

NATIONAL UNIVERSITY OF SINGAPORE

SCHOOL OF COMPUTING

MID-TERM TEST FOR

CS 5244: DIGITAL LIBRARIES

7 September 2004

Time Allowed: 1 Hour

**INSTRUCTIONS TO CANDIDATES**

1. This test paper contains FIVE (5) questions and comprises FIVE (5) printed pages, including this page. Please note that the questions are unequally weighted.
2. This is an open book exam. You may use any resources that you have brought into class but you may not consult any on-line resources during the exam. You are allowed to use a calculator for any numerical calculations.
3. Please limit your answers to the spaces provided.
4. You may use pencil or pen to do the exam.
5. Please sign below **after completing** the test to certify that your work is done independently of others.

I certify that all of the answers to this exam are my work only.

Name: \_\_\_\_\_

Signed: \_\_\_\_\_

Matriculation Number: 

--	--	--	--	--	--	--	--	--	--	--	--

EXAMINER'S USE ONLY		
Question	Max	Score
Q1	16	
Q2	4	
Q3	6	
Q4	16	
Q5	8	
Total	50	

1. (16 points) A printed text document can be converted to digital formats by a choice of methods:

- (i) digitization by scanning
- (ii) digitization plus optical character recognition
- (iii) retyping with SGML markup

(a) (12 points) What are the advantages and disadvantages of each of these three methods? You may use phrases or bullets to give your answers instead of complete sentences

(b) (4 points) Under what circumstances would a user be unsatisfied with all three digital manifestations and want to use the original printed copy?

2. (4 points) We saw that CCITT includes codes for run lengths of zero. We know that lines in CCITT can have a black run of zero length because the standard assumes the start of a line is a white run. What is the black run of zero used for?

3. (6 points) Given the query “word<sub>0</sub> word<sub>1</sub> word<sub>2</sub>”, construct two example documents  $d_1$  and  $d_2$  that would be ranked differently by the Euclidean and cosine similarity calculation methods. Show your calculations.

4. (16 points) Metadata and other annotations to a document can be stored either in-place (as part of the document) or in a separate file (standoff annotation). Baca (1998) terms this distinction as the *source* attribute of metadata (internal or external)

i) (4 points) Give an example of each: internal metadata and external metadata.

ii) (9 points) Discuss how the values of other metadata attributes might influence a designer's decision to use external or internal metadata.

iii) (3 points) Use your answer in ii) to explain why designers chose to use the source they did both examples in i). Be brief.

5. Briefly define **four** of the five terms (8 points):

Comanding:

Dublin Core:

CCITT Fax IV:

Raster image:

Word inflection:

End of Paper