CSC 2508S
Fall, 2002

October 2, 2002
DUE: October 21, 2002

Information: Quantification, Representation and Manipulation

Assignment #2
<**Sample Answers**>

1. Arithmetic Coding [20 marks]

Either by hand or using a program, carry out the tasks below involving *Arithmetic Coding*. (Show your intermediate work by writing down, or printing out from your program, the intermediate values obtained at each stage of the computations.)

Assume a source alphabet of four symbols (A, B, C, and D) with respective *independent* probabilities of occurrence (.4, .3, .2, and .1).

(a). Encode the string "DACABABACAD" producing a real number in [0,1).

letter	low	high	range
	0.0	1.0	1.0
D	.9	1.0	.1
A	.94	.90	.04
C	.928	.936	.008
A	.9280	.9312	.0032
B	.92928	.93024	.00096
A	.929280	.929664	.000384
B	.9294336	.9295488	.0001152
A	.92943360	.92947968	.00004608
C	.929465856	.929475072	.000009216
A	.9294658560	.9294695424	.0000036864
D	.92946917376	.92946954240	.00000036864

(b). Decode the value .517353856 to produce a string of length nine composed of the four symbols.

letter	low	high	range
	0.0	1.0	1.0
B	.4	.7	.3
A	.40	.52	.12
D	.508	.520	.012
C	.5164	.5188	.0024
A	.51640	.51736	.00096
D	.517264	.517360	.000096
D	.5173504	.5173600	.0000096
A	.51735040	.51735424	.00000384
D	.517353856	.517354240	.000000384

2. Lempel-Ziv-Welch Coding [40 marks]

Write a program to do encoding and decoding according to the simplified LZW approach. Your dictionary implementation may be based on any of (i) text strings (easier, but less general), (ii) linked lists representing strings (simple and efficient), or (iii) a TRIE representation of the set of strings (complex, but interesting!). Your program should read input from a file that has the following format:

line 1: D % the number of symbols in the alphabet

line 2: $S_1, S_2, S_3, \ldots, S_D$ % the symbols

line 3: ENCODE or DECODE % the operation to be carried out

The remainder of the file will be, for ENCODE, a string of characters over the alphabet **in which all spaces should be ignored**, or, for DECODE, a sequence of integers (separated by spaces) that represents an encoded text string. Conversely, the output for ENCODE should be a sequence of integers, and the output for DECODE should be a sequence of characters from the alphabet **(either with or without spaces between adjacent pairs of characters).**

Apply your program to each of the (four to six) files available for this purpose on the course web page.

For each of the six text files that you process, keep track of and display:

(i). the number of dictionary entries at the end, and

(ii) the ratio of the number of characters in the unencoded message to the number of integers in the encoded message.

```
Q2-1
INPUT:
   B O O H O O H O O C H O O C H O O B O O H O O H O O
OUTPUT:
    1    4    4    3    6    8    4    2   10   12    6    5    7    9    6
ENTRIES: 18     LETTERS:   26     INTEGERS:    15     RATIO:   1.733

Q2-2
INPUT:
    1    4    4    3    6    8    4    2   10   12    6    5    7    9    6
OUTPUT:
   B O O H O O H O O C H O O C H O O B O O H O O H O O
ENTRIES: 18     LETTERS:   26     INTEGERS:    15     RATIO:   1.733

Q2-3
INPUT:
    B E T T Y  B O T T E R  H A D  S O M E  B U T T E R
    B U T  S H E  S A I D  T H I S  B U T T E R S  B I T T E R
    I F  I  B A K E  T H I S  B I T T E R  B U T T E R
    I T  W O U L D  M A K E  M Y  B A T T E R  B I T T E R
    B U T  A  B I T  O F  B E T T E R  B U T T E R
    T H A T  W O U L D  M A K E  M Y  B A T T E R  B E T T E R
    S O  S H E  B O U G H T  A  B I T  O F  B U T T E R
```

```
    B E T T E R  T H A N  H E R  B I T T E R  B U T T E R
    A N D  S H E  B A K E D  I T  I N  H E R  B A T T E R
    A N D  T H E  B A T T E R  W A S  N O T  B I T T E R
    S O  T W A S  B E T T E R  B E T T Y  B O T T E R
    B O U G H T  A  B I T  O F  B E T T E R  B U T T E R
OUTPUT:
    2    4   16   16   19    2   13   22    4   14    7    1    3   15   13   11
    4    2   17   27   14   37   16   15    7    4   15    1    8    3   16    7
    8   15   41   16   28   53    8   39    8    5    8    2    1    9   21   51
   57   39   54   28   58   18   13   17   10    3   11   64    4   11   24    1
   69   58   55   40   38    1    2   58   13    5   20   69   38   86   50   83
   73   75   77   79   81   63   69   21   86   33   43   36   74    6    7   16
   89   91   93   96   28   94   97   30   12   44   40   85  120  119   14    1
   12   32   44   63   65    3   58    8  124  120   83   86  131   49  134  142
   28   18    1   15   12   26   90   39   33   16  149   53  107  120  107   24
   26   86   25   17  113  115  154   92  121  128   39
ENTRIES: 173     LETTERS:   290     INTEGERS:    155     RATIO:   1.8710

Q2-4
INPUT:
  12    5    3   12    3    7    7   12   19    1   16   20   22    2   15   13
  17   24   16   10   11    3   31   18   17   21   23   32   27   12    1   36
  12   14   36    7   15   33   40   22   12   10    6    4   26   19   41   26
  41   10    9   38   53   35    3    6    8   48   47   43   62    3   25    5
  69   63   22
OUTPUT:
 S H E S E L L S S E A S H E L L S B Y T H E S E A S H O R E
 T H E S H E L L S S H E S E L L S A R E S U R E L Y S E A S H E L L S
 S O I F S H E S E L L S S H E L L S O N T H E S E A S H O R E
 I M S U R E S H E S E L L S S E A S H O R E S H E L L S
ENTRIES:  81     LETTERS:   124     INTEGERS:    67     RATIO:   1.8507

Q2-5
INPUT:
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
   H O W N O W B R O W N C O W H O W N O W B R O W N C O W
OUTPUT:
```

```
 3 5 7 4 9 1 6 9 4 2 9 8 10 12 14 10 17 7 19 11 7 13 15 24 26 21 30 18 15 33
 23 35 20 28 22 16 39 27 29 38 25 36 41 34 48 40 46 43 52 45 42 31 49 54 59
 53 58 44 37 55 32 50 47 68 61 65 69 67 60 64 56 66 62 57 51 71 77 83 82 76
 86 63 88 81 47
ENTRIES:  91    LETTERS:   336     INTEGERS:    85     RATIO:   3.9529
```

