CAPSTONE PROJECT

Code Snippet Manager

Project ID: 12

Team Name: BEAT 100%

Team Members:

1) Krish Makwana (202301103)
2) Dhruv Sanjaykumar Patel (202301024)
3) Dev Trivedi (202301150)
4) Neeraj Vania (202301060)

Github Repository Link :

https://github.com/Krish-Makwana-1205/DS-project-Beat-100

Introduction

This report provides a detailed exploration of the data structures employed within a Code Snippet Manager application leveraging a Graphical User Interface (GUI). It dissects the reasoning behind these selections, investigates their time and space complexities, and offers insights into the data structures deemed unsuitable for this project.

Bonus Question:- The bonus question that we solved in making our code snippet manager was the use of natural language processing in our project though the code that we wrote for it's implementation is primitive related with modern NLP techniques. Please see the detailed description of the algorithm found in the pseudocodes section.

1. Project Overview

The project centered around the development of a Code Snippet Manager equipped with a user-friendly GUI. This application aims to streamline the organization, retrieval, and management of code snippets, fostering developer productivity.

2. Employed Data Structures: Hash Table

- **Definition:** A data structure that excels at rapid insertion, retrieval, and deletion of key-value pairs. It leverages a hash function to map unique keys (often code snippet identifiers) to their corresponding values (the actual code snippets). The hash function that we used is h(s) = tag[0] - 'a'. The rank of the first letter in the tag in alphabetical order. This function gives the same rank to 'c' and 'C' that is they collide.
- **Collision Resolution Technique:** Separate Chaining is employed to address collisions that arise when multiple keys hash to the same index within the table.

In separate chaining, each index in the hash table acts as a pointer to the head of a linked list, where elements that collide are stored.

- **Time and Space Complexity:** Under the assumption of uniform hashing (where each key has an equal probability of mapping to any slot in the table and the number of keys is not too great compared to 26(The number of letters), the average-case time complexity for search, insert, and delete operations is constant time but in terms of big Oh search = O(n), insert = O(1) and Delete = O(n), The worst case scenario being that all the incoming tags of snippets start with the same letter. However, the space complexity scales linearly with the number of elements stored (O(n)) within the hash table and also some more space is consumed in storing the synonyms for natural language processing(Or more accurately keyword searching).
- **Justification:** Hash tables were chosen due to their exceptional average-case performance in searching, inserting, and deleting code snippets. Since each code snippet likely possesses a unique identifier, hash tables provide efficient retrieval based on these identifiers, significantly accelerating the process of locating specific snippets within the manager.

3. Data Structures Not Considered

3.1 Stacks and Queues

- **Unsuitable Ordering Mechanisms:** Stacks and Queues enforce specific ordering disciplines:

    - Stacks adhere to a Last-In-First-Out (LIFO) principle, where the most recently added element is retrieved first.
    - Queues follow a First-In-First-Out (FIFO) approach, where the element that was added first is retrieved first.
- **Mismatch with Random Access Needs:** A Code Snippet Manager necessitates random access. Developers need the ability to retrieve any snippet at any given time, irrespective of when it was added. Stacks and Queues, with their predefined ordering, are not well-suited for this requirement.
- **Retrieval Challenges:** In a Stack, you can only access the element at the top. Retrieving a specific snippet within the stack would necessitate potentially removing (and potentially re-adding) multiple snippets that were added later. This becomes highly inefficient, especially for large collections.
- **Limited Functionality in Queues:** Similarly, queues only allow access to the element at the front. While you could iterate through the queue to find a specific snippet, this approach is linear in time complexity (O(n)), making it slow for large numbers of snippets.

3.2 Arrays

- **Predefined Size Limitation:** Arrays offer efficient random access (O(1) on average) – you can directly jump to any element using its index. However, arrays have a significant drawback in the context of a Code Snippet Manager: their size must be predefined.
- **Inflexible for Dynamic Collections:** As a developer's codebase grows and shrinks, the Code Snippet Manager should gracefully accommodate these changes. Arrays cannot dynamically expand or contract in size. If the predefined

size is insufficient, you'd need to create a new, larger array and copy all the existing snippets, which can be cumbersome and inefficient.

