Bielefeld University
Faculty of Technology

Big Data Analytics

Summer 2022

Number 06, Submission Deadline: June 26. 2023, 11:59 PM

1. **Bloom filter I:**
   A bloom filter is a so called probabilistic data structure and mainly used to estimate if a given data point already occurred in a continuous stream of data. (6 P)

   (a) What is probabilistic about bloom filters?

   (b) What can you say about the properties of the bloom filter with respect to precision and recall?

2. **Bloom filter II:**

   (a) Construct the bit array of a 20-bit bloom filter for the following stream of elements $S_1 = \{10, 15, 3, 7, 2, 1, 12\}$ and the three hash functions h1, h2, and h3: (4 P)

   $h1(S) = (s + 1) mod 20$
   $h2(S) = (2s + 2) mod 20$
   $h3(S) = (3s + 3) mod 20$

   (b) Consider a bloom filter given by the following 20-bit filter array and the three hash functions from the previous exercise: (4 P)
   [10001101101010111001]
   Which of the following stream elements $s_i \in S_2$ have already been recorded according to the Bloom Filter: $S_2 = \{15, 1, 10, 7, 3, 12, 2\}$

3. **Flajolet–Martin algorithm**[1]: (4 P)
   Apply the Flajolet-Martin Algorithm to count the number of distinct elements in a stream. Suppose we have ten possible elements $1, 2, ..., 10$, that could appear in the stream, but only four of them actually appeared.
   To estimate how many different elements we have seen in the stream, we hash every element to a 4-bit binary number.
   As a hash function we use $h(x) = (3x+7)mod11$. For example element $x = 8$ hashes to $3*8+7 = 31mod11 = 9$. Thus, the 4-bit string (binary representation) for element 9_{10} is 1001_2.

   Consider the following sets of elements: $A = \{2, 3, 6, 9\}; B = \{1.3.9.10\}; C = \{1, 4, 7, 9\}; D = \{4, 6, 9, 10\}$.
   Which of these sets produces the correct estimate of 4 distinct elements [2]?

   Please answer by either going through the Flajolet-Martin algorithm manually or implement a solution in code.

Important:
Please submit your group solution via LernraumPlus. You are free to hand solutions in as PDFs or Jupyter Notebooks.

[1] Here your can find some additional explanations:
`https://arpitbhayani.me/blogs/flajolet-martin`

[2] Sometimes you may read about a factor $phi = 0.77351$, which is suggested for correcting the final approximation. You don't have to use this in this task.