Task 1: B+Tree

Command to Run: java btindex -p <pagesize> <heapfile>

<pagesize> should reflect the number of bits in the heap file, such as heap.4096 where pagesize is 4096

<heapfile> should be the heapfile used from assignment that is attached to the project called 'heap.4096'

B+Tree Implementation

The B+Tree implementation uses 3 classes:

- BTree, to perform actions to the B+Tree
- RecordPair, to store the key and value pair
- Node, to store data of each node, such as keys, children, siblingpointer etc.

To insert, we:

- We initialise the root if it's null, otherwise we insert into the target leafnode
- Else if the leafnode we need to insert into is full, we replace the target leafnode with an overflow node, where the leafnode's order is order + 1 and insert the new key and value into the overflow node and then call split

To split, we:

- We create two temporary left and right nodes, and then we split the overflow node to the two temporary nodes and push the first key of the right node to the parent node.
- We then check if the parent node is full, if so we call the split function.

Task 2: Range Query

Command to Run: java btsearch <heapfile> <indexfile> <start date> <end date>

<heapfile> should be the name of the heapfile, e.g. 'heap.4096'

<indexfile> should be the name of the indexfile e.g., 'index.4096'

<start date> should be the lower bound of the range search, e.g. '19800204'

<end date> should be the upper bound of the range search, e.g. '19990415'

Range Search Implementation

- We first load all the index values [Date, lineNum] into a Hashmap
- We get all hashmap values whose keys are in range and store into an arraylist
- Sort the arraylist in ascending order

- We then get all the records by the line number from the arraylist

Trial Number	Range Search w/o Index	Range Search w/ Index	Range Search w/ Index (Only reading heapFile Time)
1	11908ms	31939ms	3887ms
2	1710ms	16580ms	294ms
3	26628ms	36171ms	5575ms

Trial 1 used the dates <18800215> to <19710623>

Trial 2 used the dates <18810215> to <18850215>

Trial 3 used the dates <16500215> to <20100623>

The range search with index will always take longer than the range search without index because there is the requirement to load in all values whose keys are in range or equal the date range. We can see that from the table above. However, if we don't consider this and only look at the time taken to compare the heapfile and the arraylist of line numbers, we can see that it is much faster compared to without index.