Carmen Díez
IDS2020

Exercise 3 - Spark

You need to write a program that uses Spark to provide an answer to the following questions.
***(1)** Calculate the following values from the dataset: **a.** Minimum **b.** Maximum **c.** Average.*

Using **min()**, **max()** and **mean()** (or dividing **sum/count**) we get these values. It is on the script, using spark as in the sample script given.

The result for the big dataset is the following:

20/10/12 00:15:48 INFO DAGScheduler: Job 3 finished: min at /wrk/users/carmendi/carmendi.py:22, took 196.055552 s
Count = 100000000.00000000
Sum = 5000495343.66150284
Mean = 50.00495344
Min = 0.00000348
Max = 99.99999933

***(2)** Explain how you would compute the mode for the data set. Additionally, explain why this dataset is not very well suited for computing the mode. What changes in the dataset would you propose to make it better?*

I would use `sorted(data.countByValue().items(), reverse=True)[0][0]` that would return a dictionary and then take the value that appears the most. This dataset is not well suited for the mode as we're using numbers with many decimal units so it's complicated that two numbers are exactly the same. It would make more sense using intervals to compute something similar to a mode.