CENG 218 Data Structures 2021 -2022- HW 1: Document Indexing

Due: 03/06/2022 - 23:59

In this project, you will design a very simple yet effective document retrieval system that will respond to single word queries. This system will allow users to enter a single word, then the system will return a list of documents containing this word.

Consider a scenario where there are 10,000 text files, and you would like to develop a program that will let users search for a particular keyword among all documents. Users are interested in finding documents containing this keyword.

Simple and first naive attempt to develop a solution to this problem would be to construct a program that read keyword from a user and then the program will go over(scan) all files and list the names of the files containing this keyword. This approach would be the most expensive method in terms of efficiency and time.

As an alternative to a naive approach that is a more efficient approach to the problem of determining documents containing the given word(or more formally a single word query) is achieved by a technique called indexing. Document indexing in its simplest definition is organizing and storing documents for later retrieval based on words they contain.

Hypothesis: Time complexity of searhing on binary search tree is O(lgn) and time complexity for linked list is O(n). On the other hand, adding an item into the binary search tree is O(logn) while inserting an item to the front of linked list is O(1).

Your task is to design an experiment that will index the documents by the words in their contents. Write a report that will compare the performance of binary search tree and linkedlist. According to the hypothesis given above, keep note of how long indexing of all the documents and search operations take on a) linkedlist b) BST using the contents of the files. You are expected to clearly present your results in tables or graphs etc. in your report.

The system is expected to be a simple menu-driven that will first start the indexing by reading file location. Once the index is created the system will indicate the end of index creating with an appropriate message and then the following menu will appear that will enable the user to benefit from the system.

Simple Document Retrieval System
1. Enter a single keyword to list the document(s)(file names) 2. Print the top 10 words that appeared most frequently 3. Print the top 10 words that appeared least frequently 4. Exit

Option:

Design Guidelines

While reading text files(or documents):

- Lowercase all words,
- Get all words, where a word is a string of alpha characters terminated by a non-alpha character (white space is not alpha). The alpha characters are defined to be [a-z]. Therefore, the sequence of characters for the word “apple+78&’^+orange” would be ‘apple’ and ‘orange’.

Limitations and Assumptions

1. The collection of documents is closed (content and number of documents are fixed and will never change).
2. Each document stored in a single text file. Hence if there are 10,000 documents, then there are 10,000 text files. (Collection of documents is provided on webonline)

Figure 1 and Figure 2 summarizes the aim of this project.

File Name (or Document)	Content of File
1.txt	Pease porridge hot, pease porridge cold.
2.txt	Pease porridge in the pot.
3.txt	Nine days old.
4.txt	Some like it hot, some like it cold.
5.txt	Some like it in the pot.
6.txt	Nine days old.

Figure 1: Small set of files and their content.