Experiment No.5
Data Stream Algorithms: Implement Bloom Filter using any programming language
Date of Performance:14/08/2023
Date of Submission:21/08/2023

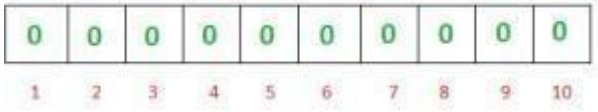
Aim: Data Stream Algorithms:

Implement Bloom Filter using any programming language

Theory:

A Bloom filter is a space-efficient probabilistic data structure that is used to test whether an element is a member of a set. For example, checking availability of username is set membership problem, where the set is the list of all registered username. The price we pay for efficiency is that it is probabilistic in nature that means, there might be some False Positive results. False positive means, it might tell that given username is already taken but actually it's not.

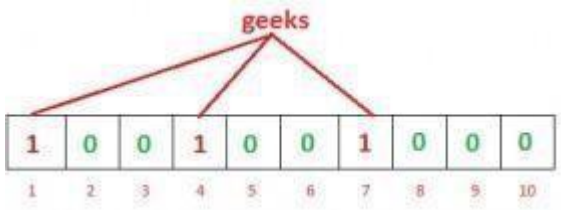
Working of Bloom Filter:-
A empty bloom filter is a bit array of m bits, all set to zero, like this –



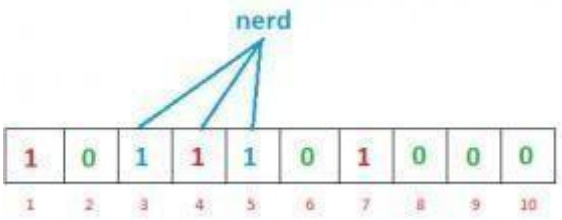
We need k number of hash functions to calculate the hashes for a given input. When we want to add an item in the filter, the bits at k indices h1(x), h2(x), … hk(x) are set, where indices are calculated using hash functions.
Example – Suppose we want to enter "geeks" in the filter, we are using 3 hash functions and a bit array of length 10, all set to 0 initially. First we'll calculate the hashes as follows:

h1("geeks") % 10 = 1
h2("geeks") % 10 = 4
h3("geeks") % 10 = 7

Note: These outputs are random for explanation only. Now we will set the bits at indices 1, 4 and 7 to 1



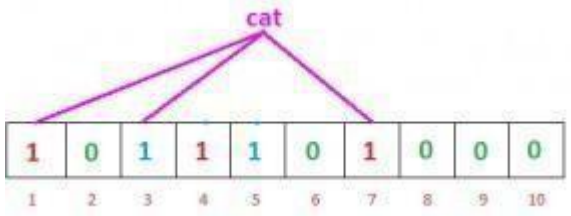
Again we want to enter "nerd", similarly, we'll calculate
hashes h1("nerd") % 10 = 3
h2("nerd") % 10 = 5
h3("nerd") % 10 = 4
Set the bits at indices 3, 5 and 4 to 1



Now if we want to check "geeks" is present in filter or not. We'll do the same process but this time in reverse order. We calculate respective hashes using h1, h2 and h3 and check if all these indices are set to 1 in the bit array. If all the bits are set then we can say that "geeks" is probably present. If any of the bit at these indices are 0 then "geeks" is definitely not present.

False Positive in Bloom Filters
The question is why we said "probably present", why this uncertainty. Let's understand this with an example. Suppose we want to check whether "cat" is present or not. We'll calculate hashes using h1, h2 and h3

h1("cat") % 10 = 1
h2("cat") % 10 = 3
h3("cat") % 10 = 7
If we check the bit array, bits at these indices are set to 1 but we know that "cat" was never added to the filter. Bit at index 1 and 7 was set when we added "geeks" and bit 3 was set we added "nerd".



So, because bits at calculated indices are already set by some other item, bloom filter erroneously claims that "cat" is present and generating a false positive result. Depending on the application, it could be huge downside or relatively okay.
We can control the probability of getting a false positive by controlling the size of the Bloom filter. More space means fewer false positives. If we want to decrease probability of false positive result, we have to use more number of hash functions and larger bit array. This would add latency in addition to the item and checking membership.
Operations that a Bloom Filter supports

insert(x) : To insert an element in the Bloom Filter.
lookup(x) : to check whether an element is already present in Bloom Filter with a positive false probability.
NOTE : We cannot delete an element in Bloom Filter.

