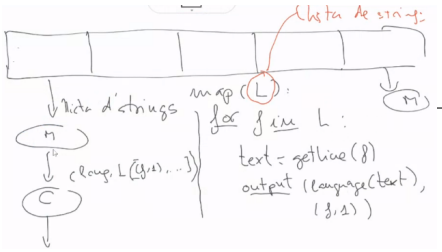
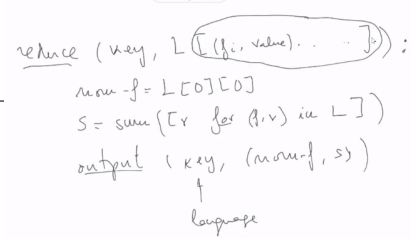
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Politècnica Superior d'Enginyeria
de Vilanova i la Geltrú

RECUPERACIÓ DE LA INFORMACIÓ



EXERCICIS DEL TEMA 6: MapReduce



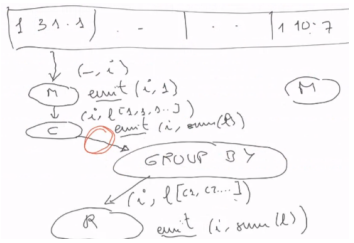
Exercici 1

Ens donen un nombre gran, N, de fitxers escrits en llengües diferents. Volem triar un subconjunt S dels fitxers que cobreixin totes les llengües i determinar la freqüència de cada llengua. Això és, per a cada llengua present entre els fitxers hi ha d'haver exactament un fitxer a S en aquesta llengua, junt amb el nombre de fitxers escrits en aquesta mateixa llengua. Expliqueu com resoldre eficientment aquesta tasca en el model de programació MapReduce.

Per ser precisos, suposem que a cada instància de map li arriba una llista de strings, on cada string és el nom d'un fitxer que ha de processar. Volem retornar, per cada llengua present al corpus, una tupla com ara `(''English'',(''The Road.txt'',2489))`, que indica que hi ha 2489 fitxers en anglès i que `''The Road.txt''` n'és un. Suposeu que hi ha una funció `language(string t)` que retorna el nom de la llengua en la què és escrit el text `t`, com ara `''English''`.

Exercici 2

Expliqueu com resoldre les tasques següents usant el model de programació MapReduce. Penseu que potser alguna tasca necessita més d'una fase de map+reduce. Suposeu que S és un *gran* multiset d'enters compresos en algun *llarg* rang $1 \dots N$.



1. Donat S, calculeu el seu histograma: per cada $i \in 1 \dots N$, doneu el nombre d'ocurrències de i a S.
2. Donat S i una funció `bool f(int n)`, digueu quants elements de S fan `f` certa.
3. Donat S i un nombre k, partiu $1 \dots N$ en k intervals de la mateixa mida i digueu quants elements de S pertanyen a cada interval.

Exercici 3

1. Suposeu que tenim dos vectors molt llargs $x, y \in \mathbb{R}^n$, i que només caben $k \ll n$ components en una màquina. Doneu una solució usant el model MapReduce per calcular en producte escalar de x i y.
2. Donada una matriu A i un vector x, expliqueu com calcular el producte Ax en el model MapReduce. Suposeu que la matriu A sencera, no pot ser tractada per una sola màquina.

Exercici 4

We say that two consecutive words in a document form a 2-phrase. A document with L words can thus contain at most L-1 different 2-phrases. Write programs in mapreduce (= map, reduce, and perhaps combine and partition functions) to do the following:

1. Given a set of documents and a parameter k, return the set of 2-phrases that appear at least k times in the set.

2. Given a set of documents and a word w, returns a list of all the pairs (d, v) for each document d and for each v such that (w, v) is a 2-phrase in document d.

Exercici 5

Give a solution in the mapreduce model for the following problem: We have a (very big) file formed by lines. Each line contains information on a blog post. More precisely, a line contains one or more names; the first name is the author of the post, and the rest (if any) are the names of the people that commented on the post. Note that a person may comment several times on the same post, and the author him/herself may comment too. For example, in the three lines

```
Jane Joe Zuzana Rosamaria Rosamaria Peter Jane Rosamaria
Jane Xabier Peter Xabier Martha Xabier Martha
Jane
```

the first one represents a post by Jane, which was commented by herself, Joe, Zuzana, Rosamaria (three times), and Peter. The second post is also authored by Jane, and was commented by Xabier (three times), Peter, and Martha (twice). The third post by Jane did not receive any comment at all.

We want to obtain, for each user X that has ever posted something, the name of the user Y that has commented on more *different* posts by X - that is, we want one pair (X,Y) for each such X. If several Y's are tied, return any of them. For example, if Jane only authored the three posts above, then we should return the pair (`Jane`,`Peter`), because Peter has commented on the largest number of posts by Jane.

Give pseudocode for map and reduce functions and, if appropriate, combine and partition functions. If it is not obvious, explain what you choose to be the input to a map instance. The efficiency of your solution will be valued. Note that a solution with a single mapreduce phase suffices, but better give one that uses several phases than nothing!