Big Data Analytics

ESSEC

Home work 4: Mining Data Streams

1. (**Exercise 4.2.1 MMDS book)** Suppose we have a stream of tuples with the schema (university, courseID, studentID, grade). Assume universities are unique, but a courseID is unique only within a university (i.e., different universities may have different courses with the same ID, e.g., "CS101") and likewise, studentID's are unique only within a university (different universities may assign the same ID to different students). Suppose we want to answer certain queries approximately from a 1/20th sample of the data. For each of the queries below, indicate how you would construct the sample. That is, tell what the key attributes should be.

   - For each university, estimate the average number of students in a course.
   - Estimate the fraction of students who have a GPA of 3.5 or more.
   - Estimate the fraction of courses where at least half the students got "A."

2. (**Exercise 4.3.1 : MMDS book)** For the situation of our running example (8 billion bits, 1 billion members of the set S), calculate the false-positive rate if we use three hash functions? What if we use four hash functions?

3. (**Exercise 4.4.1 MMDS book)** Suppose our stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Our hash functions will all be of the form $h(x) = ax + b \text{mod} 32$ for some a and b. You should treat the result as a 5-bit binary integer. Determine the tail length for each stream element and the resulting estimate of the number of distinct elements if the hash function is:

   - $h(x) = 2x + 1 \bmod 32$;
   - $h(x) = 3x + 7 \bmod 32$;
   - $h(x) = 4x \bmod 32$;