Lab 3

B-Tree and Indexing

- **Problem Statement:**

  Implementing B-tree and a simple search engine application that utilizes B-tree for data indexing.

- **Search Engine:**
  - **Design:**

    The search engine is built mainly with a B-tree used for data indexing. The B-tree have word as keys and each key is attached by a value of hash map of IDs as key and ranks as values.

    On indexing a web page, it is parsed using a dom parser then all words in the text are extracted. The value of already stored keys is updated to have the new ID with the new rank, and if a word is unfound the it is inserted.

    Indexing a directory is similar to indexing a page, where we apply the same algorithm for all files in the directory.

    On deleting a web page, the page is parsed and all words are extracted, every word is saved as a key of a map with a value of all its IDs and ranks. We apply search on B-tree for all the words to update their value by removing all the deleted IDs.

    On search, the value of the word is fetched from the B-tree, then all IDs and ranks of this word are returned in a list of Search Result.

    On search by multiple words, w apply search for all words of the sentence, then we return a list of search result with common IDs for all words and minimum rank for each.

➢ Analysis:

- **Time Complexity:**
  1. All operations in B-tree (search, delete and insert) are $O(\lg_t h)$, where h is the height of the tree.
  2. IndexingWebPage is O(n lgh), where n is the number of words in all documents and h is the height of the b-tree used for indexing.
  3. IndexingDirectory if O(mn lgh), where m is the number of files in the directory, n and h are as defined previously in 2.
  4. deleteWebPage is O(n).
  5. searchByWord is O(lg h).
  6. searchByMultipleWords is $O(m^2 + n \lg h)$, where m is the number of documents, n is the number of words of the sentence and h is the height of the B-tree.
- **Space Complexity:**
  1. For B-tree: O(n), where n is the number of keys
  2. Fore Search Engine: O(n), where n is the number of all words in all documents.

Prepared by:

• Eman Rafik	ID:11
• Reham Mohamed Naguib	ID:17
• Yomna Gamal el-Din	ID:60