CSC 2508S NAME _____
April 10, 2002

CSC 2508 -- Information: Quantification, Representation and Manipulation

TERM TEST

1	2	3	4	5	6	7	8	9	10	TOTAL
/12	/12	/12	/9	/8	/9	/7	/10	/12	/9	/100

[ADVICE: The test is long. Try all parts before spending much time on any single part.]

1. [12 marks]
Consider compressing a large body of English text (e.g., a set of newspaper articles) using each of *Huffman encoding* (**HUF**), *Arithmetic encoding* (**ARI**), and *Ziv-Lempel encoding* (**ZL**). For each question below, circle **ALL** correct answers.

(i). Which one(s) (if any) require that the relative frequencies (or probabilities) of letter occurrences be known (or at least estimated)?

HUF **ARI** **ZL**

(ii). Which one(s) (if any) send the same unique code word for each occurrence of a particular letter?

HUF **ARI** **ZL**

(iii). Which one(s) (if any) require use of a large main memory buffer for both compressing and decompressing?

HUF **ARI** **ZL**

(iv). Which one(s) (if any) may lead to an average encoding length that is smaller than the entropy of the body of text?

HUF **ARI** **ZL**

(v). Which one(s) (if any) can yield a code with lower redundancy if *pairs of characters* rather than individual characters are encoded?

HUF **ARI** **ZL**

(vi). Which one(s) (if any) can automatically adapt to *changing* letter frequencies without any special auxiliary actions?

HUF **ARI** **ZL**

2. [12 marks]
Consider the five binary encodings below for indicating the provinces in which Canadians live. (Keep in mind that many more people live in Ontario and Quebec than live in the other six provinces listed and ignore the provinces other than these eight.)

Province	Code A	Code B	Code C	Code D	Code E
Ontario	000	0000	00	00	00
Quebec	001	0001	01	01	01
Alberta	010	0010	10	100	100
Manitoba	011	0011	11	101	101
British Columbia	100	010	010	1100	0011
Saskatchewan	101	011	011	1101	0111
Newfoundland	110	10	100	1110	10011
Nova Scotia	111	11	101	1111	10111

A. Indicate in a few words the advantages and disadvantages of the five codes by filling in entries in the table below:

Code	Advantages (if any)	Disadvantages (if any)
A		
B		
C		
D		
E		

B. Rank the five codes from "best" to "worst" overall:

BEST:	2nd:	3rd:	4th:	WORST:

3. [12 marks]
Each round of a game consists of flipping three coins (all "fair"), a penny, a nickel, and a dime. Two outcomes are the *count*, which is the total number of "heads" among the three coins, and the *score*, which is the total value of the coins that come out "heads". More precisely:

$$P = \begin{cases} 1 \text{ if } \textit{the penny is heads} \\ 0 \text{ if } \textit{the penny is tails} \end{cases}$$

$$N = \begin{cases} 5 \text{ if } \textit{the nickel is heads} \\ 0 \text{ if } \textit{the nickel is tails} \end{cases}$$

$$D = \begin{cases} 10 \text{ if } \textit{the dime is heads} \\ 0 \text{ if } \textit{the dime is tails} \end{cases}$$

$$C = \textit{\# of coins that are heads}$$

$$S = P + N + D$$

For each part below, fill in a letter from A to L to indicate the correct answer from the column on the right. (All correct answers do appear on the right, but *more than one* question on the left may have the *same* answer.)

1. What is $H(P)$? _____
2. What is $H(C)$? _____
3. What is $H(S)$? _____
4. What is $H(N, D)$? _____
5. What is $H(N, C)$? _____
6. What is $H(C \mid P)$? _____
7. What is $H(S \mid N)$? _____
8. What is $H(S \mid C)$? _____
9. What is $H(C \mid S)$? _____
10. What is $H(D \mid C)$? _____
11. What is $I(N, S)$? _____
12. What is $I(C, S)$? _____

A. 0
B. .689 $= \frac{3}{4}\log_2(3) - \frac{1}{2}$
C. 1
D. 1.061 $= \frac{9}{4} - \frac{3}{4}\log_2(3)$
E. 1.189 $= \frac{3}{4}\log_2(3)$
F. 1.5
G. 1.811 $= 3 - \frac{3}{4}\log_2(3)$
H. 2
I. 3
J. 4
K. 6
L. 8

4. [9 marks]

A.

(i). Give two distinct reasons why is *is desirable* to represent information in as few bits as possible.

(ii). Give two distinct reasons why it *is not desirable* to represent information in as few bits as possible.

B.

A set of N records is to be stored in a *main memory* of size M using either a *Fredkin TRIE* or a *binary search tree*. Each record is L bytes long, and its key consists of K characters in base B. For each of these five parameters, indicate the effect of an increase in that parameter assuming that the other four remain fixed by stating whether an increase in that one parameter would

(1) favour the use of a TRIE,
(2) be neutral to the choice, or
(3) favour the use of the binary search tree (BST).

