William Kurek

CS 1501

Writing Assignment 4

**The Anagram Problem:**

The anagram problem presented in this project required a recursive backtracking algorithm to compute all possible anagram solutions of a given input sting. The algorithm beings by calling a function with the following arguments: an empty, modifiable string object, which we will call the solution string, and a list data structure containing the characters of the input string. Also, the current number of words is stored as a global variable. In this function, the algorithm adds a single letter, or character, from the list to the initially empty solution string. The resulting string is then passed to another function to determine whether it is a word, prefix, or neither.

If the solution is neither a prefix or a word, the algorithm tries the next character in the data structure. If the given string is a prefix but not a word, a recursive call is made, repeating the process with the prefix as the solution string and the prefix characters as the new input string. If the solution string is a word, there are two other cases to consider before proceeding. In the first case, if the solution string contains all the characters, the string is then added to the anagrams solutions array. In the second case, if the solution string does not contain all the characters, a space is appended to the solution string. Then, another recursive call is made with the current string and the characters remaining from the input string. The number of words, implemented via a global integer, is incremented as well. Since the number of words is greater than one, this recursive call requires a different course of action: operating after the appended space on the recursive call.

**Runtimes:**

Rather than calculating the runtime of each test file via a stopwatch, I implemented a simple timer within the program which outputs the runtime of the algorithm. But, because test files four and five required several hours to execute on the MyDictionary implementation, runtime approximations, given previous results, were necessary.

|  |  |  |
| --- | --- | --- |
| **TEST FILE RUNTIMES (milliseconds)** | | |
|  | **MyDictionary** | **DLB** |
| **Test1.txt** | 2422 | 219 |
| **Test2.txt** | 8301 | 235 |
| **Test3.txt** | 214658 | 360 |
| **Test4.txt** | 621493 (estimated) | 641 |
| **Test5.txt** | 1228806 (estimated) | 4521 |
| **Test6.txt** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | 1060415 |

After thorough analysis of the test files and their respective runtime, it was apparent that multiple variables influenced the results.

First, the number of words in the dictionary has a direct impact on the dictionary load time. Also, the numbers of words could also have a negative effect on the search time. In the worst case, it is possible that each level of the DLB must be traversed, which would increase the asymptotic runtime. which could have a negative effect on the search time. But, for the MyDictionary class, the asymptotic runtime is directly impacted by the number of words in the dictionary.

Second, the number of characters in a word also influences the runtime. Depending on the dictionary, an increased number of characters has the potential to increase the number of multiple word solutions.

Third, the number of solutions also has the potential to increase the runtime of the algorithm. For example, sorting the solutions at the end of program would increase the asymptotic runtime, because sorting is not executed in constant time. So, many solutions will have a negative impact on the runtime of the program. But, because comparison-based sorting can be run in O(nlgn) time, the impact of more solutions is minimal, as it is still far more efficient than the much larger O(k!) anagram generation algorithm.

Finally, the number of words in a solution affects the runtime of the program as well. As the number of words in a solution increases, the number of repetitions necessary increases as well. Adding up the recursive calls asymptotically will give us the overall runtime. However, this is bounded by the length of the string, as a string of length N can at most have N words to it. Therefore, in the worst case, the runtime is multiplied by N.

Following, the DLB algorithm performs better than the MyDictionary implementation. The worst-case runtime of the DLB algorithm is Theta(KS), because there are up to S comparisons to find the character on each level and K levels to get to the end of the key, where S is the size of the alphabet and K is the length of the key. But, because most of the levels will have very few characters, a DLB search will more likely require Theta(K) time. In comparison, assuming there are N values in the dictionary to search from, the worst-case runtime of MyDictionary is O(N). Due to its’ dependence on the number of values in the dictionary, MyDictionary is extremely inefficient. Therefore, because the DLB search time is independent of N, it is far more efficient than MyDictionary.

**Implementation Comparison:**

Although the MyDictionary and DLB versions of DictInterface follow a similar structure and commute the same output, their respective implementations are quite different.

In each implementation, data is stored different ways, which directly influences memory consumption. In the MyDictionary implementation, data strings are stored in a sorted list data structure. Unlike DLB, for each addition to the dictionary, the entire data string is stored at a new index in the list. As a result, as the number of strings in the dictionary increases, the amount of memory used increases as well. In the DLB implementation, dictionary entries are stored in a linked list. Each entry of the linked list contains a character, a child reference node, and a sibling reference node. But, rather than storing each string in its entirety, only the character values that are used are stored in the linked list, which can save a significant amount of memory in comparison to MyDictionary.

Further, MyDictionary and DLB implement different prefix search methods. In the MyDictionary implementation, the algorithm iterates through a list data structure until the end of the list, or until the key is found. Then, it iterates through the key and the current String in the list character by character. If the end of the string in the list has been reached, the algorithm stops with that string, and proceeds to the next one. If the current character in the key is less than the current character in the string, the search is terminated all together. Otherwise, the algorithm determines whether the characters match. Also, it is determined whether the end of the key and/or string has been reached, and respective Booleans are set. As a result, the algorithm can determine if the string is a word, prefix, both, or neither.

In comparison, the DLB implementation utilizes recursion in the search method. At each level of the linked list, the algorithm follows the right sibling references of each node until the character is matched or until NULL is reached. If the character is matched, the algorithm proceeds to the next level of the linked list. The DLB proceeds down the levels of the structure until the size of the string matches the size of the key. If the child reference of the current node is not null, then the string is a prefix. If the end of the string character matches the end of the key and the sibling reference of the current node is not null, then the string is a prefix and a word. Otherwise, the string is only a word.

As a result of this implementation, the search time at each level is potentially linear in regard to the number of characters in the alphabet. Since the number of characters in the key is equal to the number of levels necessary to get to the end of the key, the potential linear run time at each level can significantly affect the overall runtime. But, since this is worst-case runtime, it is highly unlikely. Since most of the levels in the DLB implementation will have very few characters, the overall runtime for search is equal to the number of characters in the key, in the average-case.

**Conclusion:**

Although it has a worst-case runtime of O(SK), the DLB implementation of DictInterface is superior to MyDictionary in testing for prefixes. In most cases, DLB will have a runtime of O(K), which is significantly better than that of MyDictionary. Also, since DLB only stores necessary character values rather than the entire string, it requires far less memory as well. Since the MyDictionary implementation stores each dictionary entry string in its entirety, it requires a significant amount of memory. For this reason, DLB is also more efficient regarding memory consumption as well. Overall, DLB is the superior algorithm in comparison to MyDictionary.