The Great Firewall of Santa Cruz: Design

Ralph Miller

December 3rd 2021

"Perhaps one did not want to be loved so much as to be understood."

George Orwell, 1984

1 Objective

In this assignment I will be writing a program that detects offensive speech using a Bloom filter and Hash table with a given set of bad words and new word translations. The user will be given a response detailing their transgressions and the offensive words used. To implement the Bloom filter and Hash table, a few ADT's will need to be created to support these functions. The deliverables for this assignment are listed below.

1.1 Deliverables

2 SPECK Cipher

The program containing the SPECK Cipher is included in the assignment. This is a block cipher that was publicly released by the NSA in 2013. The hash function requires two inputs. A salt and Char string. The salt for the has adds flavor to the output and will give us a repeatable result for an input Char string. We will be using the salts provided in salts.h The function will return a uint32 that we will be using for our Bloom filter and Hash table.

3 Bloom Filter

We will be using the Bloom filter to probabilistically detect if a read in word is flagged. We will be adding words to our bloom filter by hashing each bad word with three different salts. The three results of the salt are the index of the filter that will be set to 1 using our Bit Vector functions. After adding all of our bad words the bloom filter can be used to determine if a read in word from input is possibly a bad word. If all three results of the hashes exist in the filter. We will need to then check the Hash table to confirm that the word is actually a bad word. This is because it is possible for Hash Collisions to occur, even with a Bloom filter of indefinite size. If less than three results exist in the hash filter, then we can certainly conclude that the word is not a bad word and read the next word in.

3.1 Pseudocode

3.2 Salts

4 Hash Table

We will be using the Hash table to confirm if a word that was flagged as a bad word is actually a bad word. The Hash table contains the indices of nodes that hold values of bad words, improper words and their translations. When initializing the Hash table with bad words, each word is hashed with the salt for the index of the hash table.

4.1 Pseudocode

4.2 Salts

5 Bit Vectors

The Bit Vector functions are used to manipulate bits of the BitVector type. This is used in the Bloom filter and Hash table. The functions consists of a; constructor, destructor, acessor, and manipulator functions. The functions necessary are defined below.

5.1 Pseudocode

6 Nodes

The Node ADT provides functions for the Hash table and Binary search tree. Each node consists of four fields, a left right field which tracks child nodes, and a newspeak oldspeak field which holds the word of an oldspeak word and its possible newspeak translation. The functions necessary are defined below.

6.1 Pseudocode

7 Binary Search Trees

7.1 Pseudocode