CAIM, segon parcial

8 de gener de 2015. Temps: 1 hora 50 minuts

Exercici 1 (2,5 punts) Dona una solució en el model mapreduce per al següent problema. Tenim un fitxer (molt gran) on a cada línia hi ha un o més noms, separats per exemple per blancs. Cada línia representa un post en una xarxa social. El primer nom de la línia és el de l'autor del post, i cada nom de després el nom d'algú que ha fet un comentari al post. Aquests noms poden estar repetits perquè una persona pot fer diversos comentaris al mateix post, i també l'autor pot comentar el seu post. Per exemple, una línia podria ser

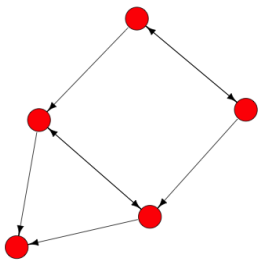
```
Joan Maria Oriol Maria Joan Pere Tomeu Rosa Maria
```

que vol dir que un post determinat d'en Joan l'han comentat el mateix Joan, la Maria, l'Oriol, en Pere, en Tomeu i la Rosa.

Cal obtenir, per a cada persona que ha fet algun post, el nombre de persones *diferents* que han comentat posts seus.

Recorda que cal donar pseudocodi per a les funcions map i reduce (i, si convé, per a funcions combiner i partition), i que en alguns problemes és necessària més d'una fase de mapreduce (i en altres no). Es valorarà l'eficiència.

Exercici 2 (2,25 punts) Determina quins nodes són els més i menys centrals segons les nocions de (a) centralitat de grau, (b) centralitat "closeness", (c) centralitat "betweenness" i (d) centralitat "Pagerank" amb damping factor $\lambda = 0.9$. Per a les tres primeres centralitats, considera la versió no dirigida del graf donat.



Exercici 3 (2,25 punts) Tenim una xarxa amb molts milions de nodes i per implementar l'algorisme jeràrquic de detecció de comunitats necessitem detectar quins parells de nodes són més similars segons la mesura de similaritat "Jaccard index". Explica com podries trobar els parells mes similars amb un cost aproximadament lineal sobre el nombre de nodes de la xarxa. Pots suposar que accedir als veïns d'un node qualsevol té cost constant.

Exercici 4 (2 punts) Pensa i explica dues maneres de combinar els enfocs Collaborative Filtering i Content-Based a la recomanació (5 línies cada explicació, orientativament).

Exercici 5 (2 punts) (Aquesta pregunta serveix també per a avaluar la competència transversal "aprenentatge autònom"). Descriu dues situacions reals on Spark seria avantatjós respecte Hadoop ("reals" vol dir que parlin d'un domini concret, no només en termes de la necessitat tècnica com ara "ser online"). Explica breument què mantindries a memòria i què en disc si usessis Spark.