HOMEWORK 1

GROUP 36

**Analysis of MRPrintStatistics:**

We start by describing in the standard way the MRPrintStatistics function, which is a 3 round function.

**Input:**

* N pairs (point, group) where each point is represented as a Vector of coordinates and each group is a Boolean (True = A, False = B);
* K points represented as Vector that are the centroids.

**Output:**

* K pairs where is the index of the centroid of the input and are the number of points of A and B, respectively, in the cluster centered in . To output them in the convenient way, these should be sorted by index.

1. **ROUND 1:**
   1. **Map phase:** for each (point, group) pair separately compute the index of the closest center to it and put a 1 on the group of the point and a 0 in the other one, achieving:  
      (point, A) -> or (point, B) ->
   2. **Reduce phase:** for each partition separately compute   
       where and are respectively the number of points of the partition in the cluster centered in the centroid that are in the group A and B.
2. **ROUND 2:**
   1. **Map phase:** empty.
   2. **Reduce phase:** for each index separately compute by summing the and of each partition .
3. **ROUND 3 (only for sorting):**
   1. **Map phase:** map every index to the same key.
   2. **Reduce phase:** sort the triplets by key.

Note that the reduce phase of round 1 and the reduce phase of round 2 are handled implicitly by Spark, as we only call the function *reduceByKey* which exploits the partitions as indicated in the official guide [1]. Note also that Round 3 is entirely handled by *sortByKey* function.

Let’s compute a bound to the amount of local space required by each phase:

1. **ROUND 1:**
   1. **Map phase:** we simply transform each point separately, so local memory is
   2. **Reduce phase:** we group together all the points of the same partition, so supposing the partitions are well defined (and they are since they are handled by Spark itself) the amount of local memory is where is the total number of partitions.
2. **ROUND 2:**
   1. **Map phase:** empty.
   2. **Reduce phase:** we have K triplets for each partition, so a total of .
3. **ROUND 3:**
   1. **Map phase:** is .
   2. **Reduce phase:** is .

In total, the required local space is , which depends on the total number of partitions, that is a parameter set externally by the user. If then the local space requirements become .