Probability of False positivity: Let m be the size of bit array, k be the number of hash functions and n be the number of expected elements to be inserted in the filter, then the probability of false positive p can be calculated as:

P=\left (1-\left [1- \frac {1}{m} \right]^{kn} \right)^k

Size of Bit Array: If expected number of elements n is known and desired false positive probability is p then the size of bit array m can be calculated as :

m= -\frac {n\ln P}{(ln 2)^2}

Optimum number of hash functions: The number of hash functions k must be a positive integer. If m is size of bit array and n is number of elements to be inserted, then k can be calculated as :

k= \frac {m}{n} ln

2 Space Efficiency

If we want to store large list of items in a set for purpose of set membership, we can store it in hashmap, tries or simple array or linked list. All these methods require storing item itself, which is not very memory efficient. For example, if we want to store "geeks" in hashmap we have to store actual string " geeks" as a key value pair {some_key : "geeks"}.
Bloom filters do not store the data item at all. As we have seen they use bit array which allow hash collision. Without hash collision, it would not be compact.

Choice of Hash Function

The hash function used in bloom filters should be independent and uniformly distributed. They should be fast as possible. Fast simple non cryptographic hashes which are independent enough include murmur, FNV series of hash functions and Jenkins hashes.
Generating hash is major operation in bloom filters. Cryptographic hash functions provide stability and guarantee but are expensive in calculation. With increase in number of hash functions k, bloom filter become slow. All though non-cryptographic hash functions do not provide guarantee but provide major performance improvement.

CODE:

```
##bloomfilter.py
import math
import mmh3
from bitarray import bitarray
class BloomFilter(object):
    def__init (self, items_count,
        fp_prob): self.fp_prob = fp_prob
        self.size = self.get_size(items_count, fp_prob)
        self.hash_count = self.get_hash_count(self.size, items_count)
        self.bit_array = bitarray(self.size)
        self.bit_array.setall(0)
    def add(self, item):
        digests = []
        for i in range(self.hash_count):
            digest = mmh3.hash(item, i) %
            self.size digests.append(digest)
            self.bit_array[digest] = True
    def check(self, item):
        for i in range(self.hash_count):
            digest = mmh3.hash(item, i) %
            self.size if self.bit_array[digest] ==
            False:
                return
        False return
        True
    @classmethod
    def get_size(self, n, p):
        m = -(n * math.log(p))/(math.log(2)**2)
        return int(m)
    @classmethod
    def get_hash_count(self, m,
        n): k = (m/n) * math.log(2)
        return int(k)
```

```
##bloom_test.py
from bloomfilter import
BloomFilter from random import
shuffle
n = 20 #no of items to add
p = 0.05 #false positive probability
bloomf = BloomFilter(n,p)
print("Size of bit array:{}".format(bloomf.size))
print("False positive Probability:{}".format(bloomf.fp_prob))
print("Number of hash functions:{}".format(bloomf.hash_count))
word_present =
['abound','abounds','abundance','abundant','accessible',
            'bloom','blossom','bolster','bonny','bonus','bonuse
            s',
            'coherent','cohesive','colorful','comely','comfort',
            'gems','generosity','generous','generously','genial'
            ]
word_absent =
           ['bluff','cheater','hate','war','humanity',
           'racism','hurt','nuke','gloomy','facebook',
           'geeksforgeeks','twitter']
for item in
   word_present:
   bloomf.add(item)
shuffle(word_present)
shuffle(word_absent)
test_words = word_present[:10] + word_absent
shuffle(test_words)
for word in test_words:
   if bloomf.check(word):
      if word in word_absent:
         print("'{}' is a false
      positive!".format(word)) else:
         print("'{}' is probably present!".format(word))
   else:
      print("'{}' is definitely not present!".format(word))
```

OUTPUT:

```
Size of bit array:124
False positive
Probability:0.05 Number of
```

hash functions:4 'war' is
definitely not present!
'gloomy' is definitely not present!
'humanity' is definitely not
present! 'abundant' is probably
present! 'bloom' is probably
present! 'coherent' is probably
present! 'cohesive' is probably
present! 'bluff' is definitely not
present!

'bolster' is probably present!
'hate' is definitely not present!
'racism' is definitely not present!
'bonus' is probably present!
'abounds' is probably present!
'genial' is probably present!
'geeksforgeeks' is definitely not
present! 'nuke' is definitely not present!
'hurt' is definitely not present!
'twitter' is a false positive!
'cheater' is definitely not present!
'generosity' is probably present!
'facebook' is definitely not present!
'abundance' is probably present!

CONCLUSION:

Hive offers a SQL-like interface designed for querying extensive datasets stored in distributed storage systems. It's a prevalent choice within the Hadoop ecosystem, particularly for data warehousing and analytical tasks. In this instance, we established a Hive database, outlined the table structure, imported data into it, and executed fundamental descriptive analytics and statistical operations. Hive's robust capabilities make it well-suited for managing large datasets, and its SQL-like syntax ensures accessibility for users well-versed in relational databases. The precise queries and analytical procedures conducted depend on the data's characteristics and the insights sought after.