

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΞΕΤΑΣΤΙΚΗ ΠΕΡΙΟΔΟΣ ΣΕΠΤΕΜΒΡΙΟΥ
ΜΑΘΗΜΑ: ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ

1. Θεωρείστε μία συλλογή 4 κειμένων ($d_1...d_4$) και έστω το λεξικό της συλλογής αποτελείται από 4 όρους ($t_1...t_4$). Στον πίνακα που ακολουθεί, φαίνεται η συχνότητα εμφάνισης των όρων στα κείμενα της συλλογής.

	t_1	t_2	t_3	t_4
d_1	4	1	12	10
d_2	8	2	0	3
d_3	5	0	6	2
d_4	0	0	1	1

Στο σύστημα υποβάλλεται η ερώτηση « $t_1 \ t_3$ ». Καθορίστε τη βαθμολογία των κειμένων ως προς τη σχετικότητα τους με το συγκεκριμένο ερώτημα σύμφωνα με το διανυσματικό μοντέλο. **(3 μονάδες)**

2. Σε ένα ανεστραμμένο αρχείο, στη λίστα εμφάνισης κάθε όρου αποθηκεύονται και οι θέσεις εμφάνισης του όρου στα κείμενα. Έτσι για τους δύο όρους "information" και "retrieval" οι λίστες εμφάνισης είναι οι εξής:

information: **3:** 12, 50, 80, 100; **5:** 3, 30, 50, 80; **10:** 70, 100, 130, 155, 180; **15:** 25, 40, 110

retrieval: **2:** 3, 20, 50, 100, 150; **5:** 15, 62, 75, 100; **7:** 15, 35, 44, 80; **10:** 82, 142, 170, 192

Με έντονα γράμματα είναι οι κωδικοί των κειμένων στα οποία εμφανίζονται οι όροι. Για κάθε τέτοιο κείμενο ακολουθεί η λίστα των θέσεων του όρου στο συγκεκριμένο κείμενο. Π.χ ο όρος "information" εμφανίζεται στο κείμενο 3 και συγκεκριμένα στον 12ο, 50ο, 80ο και 100ο χαρακτήρα του κειμένου (Positional offset). Στο συγκεκριμένο σύστημα ανάκτησης πληροφοριών υποβάλλεται η ερώτηση φράσης (phrase query) "Information retrieval". Περιγράψτε με λεπτομέρεια μία γρήγορη μέθοδο επεξεργασίας των δύο λιστών εμφάνισης για την απάντηση της ερώτησης. **(2 μονάδες)**

3. α) Μετά την εκτέλεση μίας ερώτησης σε ένα Σύστημα Ανάκτησης Πληροφοριών (ΣΑΠ), επιστρέφονται 20 κείμενα ταξινομημένα ως προς τη σχετικότητάς τους με την ερώτηση. Σύμφωνα με τη γνώμη ενός ειδικού τα πραγματικά σχετικά κείμενα είναι το 2ο, 6ο, 8ο, 9ο, 10ο, 12ο, 15ο, 16ο, 17ο, 18ο και 19ο κείμενο της κατάταξης. Άλλα σχετικά κείμενα στη συλλογή κειμένων του συστήματος ανάκτησης πληροφοριών δεν υπάρχουν. Σχεδιάστε το διάγραμμα ακρίβειας-ανάκλησης για την ερώτηση αυτή. **(1 μονάδα)**

β) Στην τεχνική Rocchio για ανάδραση σχετικότητας, η ενημέρωση του διανύσματος ερώτησης γίνεται σύμφωνα με τον ακόλουθο τύπο:

$$q_m = \alpha q_o + \beta \frac{1}{|D_r|} \sum_{d_i \in D_r} d_i - \gamma \frac{1}{|D_{nr}|} \sum_{d_i \in D_{nr}} d_i$$

Μετά την εκτέλεση της αρχικής ερώτησης q_0 επιστρέφονται τα εξής κείμενα ταξινομημένα ως προς τη σχετικότητα τους με την συγκεκριμένη ερώτηση: $d_1, d_2, \dots, d_{10}$. Ο χρήστης κρίνει ως σχετικά τα κείμενα d_2, d_3, d_5, d_6, d_9 και τα υπόλοιπα ως μη σχετικά.

Για κάθε μία από τις επόμενες περιπτώσεις, δώστε μία εκτίμηση για τις τιμές των συντελεστών α , β και γ .

i) Μετά την εκτέλεση της ερώτησης q_m , τα δέκα πρώτα αποτελέσματα είναι:

$d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}$

ii) Μετά την εκτέλεση της ερώτησης q_m , τα δέκα πρώτα αποτελέσματα είναι:

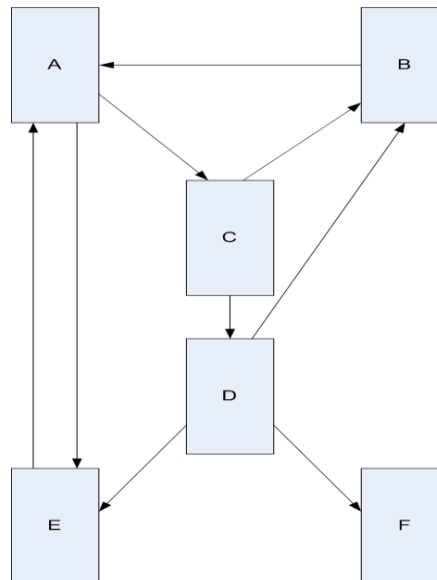
$d_2, d_3, d_5, d_6, d_9, d_1, d_4, d_7, d_8, d_{10}$

iii) Μετά την εκτέλεση της ερώτησης q_m , τα δέκα πρώτα αποτελέσματα είναι:

$d_2, d_3, d_5, d_6, d_9, d_{20}, d_{25}, d_{27}, d_{29}, d_{35}$

(1,5 μονάδα)

4. α) Δώστε τις εξισώσεις της τεχνικής PageRank για τον ιστό που ακολουθεί. **(1 μονάδα)**



β) Η τεχνική του PageRank βασίζεται στο μοντέλο της τυχαίας περιήγησης, όπου σε κάθε σελίδα, ο χρήστης είτε διαλέγει στη τύχη μία οποιαδήποτε σελίδα του ιστού ή μεταξύ των συνδέσμων που έχει η τρέχουσα σελίδα, διαλέγει με ίση πιθανότητα να ακολουθήσει έναν από αυτούς τους συνδέσμους. Μία παραλλαγή της βασικής τεχνικής θα ήταν η πιθανότητα με την οποία ο χρήστης ακολουθεί ένα συγκεκριμένο σύνδεσμο να μην είναι ίδια για όλους τους συνδέσμους της τρέχουσας σελίδας αλλά ανάλογη του πόσο δημοφιλής είναι η σελίδα στην οποία δείχνει ο σύνδεσμος. Θεωρείστε ως μέτρο "δημοφιλίας" μίας σελίδας το πλήθος των συνδέσμων που δείχνουν προς αυτή. Δώστε τις εξισώσεις της τροποποιημένης τεχνικής PageRank για τον παραπάνω ιστό. **(1,5 μονάδα)**