Εργασία 2 – Ανάκτηση Πληροφορίας και Μηχανές Αναζήτησης

1. Για κάποιον όρο του λεξικού έχουμε την εξής **postings list <44, 59, 80, 85, 99, 300, 301>**. Συμπιέστε τη λίστα με **γ-code**, **δ-code** και **variable byte code** (με 8 bit blocks) και υπολογίστε τη συμπίεση που πετυχαίνει η κάθε μια προσέγγιση. Η ασυμπίεστη εκδοχή της λίστας απαιτεί 7*4=28 bytes ή 28*8=224 bits. Υπενθυμίζω ότι ο δ-code είναι ίδιος με τον γ-code με τη μόνη διαφορά ότι κωδικοποιεί το length με γ-code.

2. Ακολουθούν ερωτήσεις σχετικά με την κωδικοποίηση variable byte.
   (α) Ποιος είναι ο μεγαλύτερος αριθμός που μπορεί να κωδικοποιηθεί με ένα byte;
   (β) Ποιος είναι ο μεγαλύτερος αριθμός που μπορεί να κωδικοποιηθεί με δυο bytes;
   (γ) Δίνεται η **postings list <4,10,11,12,15,62,63,265,268,270,400>** και η αντίστοιχη **gaps list <4,6,1,1,3,47,1,202,3,2,130>**. Με βάση τις απαντήσεις σας στα (α) και (β) πόσα bytes απαιτούνται συνολικά για την κωδικοποίηση της παραπάνω gaps list;

3. Σας δίνεται ο γ-code **1110001110101011111101101111011**. Αποκωδικοποιήσετε τον ώστε να πάρετε την gaps list και μετά δώστε την αρχική postings list.

4. Υλοποιήστε σε όποια γλώσσα προγραμματισμού θέλετε την κωδικοποίηση variable byte (με 8 bit blocks). Το πρόγραμμα θα παίρνει ως όρισμα έναν ακέραιο (>0) και θα επιστρέφει τον αντίστοιχο variable byte code.

5. Έστω ότι στη συλλογή Reuters (N=806791) έχουμε τα παρακάτω στοιχεία για τέσσερις όρους:

Πίνακας 1: df_t και idf_t των όρων

term	df_t	idf_t
car	18165	1,65
auto	6723	2,08
insurance	19241	1,62
best	25235	1,5

Πίνακας 2: tf των όρων για 3 έγγραφα

term	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

Υπολογίστε τα tf-idf βάρη των όρων για κάθε ένα από τα τρία έγγραφα και δώστε τα κανονικοποιημένα διανύσματα των εγγράφων (δείξτε αναλυτικά τους υπολογισμούς που κάνατε).

6. Διατάξετε τα τρία έγγραφα του Πίνακα 2 του προβλήματος 5 ως προς την ομοιότητά τους με το ερώτημα “car insurance” χρησιμοποιώντας ως βάρος των όρων στο ερώτημα
   (α) το 1 αν υπάρχει ο όρος και 0 αλλιώς
   (β) το κανονικοποιημένο idf όλων των όρων

Δείξτε αναλυτικά τους υπολογισμούς που κάνατε.

Καταθέστε ένα zip αρχείο με τις απαντήσεις σας και τον πηγαίο κώδικα του ερωτήματος 4.