Computer Science 320 S2 (2019)

Assignment 4

Due date Sep 28, 2019 23:59pm

Answer all of the following questions. There are 10 points, which contribute 5% of your total course marks. Submit a properly typesetted pdf file (LaTeX preferred) of your answers to Canvas before the deadline. There is no automarker for the Python program. To aid the markers, scanned handwritten solutions or late submissions are **NOT** accepted.

Email spam filtering

You are implementing a new email spam filtering for the University of Auckland. Given that you know the set A of **1 billion** trusted email addresses, e.g. ninh.pham@auckland.ac.nz. If the email comes from one of these trusted addresses, it is not a spam. Otherwise, it will be a spam. Since the server memory is very limited, you can not keep the list of trusted addresses on its memory (if an email address requires 25 bytes on average, it would take 25 GB to store A in the memory).

After searching on Google, you have found a very memory-efficient solution. That is, instead of keeping A on memory, you will construct a bit array B of size n representing for the set A. You choose $n = 8000,000,000$ bits so that you need only 1 GB of memory. The construction is as follows.

- B is initialized by 0s.
- You choose a hash function $h : a \mapsto [0, n)$. In other words, the input of h will be an string a (e.g. email address) and the hash value will be an integer ranging between 0 and n.
- For each trusted email address $a \in A$, you hash a into one of n buckets, and set that bit to 1. In order words, you simply set $B[h(a)] = 1$.

The filtering mechanism works as follows.

- When you receive a new email from the address a', you compute the hash value $h(a')$.
- If $B[h(a')] = 1$, you consider that this email is not a spam and let it go through.
- If $B[h(a')] = 0$, you consider that this email is a spam and discard it.

Theoretical questions for measuring the performance of the filtering (5 pts):

1. Illustrate that if a new email from the address $a \in A$, it always gets through (1 pts).
2. Given any position $0 \leq i < n$, what is the probability that $B[i] = 1$ (2 pts).
3. Given a spam email from the address $a' \notin A$, what is the probability that it gets through (2 pts).

Practical implementation for measuring the performance of the filtering (5 pts):

Write a Python script to implement this email spam filtering technique given the scaled setting. Your data structure B has size $n = 8000,000$ bits. For simplicity, assume that your trusted email addresses and spam email are presented as integers. In particular, trusted email addresses $A = \{1, 2, ..., 1000000\}$ and spam email addresses are any integer $x > 1000,000$. You are free to choose your hash function to hash an integer into the range $[0, n)$.

1. Verify that any email from the address $1 \leq a \leq 1000,000$ it always gets through (1 pts).
2. Compute the probability that a spam email going through your filter given this setting. (1 pts).
3. Generate 1000 random integers $x > 1000,000$ as spam email addresses and compute the number of spam emails going through your filter. Verify this value with your theoretical value from the step 2 (3 pts).