# Similitud entre documents (Part II)

Guillem Vidal

6 de desembre de 2024

## 1 Introducció

M'ha semblat adient modificar el `MapReduce` per tal d'optimitzar més les operacions, i fer-lo més genèric de manera que pogués acceptar més diversitat de tipus.

La funció que genera tot el funcionament amb actors es diu `groupMapReduce`, i té la següent signatura:

```
def groupMapReduce[K, A, B, C](
        input: Iterable[A],
        compute: A => Iterable[B],
        key: B => K,
        mapper: B => C,
        reducer: (C, C) => C,
        nmappers: Int = 16,
        nreducers: Int = 16
): Map[K, C]
```

El paràmetre `compute` és redundant: permet que durant el mapeig es generi una sèrie de valors a agrupar i reduir, que no sigui una relació 1 a 1 amb l'input, sinó 1 a N (opcionalment).

Això és útil quan hem de calcular les aparicions de les paraules en els documents, per exemple, ja que podem passar com a input els documents i generar totes les combinacions a dins del `groupMapReduce` aprofitant les seves aptituds de paral·lelisme.

## 2 Demostracions

### 2.1 La funció `timeMeasurement`

L'he definida de la següent manera:

```
def timeMeasurement[A](
        compute: Unit => A,
        name: String = "Function"
```

```
): A = {

        val beg = System.currentTimeMillis()

        val ret = compute()

        val end = System.currentTimeMillis()

        val elapsed = (end - beg).toFloat / 1000;
        println(s"$name took $elapsed seconds!")

        return ret
}
```

No he volgut utilitzar la sintaxi especial per l'avaluació *lazy* de les expressions, perquè m'agrada la claredat que el que s'està enviant és una expressió, no el valor d'aquesta.

## 2.2 Nombre promig de referències

```
112 references on average.
```

## 2.3 *Query* de recomanació mitjançant PR

Si busquem per 'Guerra', aquests són els primers resultats:

```
Natal (Rio Grande do Norte),
Heinkel He 280,
Fokker F.VII,
Raymond Collishaw,
Yokosuka MXY-7,
Diàspora basca,
Països àrabs,
Edat Contemporània als Països Catalans,
Història de Bielorússia,
...
```

## 2.4 Pàgines que s'assemblen sense referenciar-se

Donada una *query* de 'Guerra' i agafant els 100 primers elements retornats, obtinc els següents resultats:

```
(Economia de la Unió Soviètica, Èxode rural),
(Història d'Estònia, Èxode rural),
(Persona desplaçada, Èxode rural),
(Feixisme italià, Èxode rural),
(Història d'Amèrica, Èxode rural),
```

(Llista de diàspores, Èxode rural),
...