CSC 571-485B / SENG 480A: Summer 2017
Assignment 2
Due: May 31, 2017, 11:55 am

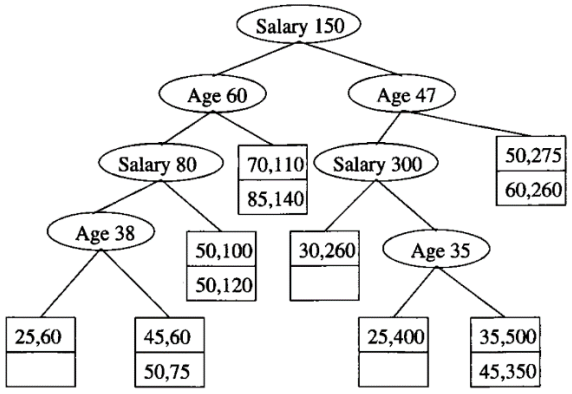
1. (4 pt) Suppose we have B-tree nodes with room for three keys and four pointers, as in the examples of this section. Suppose also that when we split a leaf, we divide the pointers 2 and 2, while when we split an interior node, the first 3 pointers go with the first (left) node, and the last 2 pointers go with the second (right) node. We start with a leaf containing pointers to records with keys 1, 2, and 3. Re then add in order, records with keys 4, 5, 6, and so on. At the insertion of what key will the B-tree first reach four levels? Draw the final tree.

2. (3 pt) Redo the example in slides titled "Attempt at using B-trees for MD-queries" under the assumption that the range query asks for a square in the middle that is $n \; x \; n$ for some *n* between 1 and 1000. How many disk I/O's are needed? For which values of *n* do indexes help?

3. (4 pt) Suppose we store a relation R(x,y) in a grid file. Both attributes have a range of values from 0 to 1000. The partitions of this grid file happen to be uniformly spaced; for x there are partitions every 20 units, at 20, 40, 60, and so on, while for y the partitions are every 50 units, at 50, 100, 150, and so on.
a) How many buckets do we have to examine to answer the range query
SELECT * FROM R
WHERE 310 < x AND x < 400 AND 520 < y AND y < 730;

b) We wish to perform a nearest-neighbor query for the point (110,205).
We begin by searching the bucket with lower-left corner at (100,200) and upper-right corner at (120,250), and we find that the closest point in this bucket is (115,220). What other buckets must be searched to verify that this point is the closest?

4. (2 pt) Show a possible evolution of the tree of in figure if we insert the points (20,110) and then (40,400).



5. (4 pt) Build an R-tree index with M=3 using the following sequence of rectangles:
(1,1,3,3) (2,2,4,4) (1,4,3,6) (5,4,7,6) (6,3,8,5) (7,2,9,4) (8,5,10,7) (8,6,9,7)
Redraw the tree each time an insertion is done.

6. (6 pt) Consider the following fragment from a collection of documents. Assume that these three documents are the only documents in the collection that contain the words “dog” or “cat”.

Document id	Document Text
234569	The phrase "fight like cats and dogs" reflects a natural tendency for the relationship between the two species to be antagonistic. However, sometimes the two species can be friends.
234578	Dogs and cats can have a bad relationship.
234839	Cats are furry.
234879	Dogs are man's best friend.

1. Write the inverted index posting lists for terms “cat” and “dog”.

2. Write the compressed form of the posting list for dog.

3. Calculate the cosine similarity of each of the above documents with query
   q: cat and dogs.
   Use stemming and remove stop words from the documents and query.
   Use only TF (ignore IDF).