Lab 3
B Tree and Indexing

1 Overview

In this assignment, you're required to implement a B-tree and a simple search engine application that utilizes the B-Tree for data indexing.

2 Introduction

2.1 B-Tree

B-trees are balanced search trees designed to work well on disks or other direct access secondary storage devices. Unlike the red-black trees, B-tree nodes can store multiple keys and have many children. If an internal B-tree node x contains x.n keys, then x has x.n + 1 children. The keys in node x serve as dividing points separating the range of keys handled by x into x.n + 1 sub-ranges, each handled by one child of x.
Please check the reference[1] for more details about the B-Trees.

2.2 Simple Search Engine

You will be given a set of Wikipedia documents in the XML format and you are required to implement a simple search engine that given a search query of one or multiple words you should return the matched documents and order them based on the frequency of the query words in each wiki document, please check the requirements section for more details.

3 Requirements

3.1 B-Tree

You are required to implement a generic B-Tree where each node stores key-value pairs and maintains the properties of the B-Trees. The IBTree and IBTreeNode interfaces (attached in the notes section) should be implemented.

3.2 Simple Search Engine

You will be given a set of Wikipedia documents in the XML format (you can download the Wikipedia data sample from here), and you are required to parse them (using Java DOM XML parser is recommended) and maintain an index of these documents content using the B-Tree to

be able to search them efficiently. The ISearchEngine and ISearchResult interfaces (attached in the notes section) should be implemented:

3.3 Report

You are required to deliver a report that contains

- Your code design for the simple search engine application
- The time and space complexity for the provided interfaces functions of both the B-Tree and the simple search engine.

4 Notes

- You can work in teams of four.
- The Wikipedia data sample can be downloaded from this link.
- The code interfaces can be downloaded from this link.
- Each team should submit the code and report online to this google form link.

5 References

1. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). Introduction to algorithms.