Autumn 2017

Big Data Management

Exercise 1

Department of Computer Science
University of Helsinki

General Instruction: You need to work independently.

1. What are the five main properties of a sketching algorithm described in the lecture? Please discuss each property with details for Bloom filters, Count-Min Sketch, Count Sketch and FM Sketch. (10 points)

2. Let S1 and S2 be two sets where the elements come from the same universe U. Let F(S1) and F(S2) be the bloom filters on S1 and S2 respectively. Recall that a bloom filter is a bit array of length x constructed using a set of hash functions from U to [x], where [x] denotes the set of integers {0, ..., x − 1}. Assume that F(S1) and F(S2) have the same length x, and are constructed with the same set of hash functions. Now, consider F = F(S1) AND F(S2), where the AND operator produces a bit array by taking the conjunction of each pair of corresponding bits. Prove that F is exactly the bloom filter on S1 ∪ S2. (10 points)

3. Given two vectors X =<x_1, x_2,… x_n> and Y=< y_1, y_2,… y_n >, the dot product of X and Y is:
$$X \cdot Y = \sum_{i=1}^{n} x_i\, y_i$$
Design an algorithm to use Count-Min sketch to estimate the dot product of V · V, where V is a vector. Analyze the probabilistic error and the space cost of your algorithm. (10 points)

4. Let S1 and S2 be two bags where the elements come from the same universe U. Let FM(S1) and FM(S2) be the FM-sketches on S1 and S2 respectively. Recall that each FM-sketch is constructed using a hash function from U to [2w], where w is set to $\log_2$ U in our context. Suppose that FM(S1) and FM(S2) are built using the same hash function. Describe an algorithm to obtain an FM-sketch on S1 ∪ S2 from FM(S1) and FM(S2). (10 points)

5. Consider two data sets F and G given as pairs (*key*, *frequency*): F{(1,2),(0,1),(4,1),(3,2)} and G{(2,1),(3,1),(0,2)}. Please estimate the size of join $|F \bowtie G|$ of two sets using Count-sketch with a 3×3 matrix. The hash function of keys and the ± 1 hashes can be found the following tables (10 points).
Hash functions of keys (j starts from 0) :
(1) h1(j) =$j \; mod$ 3
(2) h2(j) = (j mod 4) mod 3
(3) h3(j) = (2*j) mod 3

	key domain				
	0	1	2	3	4
1	+1	-1	-1	+1	+1
2	-1	+1	-1	+1	-1
3	-1	-1	+1	+1	+1