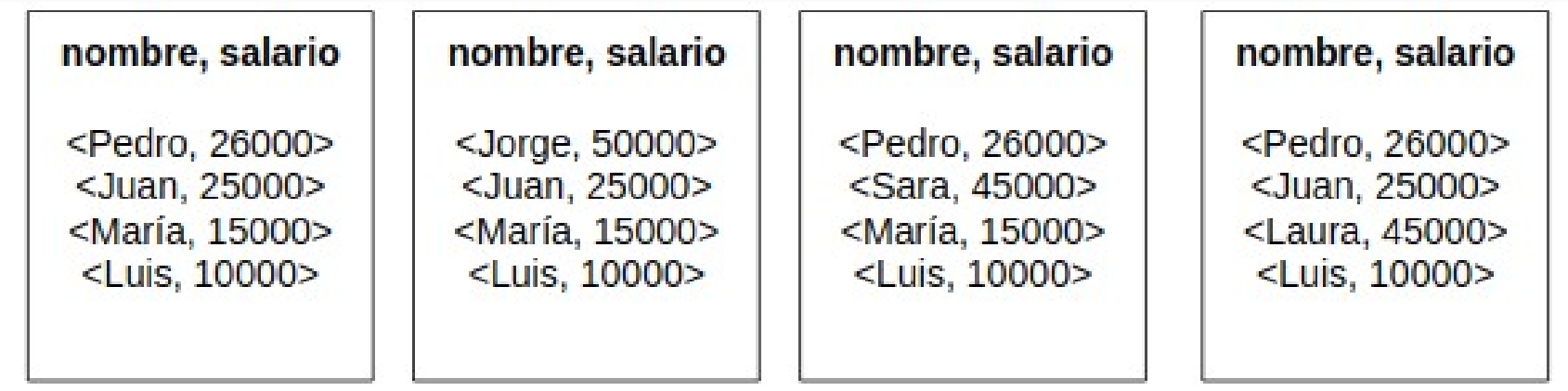
- Este algoritmo participa en la fase de agrupar por llave y en las tareas de reducción.



Localizar el empleado con el salario más alto

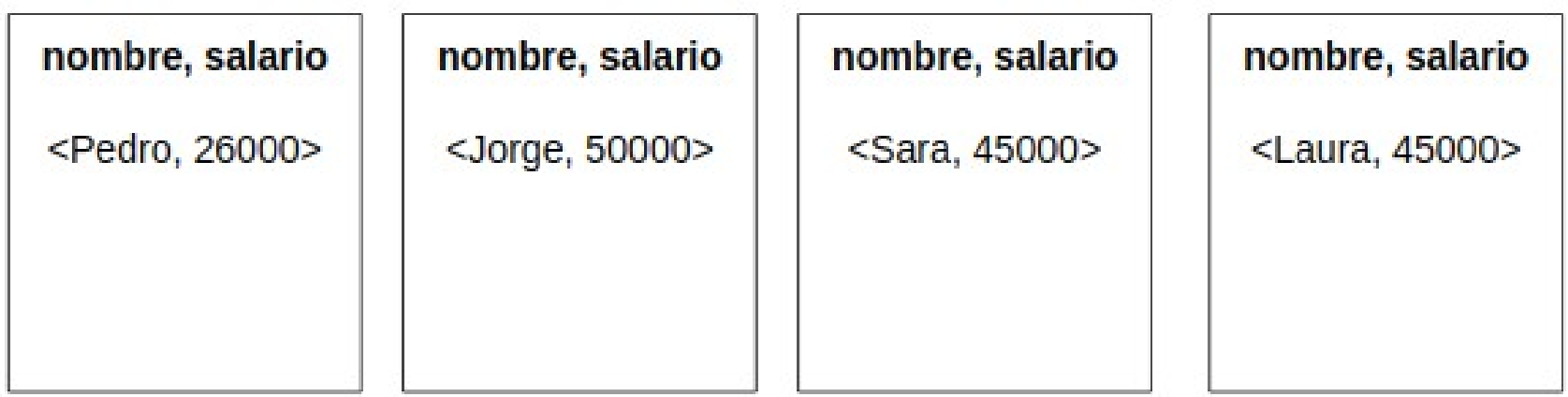
Búsqueda - searching

- La tarea de Mapeo procesa cada entrada de las tablas como un único archivo y asigna los pares (<k,v>: <nombre_empleado, salario>)



