University of Ioannina,
Department of Computer Science & Engineering

Data Mining – CS 059
Fall 2013

Assignment 1

This is the first part of Assignment 1. The deadline for this part of the assignment is November 8, 11:00 am, at the beginning of the class. You should turn in the code for question 1, and submit the remaining questions either electronically, or on paper. For late submissions the late policy on the page of the course will be applied. Details for the turn-in, and how to write reports are on the Assignments web page of the course.

Question 1 (Weighted Reservoir Sampling)

In class we described the Reservoir Sampling algorithm for sampling a single item from a stream of N items. In this question you are required to modify the algorithm to do **weighted sampling**. We assume that each item i has an integer weight w_i. You will modify the sampling algorithm so that given a stream of N weighted items we sample one item with probability proportional to its weight, that is, the probability of sampling item i is proportional to w_i. Similar to before, the algorithm should work with constant amount of memory, independent of N. Prove the correctness of your algorithm

Question 1 (Reservoir Sampling)

In this question you are required to modify the simple (no weights) Reservoir Sampling algorithm to sample $\boldsymbol{k}$ items from a stream of N items.

1. Describe the algorithm for sampling k items uniformly at random from a stream of N items. The algorithm should work in a single pass over the data, reading the items one by one, without prior knowledge of the size of the stream N, and using $O(k)$ of memory (assume the size of an item is fixed).
   (**Hint**: In a random sample each element should have probability k/N to appear in the sample).
2. Prove that your algorithm produces a uniform sample, that is, prove that for every $i, 1 \leq i \leq N$, the i-th element has probability k/N to appear in the sample.
   (**Hint**: What is the probability that after the i-th item is selected, it is later replaced in the j-th item, for $j > i$?)
3. Write a program that implements the sampling algorithm (in any language you want). Your program should sample k lines from a file. It should be possible to use the program from command line, and it should take command line parameters the value of k, the input file name, and the output file name. Therefore, your program should work as follows:
   "`sample <k> <inputfile> <outputfile>`".

Question 3

On the Assignments page of the course there is a file “stocks.txt” that contains tab-separated columns of values. Each line corresponds to the values of two stocks on a given day, and the file contains the values for 100 consecutive days. Imagine that you are a data analyst in a bank and you want to find out if there is some relationship between the two stocks. Analyze the data and write a short report about the kind of analysis that you did and your findings. The report should have convincing evidence (with numbers and/or plots) about the relationship between the two stocks.