ΑΣΚΗΣΗ

Ημερομηνία Παράδοσης: τέλος εξεταστικής περιόδου

Η άσκηση είναι **ατομική, υποχρεωτική και απαλλακτική** για την εξέταση του μαθήματος (αφορά και φοιτητές παλαιότερων ετών). Από την εκπόνηση της άσκησης εξαιρούνται οι φοιτητές που θα εξεταστούν προφορικά.

Η αναφορά που θα συντάξετε θα παραδοθεί σε ηλεκτρονική μορφή. **Προσοχή,** στην αναφορά που θα παραδώσετε να αναγράφεται το ονομά σας, το επίθετό σας, το ΑΜ σας και το έτος σπουδών σας.

Κάθε απάντηση θα είναι **αιτιολογημένη**. Η αιτιολόγηση μπορεί να γίνει είτε περιγράφοντας τη μεθοδολογία που ακολουθήσατε για την ολοκλήρωση κάθε βήματος είτε υπό τη μορφή παραδειγμάτων. Σε περίπτωση αντιγραφής μηδενίζονται και οι δύο ασκήσεις.

Ασκήσεις που θα παραδοθούν εκπρόθεσμα δεν γίνονται δεκτές!

ΕΚΦΩΝΗΣΗ

Έστω μια συλλογή κειμένων C που περιέχει 5 κείμενα:

$$C \{d_1, d_2, d_3, d_4, d_5\}$$

Όπου για κάθε κείμενο d ισχύουν τα ακόλουθα:

$d_1 \{w_a, w_b, w_c, w_d, w_e, w_m\}$

$d2 \{w_a, w_f, w_c, w_j, w_k, w_m\}$

$d_3 \{w_e, w_k, w_n, w_c, w_d, w_p, w_t\}$

$d_4 \{w_b, w_f, w_k, w_u, w_c, w_m, w_d, , w_t\}$

$d_5 \{w_a, w_c, w_f, w_k, w_y, w_z, w_t\}$

όπου όλοι οι όροι εκτός από τους w_b, w_c είναι ΟΥΣΙΑΣΤΙΚΑ.

Επιπλέον, έστω ότι:

$D_1 \{fw_a=8, fw_b=8, fw_c=10, fw_d=14, fw_e=5, fw_m=15\}$, όπου *f = frequency*

$D_2 \{fw_a=13, fw_f=19, fw_c=10, fw_j=18, fw_k=6, fw_m=14\}$, όπου *f = frequency*

D_3 {fw_e =10, fw_k=12, fw_n=18, fw_c=13, fw_d=27, fw_p=10, fw_t=10}, όπου *f* = *frequency*

D_4 {fw_b =15, fw_f=18, fw_k=17, fw_u=13, fw_c=6, fw_m=1, fw_d=8, fw_t=10}, όπου *f* = *frequency*

D_5 {fw_a =15, fw_c=18, fw_f=20, fw_k=13, fw_y=16, fw_z=18, fw_t=20}, όπου *f* = *frequency*

Για τα παραπάνω δεδομένα θέλουμε να δημιουργήσουμε ένα ευρετήριο όρων προκειμένου να μπορούμε να ανακτήσουμε πληροφορία από τα δεδομένα της συλλογής.

(i) Παρουσιάστε το ευρετήριο όπου θα αναγράφονται οι όροι που θα περιέχονται σε αυτό, οι δείκτες προς τα κείμενα που τους περιέχουν και για κάθε κείμενο ο βαθμός ευρετηρίασής τους (w score).

(ii) Ποιος είναι ο πιο σημαντικός όρος ευρετηρίασης για το d_5; Να αιτιολογήσετε την απάντησή σας και να δώσετε την αριθμητική τιμή που δείχνει το βαθμό σπουδαιότητας του όρου που επιλέξατε για το d_5

(iii) Ποιος ο συνολικός βαθμός σπουδαιότητας για το ευρετήριο του όρου που επιλέξατε στον προηγούμενο ερώτημα; (υπολογίστε τη σπουδαιότητα του όρου με τη μετρική TF*IDF για το d_5).

(iv) Έστω το ερώτημα *q* που περιέχει τους όρους *q* = w_a, w_f, w_z. Ποια κείμενα θα επιστρέφονταν ως απάντηση για το ερώτημα και με ποια σειρά; Αιτιολογήστε την απάντησή σας.

Καλή επιτυχία!