p343
MR

1 Data structures and algorithms

2004

(a) Describe how the Lempel Ziv text compression algorithm works, illustrating your answer by deriving the sequence of numbers and corresponding bit patterns it would generate when applied to a string starting with the follow 24 characters:

```
ABCDABCDABCDABCDABCDABCD ...
```

You may assume that the initial table is of size 256 (containing bytes 0 to 255) and that the codes for 'A', 'B', 'C' and 'D' are 65, 66, 67 and 68, rerspectively. [12 marks]

(b) Estimate how many bits the algorithm would use to encode a string consisting of 1000 repetitions of the character 'A' [8 marks]

ANSWER NOTES:

(a) Description of Lempel-Ziv is bookwork.

The initial table holds codes 0 to 255 corresponding to the 256 different strings of length one. The encoding goes as follows:

string	code	bits	new table entry	represented by	
A	65	01000001	256: AB	65(A)	B
B	66	001000010	257: BC	66(B)	C
C	67	001000011	258: CD	67(C)	D
D	68	001000100	259: DA	68(D)	A
AB	256	100000000	260: ABC	256(AB)	C
CD	258	100000010	261: CDA	258(CD)	A
ABC	260	100000100	262: ABCD	260(ABC)	D
DA	259	100000011	263: DAB	259(DA)	B
BC	257	100000001	264: BCD	257(BC)	D
DAB	263	100000111	265: DABC	263(DAB)	C
CDA	261	100000101	266: CDAB	261(CDA)	B
BCD	264	100001000	267: BCDA	264(BCD)	A
...					

(b) The encoding of AAAAAAAAAAAAAAAAAAAA... is as follows

string	code	bits	new table entry
A	65	01000001	256: AA
AA	256	100000000	257: AAA
AAA	257	100000001	258: AAAA
AAAA	258	100000010	259: AAAAA
...			

So length of string encoded by n codes is 1+2+3+4+...+n = n(n+1)/2

45*46/2 = 1035
so a sequence of 45 codes can represent a string of 1035 As.
If the last code were changed the sequence of 45 codes could represent a string of exactly 1000 As. The first code is 8 bits the remaing codes are all 9 bits so the total length is

8 + 44*9 = 468 bit

Any answer between 420 and 530 would gain full marks provided the explanation was ok.