Challenge: BLOOM FILTERS

1 Preliminaries

1.1 Research

A Bloom filter is a probabilistic data structure used to test whether an element is a member of a given set. Do your research about bloom filter and answer the following questions (with explanation):

1. State the factors to estimate the optimal size of the bit array and the number of hash functions for a Bloom filter.

   1a. How do hash functions impact the performance and accuracy of a Bloom filter?

2. Do lookup operations on bloom filter always give correct results? If not, indicate the estimation method for false positive rate.

   2a. Does the insertion of elements affect the false positive rate in a Bloom filter?

   2b. Are there any known techniques or strategies to reduce the false positive rate in a Bloom filter?

3. Is it possible to estimate the number of elements stored in a Bloom filter?

   3a. Can a Bloom filter handle deletions of elements from the set it represents?

   3b. Can a Bloom filter be dynamically resized to accommodate more elements?

4. Build a comparison table between Bloom filter and Hashtable in terms of Space efficiency, Time efficiency, Insertion and Deletion possibly, queries's probabilistic answer and limitations.

1.2 Programming You are requested to simulate a simple account management system of a website. System includes Username and Password. The program should have the following requirements:

1. **Constraints:**
   - Username:
     - $5 <$ sizeof(Username) < 10.
     - Username must not contain spaces.
     - Username must not be the same as any registered Username.
   - Password:
     - $10 <$ sizeof(Password) < 20.
     - Password must not contain spaces and cannot be the same as username.
     - Password must include uppercase, lowercase, numbers and special characters.
     - Password must not match the weak password listed in the file <*WeakPass.txt*>.
2. **Operations**
   - *Registration*: Verify the validity of the account and store created information in the database.
   - *Multiple registrations*: Verify the validity of multiple accounts and store created information in the database.
     - The accounts are stored in the file <*SignUp.txt*>. Each line of the file contain 1 account with Username and Password are seperated by a single space.
     - Ouput file <*Fail.txt*> for accounts that cannot be created due to constraints violation.
   - *Log-in*: User needs to enter correct Username and Password to login. If success, give users the permission to change their password.
   - *Change Password.*

2 Group registration and Submission regulations

2.1 Group Registration

- This project requires a group of 3 - 4 students.
- Group ID is generated by concatenating the last 3 digits of each member's Student ID in ascending order.
-

  Example:

  - Given the student codes: *22127666 - 22127888 - 22127991 - 22127999.*

  → **Generated ID**: *666-888-991-999.*

2.2 Submission regulations

- The submission file must be in the following format: [**Group_ID.rar**] or [**Group_ID.zip**], is the compression of the [**Group_ID**] folder. This folder contains:
  - The report file must be presented as a document [**Group_ID.pdf**]. This file presented research answers from **1.1** and the information of code fragment (data structures, algorithms, functions) from **1.2**.
    - If your submission is a slideshow, there must be explanation in the *Note* part of each slide.
    - Information (Names, Student IDs) must be declared clearly on the first page (or first slide) of your report. Your working progress (Which option did you choose? How much work have you completed?) should be demonstrated on this page, too.
    - The report file should be **structured, logical, clear** and **coherent**. The length of the submission should not exceed 15 pages for the document file, and 30 pages for the presentation slide.
    - All links and books related to your submission must be mentioned.

  - The programming file must be put into a [**Group_ID**] folder. The code fragment must be clear, logical and commented.

- Submission with wrong regulation will result in a "0" (zero).
- Plagiarism and Cheating will result in a "0" (zero) for the entire course and will be subject to appropriate referral to the Management Board of the CLC program for further action.