IR Assignment III

1. Consider a query and a document collection consisting of three documents. Rank the documents using vector space model. Assume tf-idf weighing scheme.

Query: "gold silver truck"

Document Collection:

d1: "Shipment of gold arrived in a truck."

d2: "Shipment of gold damaged in a fire."

d3: "Delivery of silver arrived in a silver truck."

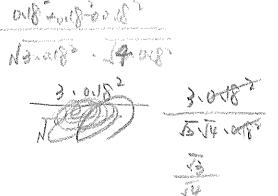
term freq								tf-idf			
	term	Q	D1	D2	D3	df	idf ≠ log (N/df)	Q	D1	D2	D3
1	a	0	1	1	1	3	0	/6\	0	0	0
	arrived	0	1.	0	1	2	0.18	0	(0.18)	0	0.18
	damaged	0	0.	1	0	1	0.48	0	0	0.48	0
	delivery	0	0	0	1	1	0.48	0	0	0	0.48
	fire	0	0	1	0	1	0.48	<u>_</u>	0	0.48	0
	(gold)	1	1	1	0	2	0.18	(0.18)	(0.18)	0.18	0
	in	0	1	1	1	3	0	0	0	0	0 :
***	of	0	1	1	1	3	0	0	0	0	0
	shipment	1	1	1	0	2	(18)	0\0	0.18	0.18	0
	silver	0	0	0	. 1	1	0.48	0.48	0	0	0.96
	truck		1	0	1	2	(0.18)	0,0	0.18	0	0.18

$$S(Q, D1) = (Q \cdot D1) / (|Q| * |D1|) = 0.33$$

$$S(Q, D2) = 0.08$$

$$S(Q, D3) = 0.83$$

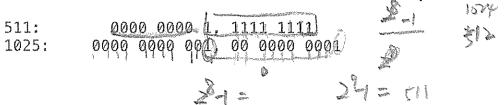
Ranking: D3, D1, D2



2. γ Codes are relatively inefficient for large numbers as they encode the length of the offset in inefficient unary code. δ codes δ codes differ from γ codes in that they encode the first part of the code (length) in γ code instead of unary code. The encoding of offset is the same. For example, the δ code of 7 is 10,0,11 (again, we add commas for readability). 10,0 is the γ code for length (2 in this case) and the encoding of offset (11) is unchanged. (i) Compute the δ codes for the numbers 511 and 1025.

- 4. We have defined unary codes as being "10": sequences of 1s terminated by a 0. Interchanging the roles of 0s and 1s yields an equivalent "01" unary code. When this 01 unary code is used, the construction of a y code can be stated as follows:
 - 1. Write G down in binary using $(b) = \lfloor \log 2i \rfloor + 1$ bits.
 - 2. Prepend (b-1) 0s.

Encode the numbers 511 and 1025 in this alternative y code.



- 5. Consider the postings <u>list</u> <4, 10, 11, 12, 15, 62, 63, 265, 268, 270, 400> with a corresponding list of gaps <4, 6, 1, 1, 3, 47, 1, 202, 3, 2, 130>. Using variable byte encoding:
 - i. What is the largest gap you can encode in 1 byte?

$$127(2^7 - 1)$$
 (1 byte)

ii. What is the largest gap you can encode in 2 bytes?

In 2 bytes,
$$2^{14} - 1 = 16383$$

iii. How many bytes will the above postings list require under this encoding?

13