CSC 255

Unit 5 Programming Assignment 1 [50 points]

Inverted File

An inverted file or inverted index is a useful data structure for maintaining the index of a book. Any book B, which can be viewed as an unordered list of words, has an associated index, I, which maintains a lexicographically ordered list of words such that each word, w, in I, is paired with the associated page number(s) in B where w appears. Implement an efficient algorithm for constructing I from B using the built-in (i.e., standard library) map or dictionary data structure from your preferred programming language (i.e., python, C++, or Java).

First, we must consider how we will transform B into I. I is a map or dictionary such that the keys are the distinct words in B and the values will be a set of the associated page number(s) where the words appear. Observe the following page from the index of an algorithms textbook (Figure 1) and the respective key, value pairings where each key is a unique word and the set of values associated with the word are the page number(s) where the word appears (e.g., double hashing, 419; CPU, 111; edge, 302, 620). We will use a set data structure to store the page numbers. Since a set does not repeat replicates, it would be the best option for storing the page numbers. Use the built-in set data structure from your preferred programming language to store the values of the associated keys for I.

Index

Figure 1

For practical purposes let's use the file *"Orwell 1984.txt"* and assume each page has 250 words.

Write a program that will create the index, I, for this book and write the index to a file named *"Orwell 1984 Index.rtf"* so that the format is similar to the example index above. **[40 points]**

Include functionality testing of the index, I, which includes a lookup function for a desired key/word which returns the set of pages associated with that word. If the word is not found, an appropriate message will be returned. **[10 points]**