Harris Seth
Von Quilon

PA4 ReadMe

Welcome to the read me for the fourth assignment: search. In this assignment we were to implement a grep-like search utility that would scan for words and tell us which files contained those words. The program starts by running the indexer program from PA3 to generate the file that we will use in PA4.

The internal data structure that is used to search through the file returned by indexer is a hash table. The hash table is a typical array of linked lists and collisions are dealt with via chaining. Strings are hashed through the Fowler-Noll-Vo (FNV) hash function.

So first we go through the file once in order to determine how many words there are and we use that to determine the size of the hash table. We then go through the file again, after rewinding, and based on the files structure insert into the hash table. So for example a word entry is of the format "<list>" + word + file1 + count1 + … +filen + countn + "</list>". So once we encounter "<list>" we run word through the hash function and find a place in the array to insert the word. We then find the location of word in the linked list. Now each node in the linked list also has a linked list inside of it, a linked list of files. We then simply insert the file name into the file linked list. We go through all the files in this fashion until we land on "</list>", at which point we stop the loop and seek out the next word. This continues until we reach the end of the file.

The program continually asks the user for input, and returns the words that they search for. If the input is not in the correct format then we simply tell them so and ask for input again. If, say the file is not correctly formatted then the hash table build up won't run. If the file is empty then the hash table will simply be empty. These properties are thanks to the fscanf() which will not run if we have reached the EOF.

The efficiencies are the same as the efficiency of a typical hash table implementation. Average space O(n), average search O(1), average insert O(1). And the worst cases are O(n) for all.