Búsqueda - searching

La salida del algoritmo de búsqueda será el salario más alto por cada tabla



- Para evitar la redundancia o duplicados, la tarea de reducción usa el mismo algoritmo de búsqueda sobre los pares <k,v>)
  Salida: <Jorge,50000>

Índice invertido

Este algoritmo es el más utilizado en los sistemas de recuperación de información (ej., motores de búsqueda).

¿Cómo trabaja este algoritmo?

- Para cada término **t**, guardamos todos los documentos que contienen **t**
- Identificamos cada documento por un **docID**, el cuál es un número incremental

Índice invertido

Supongamos que tenemos 3 documentos:

- Doc 1: El perro corre rápido
- Doc 2: El niño juega con el perro en el parque
- Doc 3: El niño nada rápidamente

Índice invertido

Paso 0: preprocesamiento del texto

- Doc 1: el perro corre rapido
- Doc 2: el nino juega con el perro en el parque
- Doc 3: el nino nada rapidamente

Índice invertido

Paso 1: Identificamos cada una de las palabras (k) e indicamos al documento que pertenecen:

Doc 1: el perro corre rapido

Doc 2: el nino juega con el perro en el parque

Doc 3: el nino nada rapidamente

Palabra	Documento
el	1
perro	1
corre	1
rapido	1
el	2
niño	2
juega	2
con	2
el	2
perro	2
en	2
el	2
parque	2
el	3
nino	3
nada	3
rapidamente	3

Índice invertido

Paso 2: ordenamos la lista alfabéticamente

Palabra	Documento
con	2
corre	1
el	1
el	2
el	2
el	2
el	3
en	2
juega	2
nada	3
nino	2
nino	3
parque	2
perro	1
perro	2
rapidamente	3
rapido	1

Índice invertido

Paso 3: creamos el índice invertido

Palabra	Documento
con	2
corre	1
el	1
el	2
el	2
el	2
el	3
en	2
juega	2
nada	3
nino	2
nino	3
parque	2
perro	1
perro	2
rapidamente	3
rapido	1

Palabra	Documento
con	{2}
corre	{1}
el	{1,2,3}
en	{2}
juega	{2}
nada	{3}
nino	{2,3}
parque	{2}
perro	{1,2}
rapidamente	{3}
rapido	{1}

Mejoras: eliminación de stopwords, uso de lematización y stemming.

Índice invertido

Ahora que hemos calculado el índice invertido, vamos a buscar los términos:

Palabra	Documento
con	{2}
corre	{1}
el	{1,2,3}
en	{2}
juega	{2}
nada	{3}
nino	{2,3}
parque	{2}
perro	{1,2}
rapidamente	{3}
rapido	{1}

Buscamos los términos: “el”, “perro”, “nino”

{1,2,3} ∩ {1,2} ∩ {2,3} = {2}

El resultado indica que el documento 2 cumple con la consulta

Índice invertido

- Uno de los objetivos de implementar un índice invertido es **optimizar la velocidad de una consulta**
- A partir de un índice invertido es posible reconstruir el documento original
- En nuestro ejemplo, encontramos que las palabras 'el','perro','niño' se encuentran en el documento 2.
- **Ejemplos de uso:** página web, secuenciación del ADN (3,200 millones de pares de base)

Índice invertido

Frecuentemente el índice invertido es usado para la creación de **bolsas de palabras:** se coloca un 1 si la palabra aparece en el documento y 0 si no aparece.

