Lab 4

In this lab, we keep on our thoughts on the Amazon food dataset we have been using so far, and we start using the rating the users give to the products.

A rating is a number of stars between 1 and 5, where 5 is “I love it” and 1 is “I hate it”. There are a number of philosophical investigations and research papers on what the other rating really means. For some really critical users, 3 stars could be really a good rating, maybe the maximum they would ever give to a product; others instead, have never gone below 4, and only in exceptional cases, when the product bought was really unsatisfactory.

What we will try in this laboratory is to normalize the ratings according to the user’s proclivity to give a 5 star.

Let’s see an example:

	A1	A2	A3	A4	A5
B1		5		4	1
B2	3				
B3		5	4	5	4
B4				5	
B5		5		2	3

The column of this matrix are the users, while the rows are the products. User A2 has given 3 reviews: to product B1, to B3 and to B5, all of which are 5 stars. A5 instead has given 1 star to B1, 4 stars to B3 and 3 stars to B5, etcetera.

We can normalize this matrix subtracting, from each column, its mean value, obtaining thus:

	A1	A2	A3	A4	A5
B1		0		0	-1.67
B2	0				
B3		0	0	1	1.33
B4				1	
B5		0		-2	0.33

Now we see that, for example, A4 has given 1 more than its personal average to product B3, that is to say he likes it, while she has given 2 less than the average to B5, so she definitely did not like it. A5 instead was not so unsatisfied by B5, since she gave 0.33 more than its average rating to this product.

Now for each of the products we can compute the average of such normalized ratings, obtaining a “normalized average rating” for each product :

B1	-0.56
B2	0
B3	0.58
B4	1
B5	-0.56

Notice how the ranking of B5 was affected by this transformation.

Your task for this lab is to write an Hadoop Application to compute the normalized ratings of the products of the Amazon food dataset by considering the (sparse) products/users matrix based on the ratings available in the file /data/students/bigdata-01QYD/Lab3/Reviews.csv that you used already in the previous lab.

For the initial test of your application you can use the small sample dataset ReviewsSample.csv, which contains a set of reviewes related to the same users, products, and ratings of the small example products/users matrix reported in this document.
ReviewSample.csv is available at the following address:
http://dbdmg.polito.it/wordpress/wp-content/uploads/2016/04/ReviewsSample.csv

Hints

Beware that the products/users matrix is sparse, i.e. many of its values are null/unknown. In your application try to exploit this fact, avoiding shuffling unuseful key/value pairs to the network.