Fall 2016

Big Data Management

Learning objectives of Lecture 4 (21.11.2016)

Department of Computer Science
University of Helsinki

General Instruction: After the fourth lecture of big data course, you should be able to answer the following questions.

1. What is data stream? What are typical applications of data stream?

2. Consider a small set {1,3,4}. Assume that the size of a bit array is 3. Give two hash functions H1 and H2 such that if we only use H1, then we will return a false positive for 6. But if we use both H1 and H2 in a bloom filter, then we can avoid this false positive for 6.

3. How can we extend Bloom filter to support the deletion of an element in a set?

4. Try to prove the following theorem for the space complexity of Count-min Sketch by yourself.

   Theorem: Give an $\varepsilon||x||_1$ error with probability $1-\delta$, the count-min sketch needs to have size $\frac{e}{\varepsilon} \times \ln\frac{1}{\delta}$

5. Will the Count-min make under-estimation? Why?

6. Count-min is a biased estimation or an unbiased estimation?