- **Resizing Bottleneck:** Resizing an array to accommodate new snippets often involves creating a new array, copying the existing elements, and then deleting the old array. This can be a time-consuming operation, especially for large collections of snippets.
- **Wastage with Unfilled Space:** Conversely, if the predefined size is too large, the array might have a significant amount of unused space. This is inefficient memory utilization, as the application is essentially reserving memory that isn't actively being used.
- **Unsuitable for Deletions:** Deleting elements from the middle of arrays can be complex and inefficient. It often involves shifting elements down the array to fill the gap, potentially impacting performance for frequent modifications.

3.3 Singly Linked Lists

While singly linked lists offer efficient insertion at the head (O(1)), their overall performance for random access and frequent modifications throughout the list can be less desirable compared to hash tables. Here are the reasons why singly linked lists wasn’t chosen, but we can see use of linked lists for separate chaining in the implementation of hash table:

- **Limited Random Access:** Traversing a singly linked list to locate a specific snippet requires starting from the head and iterating through each node until the target is found. This operation can become time-consuming, especially for large collections of snippets. Imagine a list containing hundreds or even thousands of snippets. Finding a specific one using a linear search through a singly linked list would be inefficient.

- **Unfavourable for Frequent Lookups and Modifications:** Since code snippet managers involve frequent retrieval and modification of snippets in no order, the linear search and pointer manipulation associated with singly linked lists become significant performance bottlenecks. Developers expect quick access to their snippets, and hash tables excel in this regard.

4. Pseudocodes

4.1 Insertion

Inserting takes O(1) time the first step is to calculate the hash function which is index = rank of first letter of tag, here lowercase ‘a’ and uppercase ‘A’ occupy the same rank. So we have a global hashing array with 26 spaces for which letter. In this array each pointer points to the head of a singly linked list which will come into play in the case of collision. Upon the condition of collision we insert in the linked list from the head to obtain an O(1) time complexity.

4.2 Searching

The search is an algorithm where in terms of big oh the time complexity if O(n). But the average time taken is much less than linear search and better than binary search when inputs are 100 or lower. And thus the hash function can benefit us to a greater extent since the snippets stored by users would be in around 100 are a bit more in usual cases.

4.3 Deletion

We only have a single type of deletion which is deletion by searching the tag. Since according to me deletion by index or rank of insertion makes little sense when managing snippets. The time complexity of Deletion would be O(n) + O(1) = O(n) since we are first searching and then deleting the node.

4.4 Editing

The user can change the code stored under a tag through this functionality; it would also take O(n) + O(1) time. In the same manner as deletion.

5.Natural Language Processing

Natural Language Processing is probably a too big name for the algorithm we are using. All the algorithm is doing is that it breaks the sentence into words then the words are individually compared with the words in the key which are separated by '_' and also synonyms stored in each snippet which are the words which could be replaced by them for example "insert" is could be replaced by "add". Now the number of words which match for each key is counted and stored as a pair with a snippet and the number of words which have matched into a vector now this vector is sorted in decreasing order with respect to the number of words. Finally this is pushed into the List Box so the most relevant option would come first. When used by a non - coder it gives a success rate of 50% where the snippet which they wanted came at the highest priority.

6. Additional Considerations

This report offers in-depth analyses of the time and space complexities associated with each implemented function within the chosen data structures. Furthermore, it explores the data structures that can be effectively utilized for separate chaining within hash tables. Lastly, the report delves into the potential drawbacks of using arrays and singly linked lists in this particular scenario.

7. Conclusion

The strategic selection of hash tables, coupled with a thorough understanding of their time and space complexities, has demonstrably contributed to the development of an efficient and adaptable Code Snippet Manager. This application empowers developers

to organize, retrieve, and manage their code snippets effectively, fostering a more productive development environment.

6. Reference

- **https://www.geeksforgeeks.org/hash-table-data-structure/**
- **https://www.geeksforgeeks.org/inserting-elements-in-an-array-array-operations/**
- **https://www.geeksforgeeks.org/time-complexities-of-different-data-structures/**