Sapienza - Università di Roma
Data Mining - Fall 2016
Apache Storm Tutorial - Homework

Sara Veterini 1536622
Davide Mazza 1520961
Lorenzo Rutigliano 1449848
Federico Croce 1546488
Riccardo Di Stefano 1528140
Roberto Gaudenzi 1527361

December 19, 2016

Instructions – Read carefully!
You must hand in the homeworks electronically and before the due date and time (December 27, 23:59).
Handing in: You must hand in the homeworks by the due date and time by an email to apachestormtutorial@gmail.com that will contain as attachment (not links!) a .zip or .tar.gz file with the Eclipse project that implements your solution.

1 Assignment

We mentioned in class the algorithm for sampling uniformly on a stream of data. Let assume we've read a set of values $x_1, x_2, \ldots, x_{i-1}$ so far, then the next value x_i will become the current sample with probability $1/i$, otherwise the sample will remain the previous one.

You are asked to implement an Apache Storm Trident topology that performs sampling uniformly on a stream of integers. You'll have a Spout that generates random numbers, and a BaseFilter that decides whether to keep the new number as a sample or not.

Within the virtual machine provided during the tutorial, you can find an archive containing an Eclipse project to be completed. To import it into Eclipse:

1. Extract the project from the archive and move it within the "workspace" folder.

2. Open Eclipse, and import the project by clicking on "File/Import/Existing Project into workspace" and selecting the folder cited above.

Hint: to generate a random number, you can use the $nextInt()$ method from java.util.Random class. You can use this function to decide when to update the sample too.