Circle one in each row:

(i).	Increasing N:	favours TRIE	neutral	favours BST
(ii).	Increasing M:	favours TRIE	neutral	favours BST
(iii).	Increasing L:	favours TRIE	neutral	favours BST
(iv).	Increasing K:	favours TRIE	neutral	favours BST
(v).	Increasing B:	favours TRIE	neutral	favours BST

5. [8 marks]

A. Each of the situations described below motivates one particular multiple-key file structure more than the others. For each situation circle that one from the set:

(i). Many of the queries are range queries and it is required that each record access require at most two disk transfers:

transposed files	k-d trees	k-d-B trees	multidim. TRIEs	grid files

(ii). The database can be held entirely in main memory:

transposed files	k-d trees	k-d-B trees	multidim. TRIEs	grid files

(iii). Most queries treat a few attributes of a large number of large records:

transposed files	k-d trees	k-d-B trees	multidim. TRIEs	grid files

(iv). Many of the queries are of the range type, and several pairs of attributes are highly correlated:

transposed files	k-d trees	k-d-B trees	multidim. TRIEs	grid files

(v). All queries are of the exact match or partial match type:

transposed files	k-d trees	k-d-B trees	multidim. TRIEs	grid files

B. Suppose that in a certain context, a multidimensional file structure is used to support queries involving various combinations of keys, A, B, and C, where A is specified in 75% of the queries, B is specified in 50% of the queries, and C is specified in 25% of the queries (i.e., on average, two of the three attributes are specified in each query). In a few words, indicate in general how each of the file structures below can be "tuned" to take advantage of the known asymmetry in frequencies of key use.

(i). k-d Tree:

(ii). Multikey Hashing:

(iii). Grid Files:

6. [9 marks]

A. In what way to *dynamic hashing methods* (such as Extendible Hashing, Linear Hashing, and Spiral Hashing) improve on the performance of standard Bucket Hashing?

B. Neither $B+$-trees nor Bucket Hashing take into account the relative *frequencies* of access to different records in a file. If it is known that the access frequencies actually follow the "90-10 rule" (i.e., 90% of all accesses go to 10% of the records in the file), suggest an approach by which average access time could be reduced relative to those of $B+$-trees or Bucket Hashing alone.

C. *Chaining* is commonly used to handle bucket overflows in standard Bucket Hashing. When an insertion is attempted to a full bucket, the new record is placed in a new bucket that is linked from the previous one by a pointer. In some cases, this means that multiple block accesses are required to locate a requested record. Identify and describe very briefly a mechanism that can be used in this situation to assure that at most one block transfer is required to access any record in the file.

7. [7 marks]

A. Real numbers can be represented (or at least approximated) in computers in either *single-precision* or *double-precision* format. In a few words, either defend or refute the following claim from each of the two viewpoints indicated below:

CLAIM: "Twice as much information is conveyed about a real number when it is represented in double-precision rather than single-precision format."

(i). From the point of view of standard *Discrete Information Theory*:

(ii). From the point of view of *Discrimination Theory*:

B. Consider a sequence of real numbers selected independently and uniformly from the interval (0,1). Let the *risk* of encoding a real number, x, by representation, x', be given by $risk\ (x, x') = | x - x' |$. If it is required that the *average risk* of encoding each number is to be kept to at most .001, what is the minimum number of bits (to the nearest bit) required to encode each number of the sequence? Justify your answer briefly.

8. [10 marks]

A. Consider R^*-trees, SS-trees, and SR-trees.

(i). What is the fundamental difference between R^*-trees and SS-trees?

(ii). In what way do SR-trees represent a combination of R^*-trees and SS-trees?

(iii). As dimensionality of the vector space increases, how would you expect the performance of SR-trees in supporting range queries or nearest-neighbour queries to change relative to R^*-trees and SS-trees? Circle one.

IMPROVE **SAME** **WORSE**

B. One technique devised recently and used in the VA-file, IQ-trees, and A-trees to address the "Curse of Dimensionality" is the use of *vector quantization*. Assume that the values of each of k attributes in a record are normalized into the range [0,1], and then represented as 32-bit binary fractions. Explain briefly the simplest approach to obtaining a b-bit quantization of the record.

9. [12 marks]
A relation with k *dimension* attributes and one or more *measure* attributes can be represented in a k-dimensional datacube.

A. What are the advantages of storing the full datacube rather than just the data from the base relation?

B. What are the disadvantages of storing the full datacube?

C. If queries can be answered *approximately* rather than *exactly*, describe how one significant disadvantage can be avoided.

10. [9 marks]
In Kanerva's SDM, explain concisely each of the following characteristics:

A. Explain why repeatedly storing a concept makes that concept more likely to be recalled from the memory.

B. Explain why the success of doing a *best match* retrieval decreases as more and more concepts are recorded in the memory.

C. Explain why some best match retrievals usually require iterated reads to succeed.