Assignment 4

Due date: March 27, 2002 **Due in class**

Question 1:

You are given these documents D1..D6, and you are asked to generate an inverted index for all these documents (messages) without stemming words but after eliminating the following stopwords: the, this, with, a, of, and, it, for, is, but, has, no, in.

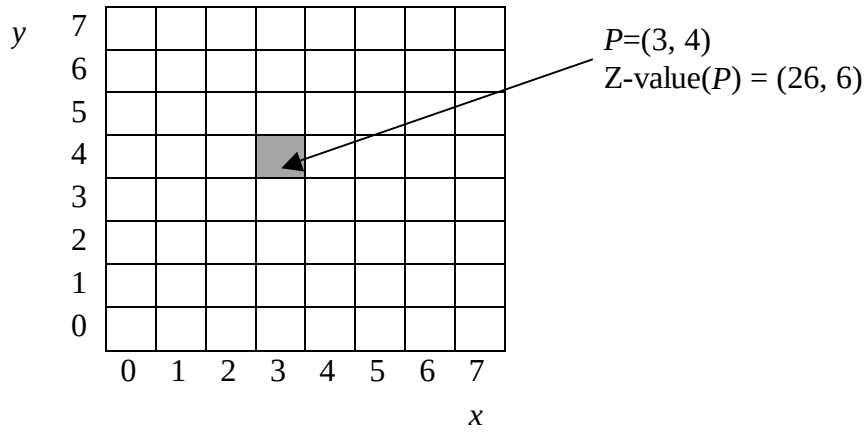
D1: This tutorial gives a brief overview of working with AWT and Swing.
D2: This book on Java GUI introduces programming with java and Swing.
D3: The tutorial is mandatory. It introduces the book.
D4: Java programming gives expertise in Java and opportunities for working.
D5: The book introduces Java but is brief and has no Swing or AWT in it.
D6: The mandatory tutorial and gives no opportunities for programming.

1) Explain how you would generate the Inverted Index and give the generated inverted index for these 6 documents assuming the inverted index would also keep track of the term frequencies.

2) If we submit the query “Java book with AWT and Swing”, what are the documents that would match if we use disjunctions of words and use the same stopwords? Rank the results by, first, the query completeness, and second, term frequency.

Question 2

1) Assume a $2^L \times 2^L$ grid where L is an integer. Give a pseudo-code algorithm that computes the decimal c of the Z-value(P)=(c, l) for any given cell P=(x, y) in the grid. E.g.:

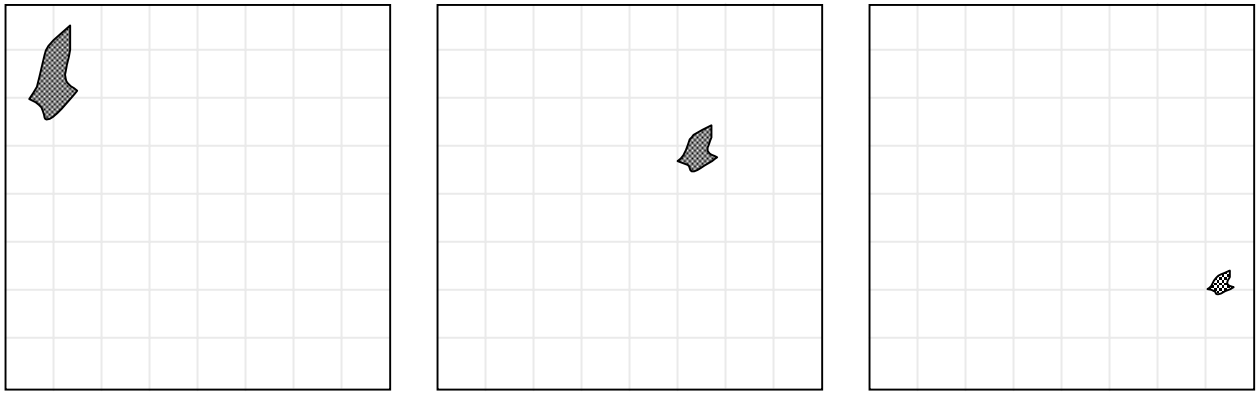


The algorithm should not be based on a recursive procedure that partitions the space, but on the following two procedures, which you can call in your pseudo-code:

- *decimal_to_binary*(n, b): computes the binary representation b of an integer n.
- *binary_to_decimal*(b, n): computes the decimal value n of a bit-array b.

where $0 \leq n < 2^{2L}$ is an integer and b is a bit-array of length $2L$, which contains the binary representation of n – including leading zeros if necessary.
Tip: Think about the binary representation of the x and y coordinates of a cell and their relation to a recursive partitioning of the space into halves.

2) Compute the Z-Values of the minimum enclosing cell (as explained in class) for the following polygons. Draw a line on the grid whenever a division of the space is necessary.



3) Give a simple pseudo-code algorithm that computes the spatial join of objects stored in two R-trees R_1 and R_2. You can assume that both R-trees R_1 and R_2 have the same height.

Question 3:

Assume the following DTD stored at http://www.csbooks.biz/bks.xml:

```
<!ELEMENT Reviews  (book* )>
<!ELEMENT book  (title,  (author+ | editor+ ), publisher, price, reviewer, review )>
<!ATTLIST book  year CDATA  #REQUIRED >
<!ELEMENT author  (last, first )>
<!ELEMENT editor  (last, first, affiliation )>
<!ELEMENT reviewer  (last, first)>
<!ELEMENT review (#PCDATA)>
<!ATTLIST review score (high | low | average) #REQUIRED>
<!ELEMENT title  (#PCDATA )>
<!ELEMENT last  (#PCDATA )>
<!ELEMENT first  (#PCDATA )>
<!ELEMENT affiliation  (#PCDATA )>
<!ELEMENT publisher  (#PCDATA )>
<!ELEMENT price  (#PCDATA )>
```

1) Write an XML-QL query to find all book titles and reviewer name of books published by MIT-press in 2001 and with a high review score.

2) Write an XML-QL query to find all book titles by reviewer. The required result should have a list of reviewers (first and last name) with their respective book titles they reviewed and the review score they gave.

3) Write an XPath query to find all book titles of books reviewed with a low score.

4) Write an XPath query to find author names of books published in 2002 and reviewed by Paul Leong.