```
Indicacions sobre les solucions d'alguns problemes del segon parcial i el final
Gener 2015

Exercici 1 del parcial, 4 del final

Amb una sola fase de mapreduce, suposant que es dona una línia
a cada instància de mapreduce)

map(string linia)
   l = l.split(" ")
   autor = l[0];
   l = l[1..final];
   eliminar repetits de l;
   per cada x de l, output (autor,x)

combiner(autor,l)
  eliminar repetits de l;
  output (autor,l);

reduce(autor,l)
  // L és una llista de llistes
  eliminar repetits de l;
  output (autor,longitud(l));


Eliminar repetits al map i al combiner
és opcional, però molt convenient per eficiència.
Notem que un combiner rep elements de molts maps,
i per tant és *més* eficient fer-ho a tots dos
llocs que només al combiner.

Una alternativa, probablement més eficient, és
no emetre una tupla per cada comentari, sinó una
per cada post:

map(string linia)
   l = l.split(" ")
   autor = l[0];
   l = l[1..final];
   eliminar repetits de l;
   output (autor,l);

combiner(autor,L)
  // L és una llista de llistes
  fusionar totes les llistes de L
  i eliminar repetits; sigui l el resultat
  output (autor,l);

reduce(autor,L)
  // L és una llista de llistes
  fusionar totes les llistes de L
  i eliminar repetits; sigui l el resultat;
  output (autor,l);

Exercici 4 del parcial

Manera 1: executar els dos algorismes separadament,
i combinar els ratings (p.ex. fent la mitjana, triant el màxim)
en un nou ranking, o bé triant els top k/2 de cadascun
```

```
i per donar k recomanacions.

Manera 2: executar primer CB per seleccionar un conjunt
d'items prometedors, i llavors fer CF només sobre aquest subset.
O en l'ordre invers (es consideren el mateix mètode).

En aquest problema hi ha hagut la tendència de donar solucions
força vagues.

Exercici 5 del parcial, 6 del final

No s'hi valien declaracions generals que no eren
escenaris sinó problemes, en especial si només
es repetia el que diuen les descripcions de spark
en les primeres línies
- "per a problemes de machine learning": hadoop va molt
bé per a molts problemes de machine learning. Problemes
de machine learning en temps real és una altra cosa.
- "per a problemes molt grans": idem, hadoop va bé
per a problemes molt grans també.
- "per a problemes de molta velocitat, spark diu que
pot ser fins a 100 vegades més ràpid que hadoop"...
en alguns problemes, es tractava de saber quins i dir-ho
(o si no se sap, buscar altres raons).

Escenaris concrets podria ser:

- anàlisi de seguretat en temps real
- anàlisi de sèries financeres en temps real
- recomanacions on productes i preferències varien en el temps
- servei transaccional alhora que anàlisi de tendències
  en temps real d'un lloc com Twitter
- ...

Exercici 1 del final:
1. fals (els tf no cal recalcular-los)
2. cert
3. cert
4. fals
5. (no comptava, però fals)
6. cert
7. fals (de fet, el que es diu no té sentit)
8. fals (no té solució)
9. fals: afegir damping factor NO arregla gens aquest problema.
   se seguirà perdent pagerank i només la solució amb
   tots els pageranks 0 és factible
10. cert
```