CAIM, segon parcial

18 de gener de 2016. Temps: 1 hora 50 minuts

Exercici 1 (2.5 punts) We are a company. We keep records of the money that we exchange with customers, either in payments or in reimbursements. At any given moment, a customer may owe us some money, or we may owe him/her some money. The records are large set of files. Each line of the file may look like the following two:

```
ricard tarragona +100 -50 -200 +30 +10
marta barcelona +100 +20 -50 -50
```

These lines imply that ricard, who lives in tarragona, currently owes us 250 - 140 = 110 euros, but we owe marta, who lives in tarragona, 120-100 = 20 euros. We say that ricard is "in the red" but marta is not. We are promised that for each customer there is exactly one line among all the files, i.e., there are no two lines anywhere with the same customer name.

Explain how to solve in the mapreduce paradigm the following two problems. Tell if you solve them in one mapreduce phase or more than one, and give map and reduce functions (and, if applicable, combine and partition functions) for each phase.

- Give the total amount of money that we have received in payments and the total amount of money that we have paid to customers (we want two separate numbers, not just their difference! The output for the input consisting only of the two lines above should be the two numbers 260 and 350).
- Give a list of the cities where over 10% of our customers in the city are the red.

Exercici 2 (2.5 punts) Compute diameter, local clustering coefficient and degree distributions of the following networks. Compute centralities (degree, closeness and betweenness) for all of their nodes as well.

- Watts-Strogatz network with $p = 0$ and 6 nodes.
- 3×2 grid.

Exercici 3 (2.5 punts) Say whether each of the following statements is true or false, and justify in 1-2 sentences your answer.

1. HyperLogLog and SpaceSaving sketches can be used to find the highly frequent items in a stream.
2. A requirement for streaming algorithms is that after reading a stream of t items they use memory less than ct for every constant c.

3. In locality sensitive hashing, the technique of *stacking* k hash functions is used to avoid collisions between dissimilar items

4. In locality sensitive hashing, the technique of *repeating* is to improve efficiency and performance.

5. Content-based recommender systems require longer customization time than Collaborative Filtering approaches in new application scenarios.

6. Matrix factorization methods based on SVD can be used to solve the coldstart problem in Collaborative Filtering recommendation systems.

Exercici 4 (2.5 punts) Explain two scenarios where Spark would be better than Hadoop for processing your data.

Note: "Machine learning" or "real-time processing" are *not* scenarios. A scenario is something like "we manufacture umbrellas and we need to suddenly increase our production and send umbrellas to the shops quickly when the weather report announces rain". Explain the two scenarios and then justify them briefly. Be specific. You definitely do not need to explain what Spark or why it is good in abstract.

Note 2: This exercise will be used to compute the grade of the Competència Transversal Aprenentatge Autònom.