Assignment 7: The Great Firewall of Santa Cruz: Bloom Filters, Linked Lists, Binary Trees and Hash Tables

Nelly Sin

November 2021

1 Introduction

In this assignment, we are detecting and blocking words that people people use. At first, I was very confused on how this works, but I slowly by bits and pieces are able to understand the result of the assignment with the help of the ADTs. This assignment deals more with bit vectors considering the the hash tables and bloom filters are the bit vectors.
Imagine ruling over a country and you decided to block out certain things your citizens say on the internet. In order to do so, we must provide a way to reward their good deed of not using the words and discourage them to ever use such words that can be hurtful or offensive.

2 Bloom Filters

Bloom filter is a bit vector ...
Operate operate more than one hash functions.
Primary secondary and tertiary. Salts is for changing up for the hash functions. A good hash functions must be fast, every single time h(x) => y it will produce the same number. Must be deterministic. How do we fake the hash functions?
By providing the salt.
The Bloom filter will added first and then the hash tables.
Neither BF or the HT should grow.
Example input: ”I use python in cse13s”
Parsing all the words then for each word we put in the bloom filter
How do we check if the word is in the BF?
Use BF probe, when it's hashed 3 times, and if it is inserted it 3 times and we would know if it's in the bloom filter.
We check 3 indices because it is less likely for a collision – all three having the same index, thus it collides.

Every time we see all 3 indices are set in the bf, we check Hash Table for confirmation that it’s not a false positive.

bad = [”cse13s”] b/c it does not have a newspeak

revise = [”python” $- >$ ”slow”]

We have a list of words to revise to print.

3 Hashing with the SPECK

SPECK is a lightweight block of ciphers publicly released by the National Security Agency.

Encryption is the process of taking some file you wish to protect, usually called plaintext, and transforming its data such that only authorized parties can access it.

The encryption algorithms to utilize the same key for both the encryption and decryption. And SPECK are symmetric-key algorithms. This means algorithms for cryptography that use the same cryptographic keys for both the encryption of plaintext and the decryption of ciphertext. The keys may be identical, or there may be a simple transformation to go between the two keys.

The SPECK block cipher has been provided for us. However we will be implementing them for our hash tables.

4 Bit Vectors

A bit vector is an ADT that represents a one dimensional array of bits, the bits in which are used to denote if something is true or false (1 or 0).

Most of the bit vector understanding and code are derived from assignment 5 and the professor’s code comments repository.

5 Hash Tables

Hashing is just a double checker of the bloom filter, however hash table can never return false positives.

Hash tables are used in data storage and retrieval applications to access data at a nearly constant time per retrieval. They require a storage space that is only greater than total space for the needed data. However, it’s useful because hashing avoids the non-linear access time of order and unordered lists and often the exponential storage requirements and direct access of large state spaces.

Chaining by using binary search trees.

Similar to a bloom filter, the hash table will be using salt and whenever a new oldspeak is being inserted we will be using binary search trees to resolve oldpseak hash collisions, another reason why the hash table contains an array of trees.

6 Nodes

Because binary search trees will be used to resolve hash collisions. And to initialize and prepare for the binary search trees for this assignment, each node contains oldspeak and its newspeak translation if it exists. The key to search with in a binary search tree is oldspeak. Each node, in typical binary search tree fashion, will contain pointers to its left and right children.

7 Binary Search Trees

Binary search trees, starting off the root if it is greater than then then you go to the right side (total ordering).
A Bloom filter can be represented as an array of m bits, or a bit vector. A Bloom filter should utilize k different hash functions. Using these hash functions, a set element added to the Bloom filter is mapped to at most k of the m bit indices, generating a uniform pseudo-random distribution. Typically, k is a small constant which depends on the desired false error rate , while m is proportional to k and the number of elements to be added.
In this case we will be taking the list of the prescribed words such as oldspeak and add the word into our bloom filter. If the words that the citizens use is added to the bloom filter, then we must give them a warning message.
In the process to reduce the chance of a false positive we must use three salts for three different hash functions. Salt can be a initialization vector or a key. Using the different salts with the same hash function results in different hashing.
Lecture slide 18 will be useful

8 Lexical Analysis with Regular Expression

We are using this to lowercase all the files. According the the assignment doc. the lexicon of badspeak and oldspeak/newspeak translations has been populated and then we can start filtering the words. The parsing should be before checking the word that is read in.
BF + BV + BST + HT => data table of bad words + fast querying
Given a parsing module, (given) if you give it a regular expression it will give a regular expression for that.
They are based off of state machines – Regular Expression:
”+” = 1 or more
”x” = 0 or more
”1” = or
”?” = optional (0 or 1)
= symbol set
(0—1)+ = binary
Means 1 or 0 in one of more times
0x([a-f 0-9])+ = hex

a-f means (a,b,c,d,e,f)
same with 0-9 (0, 1, 2, 3, 4, 5, 6, 7, 8, 9)
If the word is likely to be added to the bloom filter, using bfprobe we do not need to take action. However, if the hash table contains the word and the word does not have newspeak translation then we must notify the citizen who used the word is guilty and must send a message to and list the badspeak words that are used.
If the hash table contains the word, then the word does have newspeak translation –indicating the usage of rightspeak. We then reward the citizen by sending a message to them.
However, if the citizen is accused of using badspeak and newspeak. They are warned with a message that contains the usaged of mix speak and list the oldspeak words and the newspeak words that are used.

9 Citation

Special acknowledgements to the TA's, tutors, and Professor Long for helping me throughout the concepts and providing some advice for my pseudo code and publishing some pseudo code.
The code I have provided in my work are not all written by me, but by tutors and TA's that provided further help for me. And credit to the professor for providing additional help from his public repository and code comments that he sent to the class.