Q2-6

INPUT:

```
  13    1   18   25    8    1    4    1   12    9   20   20   12    5   12    1
  13    2   35   37   39   41   43   45   38   40   42    2   27   29   31   33
  50   47   53   36   19    6   39    5    3    5   23    1   19   23    8   36
   5   70   19   14   15   23    5   22   29   72    5   18    5   20   31   20
  55   25   80   14   90   28   92    5   94   91   93   20   81   83    8   85
  87   89  100   98   37  105   48    2   69   19   19   21   86   20   15    7
  15   91   57   34   36   51  113   59   52   49  127   60   43  124   32  126
  46  131    2   62   64    5   66   68   70   72   74   76   78   80   82   92
 105  119  108   96  101  109   99  158  110  103  154  106   88    1   95   83
 110   88  140  115  117  119  121
```

OUTPUT:

```
 M A R Y H A D A L I T T L E L A M B
 L I T T L E L A M B L I T T L E L A M B
 M A R Y H A D A L I T T L E L A M B
 I T S F L E E C E W A S W H I T E A S S N O W
 E V R Y W H E R E T H A T M A R Y W E N T
 M A R Y W E N T M A R Y W E N T
 E V R Y W H E R E T H A T M A R Y W E N T
 T H E L A M B W A S S U R E T O G O
 M A R Y H A D A L I T T L E L A M B
 L I T T L E L A M B L I T T L E L A M B
 M A R Y H A D A L I T T L E L A M B
 I T S F L E E C E W A S W H I T E A S S N O W
 E V R Y W H E R E T H A T M A R Y W E N T
 M A R Y W E N T M A R Y W E N T
 E V R Y W H E R E T H A T M A R Y W E N T
 T H E L A M B W A S S U R E T O G O
ENTRIES:  176    LETTERS:   309     INTEGERS:    151     RATIO:   2.0464
```

3. Calculating Entropy Rates Based on Context [40 marks]

Write a program to calculate an estimate of the *Entropy Rate* of an arbitrary text file by creating Markov models of **orders 0, 1, 2, 3 and 4.**

If you find it difficult to base your program on a single scan of the file, you may use a first pass to determine the symbol alphabet, and a second pass to gather the relevant probabilities.

Once the transition probabilities for various amounts of context are found, your program should use them to produce estimates of the entropy rate of the document.

(a). Apply your program to the four test files for this question available on the course web page, and give an estimate of the entropy rate for each one.

file	Q3-1	Q3-2	Q3-3	Q3-4
observations	4997	9995	19995	49995
context length				
0	2.0493	3.1891	2.9998	3.3218
1	2.0477	1.5537	2.9981	3.3206
2	2.0379	1.5476	1.2855	3.3096
3	1.9738	1.5215	1.2750	1.5280
4	1.7410	1.4222	1.2459	1.3862

(b). Next, apply your program to at least three other text files (of diverse natures) found in your directories (e.g., mail, latex, html, data or system files), and report your estimates of their entropy rates.

file type	troff	postscript	uuencoded
observations	3274	12760	18140
context length			
0	4.9378	5.3544	5.5431
1	3.4447	3.5801	3.7618
2	1.6329	2.1038	2.1007
3	.6394	.9546	1.0353
4	.3262	.4919	.4537

(c). Briefly discuss your findings from the experiments carried out in part (b).

These files were far too small to truly reflect entropy rates, particularly for the higher orders of dependence. With an alphabet of 256 characters (which these files were treated as having), 18,000 is a very small number of observations to make over all the possible contexts that can be encountered. When a context is encountered only a few times (say, four or less), it cannot possibly have a very high entropy. Note that, as the order of the context increases, the perceived entropy rate drops of radically.