**BIG DATA COMPUTING 2019-20 – HOMEWORK 3 – GROUP 49**

**(USE ONLY 1 PAGE)**

1. **Required tests.** Do the tests with the parameters indicated in the first 4 columns of the following table. Fill the table reporting, for each test, the following values: **Init** = *time to read the input and create the RDD*;**Ti** = *time of Round i of runMapReduce* (i=1,2), **AvgDist** = *average distance among the solution points*. **Times should be in ms.** If you notice anomalies in the values of a test try to repeat the test 3 times and take the median values*.*

*Some anomalies are to be expected!*

**Table to be filled by Java users**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **K** | **L** | **Num Exec** | **Init** | **T1** | **T2** | **AvgDist** |
| **Glove2M300d.txt** | **100** | **16** | **16** | 22441 | 10455 | 19699 | 29.040733847821866 |
| **Glove2M300d.txt** | **100** | **16** | **8** | 36149 | 19947 | 19863 | 29.040733847821866 |
| **Glove2M300d.txt** | **100** | **16** | **4** | 63788 | 39578 | 19884 | 29.040733847821866 |
| **Glove2M300d.txt** | **100** | **4** | **4** | 64254 | 48036 | 1107 | 29.066280389112293 |

1. **Observations:** *write here a brief comparison among the results of the experiments reported in the table (and perhaps of other experiments that you did on your own), and try to justify the differences that you may have observed.*

For brevity we will call the number of executors NE, while the initialization time TINIT.

As we can see from the table, obviously NE impacts the overall running time. In particular, for round1 we can see that NE is inversely proportional to T1, indeed if NE doubles, the runtime halves and vice-versa: here we have an almost linear relation. For T2 instead, the computational time is almost the same for every NE. This happens because round2 can be executed with only one reducer, so having more than one executor is useless. Here we do not have a relation between NE and T2.

Instead, for TINIT we can notice that NE impacts the running time. In fact, as NE increases, TINIT decreases. As one would expect, reading the dataset with bigger NE would take shorter compared to small NE.

The chart below is a summary of what was said above and it can be used as an alternative way of reading the table.

Another remark on some tests that we have done is regarding the cache. In particular, we tried to use and do not use cache in round 1 and results are interesting. With the cache we get round 1 with greater running time wrt round 2 (as you can see from table), while without cache we get running times more equally distributed but the round 2 takes more time that the one reported in the table. However, in general, results are better with cache.