Palabra	Documento
con	{2}
corre	{1}
el	{1,2,3}
en	{2}
juega	{2}
nada	{3}
nino	{2,3}
parque	{2}
perro	{1,2}
rapidamente	{3}
rapido	{1}

*Eliminamos stopwords e hicimos lematización y stemming

Bolsa de palabras

Palabra	Doc 1	Doc 2	Doc 3
corr	1	0	0
jug	0	1	0
nad	0	0	1
nin	0	1	1
parque	0	1	0
perr	1	1	0
rapid	1	0	0

Índice invertido

La columnas indican los términos que están en cada documento

Las filas representan los términos y los documentos dónde se encuentran

Palabra	Doc 1	Doc 2	Doc 3
corr	1	0	0
jug	0	1	0
nad	0	0	1
nin	0	1	1
parque	0	1	0
perr	1	1	0
rapid	1	0	0

La bolsa de palabras nos sirve para representar documentos

Ejercicio: índice invertido

- Doc 1: vendo camionetas y coches
- Doc 2: coches usados
- Doc 3: camionetas y motocicletas de ocasión
- Doc 4: coches de segunda mano
- Doc 5: coches y camionetas de ocasión
- Doc 6: permuto coche por camioneta

Buscar los términos: 'coche', 'camioneta'

TF-IDF

- Para la recuperación de información, uno de los pasos iniciales es comprobar la presencia o ausencia de la palabras que forman la cadena de consulta.

TF-IDF

- Para hacer la recuperación de la información, sería necesario 'mirar' cada documento y ver si las palabras que forman la consulta aparecen dentro del documento.
- ¿podría servir la frecuencia de aparición de las palabras en cada documento?
- ¿es posible estimar el valor de cada término dentro de un documento para mejorar el resultado de la consulta?

TF-IDF

- La ponderación de los términos tienen como finalidad conocer la importancia de los términos (palabras) para representar un documento y permitir su posterior recuperación.
- Para lograr esto es necesario calcular la capacidad de los términos para representar el contenido de de un documento en una colección de documentos.
- Objetivo: identificar cuáles documentos son relevantes o no a partir del cálculo de la ponderación de los términos.

TF-IDF

El algoritmo de TF-IDF (Term frequency - Inverse Document Frecuency) calcula el peso de un término en un documento en una colección de documentos

- A través de esta medición se expresa qué tan relevante es una palabra para un documento en una colección (factor de ponderación)

TF-IDF

¿Cómo se calcula: Term frequency?

- Term frecuency: es la suma de todas las ocurrencias ó el número de veces que aparece un término en un documento.
- También se le conoce como frecuencia relativa, porque nos referimos a un documento específico y no a toda la colección

TF-IDF

¿Cómo se calcula: Term frequency?

- Ejemplo: *Imaginemos un documento que contiene 1000 palabras, donde la palabra 'coche' aparece 50 veces*

$$TF = \frac{\text{Número de veces que aparece la palabra}}{\text{Número total de palabras en el documento}} \qquad TF = \frac{50}{1000} = 0.05$$

TF-IDF

¿Cómo se calcula: Term frequency - Inverse Document Frecuency?

- Inverse Document Frecuency: El factor IDF de un término es inversamente proporcional al número de documentos en los que aparece dicho término.
- Cuanto menor sea la cantidad de documentos, así como la frecuencia absoluta de aparición del término, mayor será su factor IDF

TF-IDF

¿Cómo se calcula: Inverse Document Frecuency?

- Ejemplo: *Imaginemos que tenemos 10 millones de documentos y la palabra coche aparece en 1000 de éstos documentos*

$$IDF = \log \left(\frac{\text{Número total de documentos}}{\text{Número total de términos en el documento}} \right) \qquad TF = \log \left(\frac{10000000}{1000} \right) = 4$$

TF-IDF

¿Cómo se calcula: Term frequency - Inverse Document Frecuency?

$$TF = \frac{\text{Número de veces que aparece la palabra}}{\text{Número total de palabras en el documento}} \qquad TF = \frac{50}{1000} = 0.05$$

$$IDF = \log\left(\frac{\text{Número total de documentos}}{\text{Número total de términos en el documento}}\right) \qquad TF = \log\left(\frac{10000000}{1000}\right) = 4$$

$$TF - IDF = TF * IDF \qquad TF - IDF = (0.05) * (4) = 0.2$$

TF-IDF determina la relevancia de un documento en comparación con los demás documentos que contienen ese mismo término

TF-IDF

¿Cómo se calcula: Term frequency - Inverse Document Frecuency?

