BDA - Assignment 6

by Group 9 -Prince Prince, Bhautikkumar Ashokbhai Lukhi
Lyon Brown

23rd June 2023

1. Bloom filter I

A Bloom filter is a so-called probabilistic data structure mainly used to estimate if a given data point has already occurred in a continuous stream of data.

(a) **What is probabilistic about bloom filters?**

The Bloom filters are probabilistic in nature because of their ability to provide approximate answers with some level of uncertainty. Bloom filters can have false positives, meaning they might incorrectly indicate that an element is in the filter when it is not. However, they can definitively determine if an element is not present. The probability of false positives depends on factors such as the size of the filter, the number of hash functions used, and the number of elements inserted into the filter.

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

Bloom filters prioritize high recall, which means they have a low probability of false negatives. If a Bloom filter indicates that an element is not present, it is guaranteed to be accurate. However, they have lower precision due to the possibility of false positives. There is a chance that the filter may indicate an element is present when it is not.

2. Bloom filter II

(a) Construct the bit array of a 20-bit bloom filter for the stream of elements $S1 = \{10, 15, 3, 7, 2, 1, 12\}$ and the three hash functions:

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

Initialize the 20-bit array with all bits set to 0:

Bit Array: 00000000000000000000

For element 10:
$\text{h1}(10) = (10 + 1) \mod 20 = 11$
$h2(10) = (2 \cdot 10 + 2) \mod 20 = 2$
$h3(10) = (3 \cdot 10 + 3) \mod 20 = 13$
Set positions 11, 2, and 13 to 1 in the bit array:
Bit Array: 00000010100000000100

For element 15:
$\text{h1}(15) = (15 + 1) \mod 20 = 16$
$h2(15) = (2 \cdot 15 + 2) \mod 20 = 2$
$h3(15) = (3 \cdot 15 + 3) \mod 20 = 18$
Set positions 16, 2, and 18 to 1 in the bit array:
Bit Array: 01010010100000000100

For element 3:
$h1(3) = (3 + 1) \mod 20 = 4$
$h2(3) = (2 \cdot 3 + 2) \mod 20 = 8$
$h3(3) = (3 \cdot 3 + 3) \mod 20 = 12$
Set positions 4, 8, and 12 to 1 in the bit array:
Bit Array: 01010011100100010100

For element 7:
$h1(7) = (7 + 1) \mod 20 = 8$
$h2(7) = (2 \cdot 7 + 2) \mod 20 = 16$
$h3(7) = (3 \cdot 7 + 3) \mod 20 = 4$
Set positions 8, 16, and 4 to 1 in the bit array:
Bit Array: 01010011100100010100

For element 2:
$h1(2) = (2 + 1) \mod 20 = 3$
$h2(2) = (2 \cdot 2 + 2) \mod 20 = 6$
$h3(2) = (3 \cdot 2 + 3) \mod 20 = 9$
Set positions 3, 6, and 9 to 1 in the bit array:
Bit Array: 01010011101101011100

For element 1:
$h1(1) = (1 + 1) \mod 20 = 2$
$h2(1) = (2 \cdot 1 + 2) \mod 20 = 4$
$h3(1) = (3 \cdot 1 + 3) \mod 20 = 6$
Set positions 2, 4, and 6 to 1 in the bit array:
Bit Array: 01010011101101011100

For element 12:
$h1(12) = (12 + 1) \mod 20 = 13$
$h2(12) = (2 \cdot 12 + 2) \mod 20 = 6$
$h3(12) = (3 \cdot 12 + 3) \mod 20 = 9$

Set positions 13, 6, and 9 to 1 in the bit array:
Bit Array: 01010011101101011100

The resulting bit array represents the Bloom filter for the given stream of elements and hash functions.

(b) Consider a bloom filter given by the following 20-bit filter array and the three hash functions from the previous exercise Which of the following stream elements have already been recorded according to the Bloom Filter:

Filter Array: [10001101101010111001]

Stream of Elements: $S2 = \{15, 1, 10, 7, 3, 12, 2\}$

To determine which elements from $S2$ have already been recorded according to the Bloom Filter, we will apply the three hash functions:

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

For element 15:

$$h1(15) = (15 + 1) \mod 20 = 16$$
$$h2(15) = (2 \cdot 15 + 2) \mod 20 = 2$$
$$h3(15) = (3 \cdot 15 + 3) \mod 20 = 18$$

Since for all positions 16, 2, and 18 in the filter array are not set to 1, element 15 may have not been recorded.

For element 1:

$$h1(1) = (1 + 1) \mod 20 = 2$$
$$h2(1) = (2 \cdot 1 + 2) \mod 20 = 4$$
$$h3(1) = (3 \cdot 1 + 3) \mod 20 = 6$$

Since for all positions 2, 4, and 6 in the filter array are not set to 1, element 1 may have not been recorded.

For element 10:

$$h1(10) = (10 + 1) \mod 20 = 11$$
$$h2(10) = (2 \cdot 10 + 2) \mod 20 = 2$$
$$h3(10) = (3 \cdot 10 + 3) \mod 20 = 13$$

Since for all positions 11, 2, and 13 in the filter array are not set to 1, element 10 is not recorded.

For element 7:

$$h1(7) = (7 + 1) \mod 20 = 8$$
$$h2(7) = (2 \cdot 7 + 2) \mod 20 = 16$$
$$h3(7) = (3 \cdot 7 + 3) \mod 20 = 4$$

Since for all positions 8, 16, and 4 in the filter array are not set to 1, element 7 is not recorded.

For element 3:

$$h1(3) = (3 + 1) \mod 20 = 4$$
$$h2(3) = (2 \cdot 3 + 2) \mod 20 = 8$$
$$h3(3) = (3 \cdot 3 + 3) \mod 20 = 12$$

Since for all positions 4, 8, and 12 in the filter array are not set to 1, element 3 is not recorded.

For element 12:

$$h1(12) = (12 + 1) \mod 20 = 13$$
$$h2(12) = (2 \cdot 12 + 2) \mod 20 = 6$$
$$h3(12) = (3 \cdot 12 + 3) \mod 20 = 9$$

Since for all positions 13, 6, and 9 in the filter array are not set to 1, element 12 is not recorded.

For element 2:

$$h1(2) = (2 + 1) \mod 20 = 3$$
$$h2(2) = (2 \cdot 2 + 2) \mod 20 = 6$$
$$h3(2) = (3 \cdot 2 + 3) \mod 20 = 9$$

Since positions 3, 6, and 9 in the filter array are not set to 1, element 2 may have not been recorded.

Answer : None of the elements from 15, 1, 10, 3, 12, and 2 may have already been recorded according to the Bloom Filter.