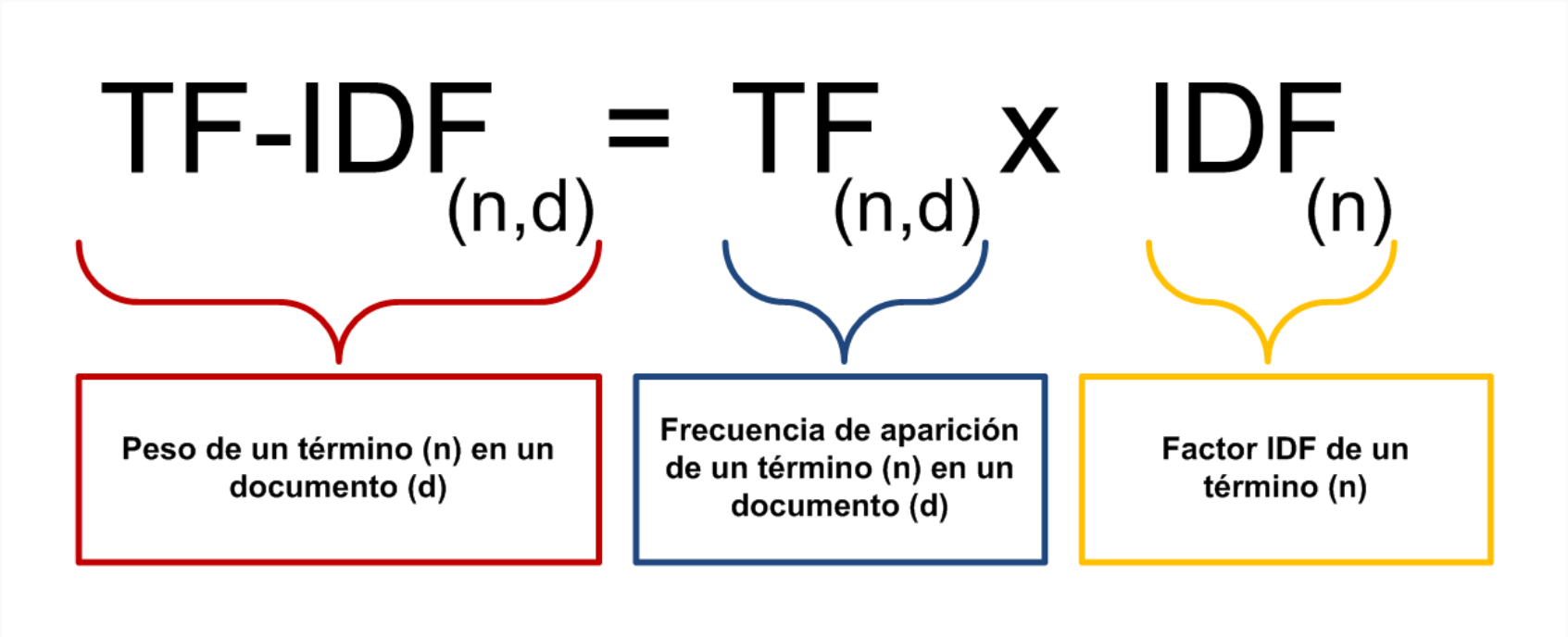


Imagen obtenida de Blázquez Ochando

Referencias

http://ccdoc-tecnicasrecuperacioninformacion.blogspot.com/2012/11/frecuencias-y-pesos-de-los-terminos-de.html