

Εργασία 1 – Ανάκτηση Πληροφορίας και Μηχανές Αναζήτησης

1. Δίνεται η συλλογή εγγράφων:

Doc1 οι πωλήσεις νεόκτιστων διαμερισμάτων ξεπέρασαν τις προβλέψεις

Doc2 οι πωλήσεις διαμερισμάτων παρουσιάζουν άνοδο τον Ιούλιο

Doc3 αύξηση στις πωλήσεις διαμερισμάτων τον Ιούλιο

Doc4 τον Ιούλιο υπάρχει αύξηση στις πωλήσεις νεόκτιστων διαμερισμάτων

(α) Δώστε το ανεστραμμένο αρχείο που θα κατασκευαστεί.

(β) Ποια από τα έγγραφα επιστρέφονται από τα παρακάτω ερωτήματα:

(i) πωλήσεις AND αύξηση

(ii) Ιούλιο AND NOT (άνοδο OR αύξηση)

2. Προτείνετε τη σειρά επεξεργασίας για το παρακάτω ερώτημα:

(τραπέζι OR καρέκλα) AND (λάμπα OR κουρτίνα) AND (υπολογιστής OR μολύβι)

αν ισχύουν τα παρακάτω:

Term	Postings size
------	---------------

μολύβι	31000
--------	-------

υπολογιστής	18000
-------------	-------

λάμπα	5000
-------	------

κουρτίνα	50000
----------	-------

τραπέζι	25000
---------	-------

καρέκλα	32000
---------	-------

Αφού απαντήσετε, ξανασκεφτείτε αν η συντηρητική προσέγγιση ότι το μήκος της ένωσης δυο postings lists είναι $O(x+y)$ θα είναι κοντά στην πραγματικότητα στο συγκεκριμένο πρόβλημα και προτείνετε εναλλακτική σειρά επεξεργασίας.

3. Θεωρήστε το query **X AND Y AND Z**. Γνωρίζουμε ότι τα μεγέθη των postings lists για τα X, Y και Z είναι 100, 105 και 110 αντίστοιχα.

(α) Με ποια σειρά θα γίνει η επεξεργασία του ερωτήματος και ποιο θα είναι το κόστος;

(β) Αν γνωρίζουμε ότι η τομή των X και Y έχει μήκος 100 και η τομή των X και Z έχει μήκος 0, ποια θα πρέπει να είναι η σειρά επεξεργασίας του ερωτήματος και ποιο θα είναι το κόστος;

4. Παρακάτω δίνονται postings lists με positional πληροφορία (μορφή: term: doc1: [position1, position2, . . .]; doc2: [position1, position2, . . .]; κλπ.) κάποιων όρων ενός λεξικού.

υπάρχει:	2: [36,174,252,651];	4: [12,22,102,432];	7: [17];
----------	----------------------	---------------------	----------

μην:	2: [1,17,74,222];	4: [8,78,108,458];	7: [3,13,23,193];
------	-------------------	--------------------	-------------------

κίνδυνος:	2: [87,704,722,901];	4: [13,43,113,433];	7: [18,328,528];
-----------	----------------------	---------------------	------------------

πολύ:	2: [3,37,76,444,851];	4: [10,20,110,470,500];	7: [5,15,25,195];
-------	-----------------------	-------------------------	-------------------

τρέχεις:	2: [2,66,194,321,702];	4: [9,69,149,429,569];	7: [4,14,404];
----------	------------------------	------------------------	----------------

να:	2: [47,86,234,999];	4: [14,24,774,944];	7: [199,319,599,709];
-----	---------------------	---------------------	-----------------------

σκοτωθείς:	2: [57,94,333];	4: [15,35,155];	7: [20,320];
------------	-----------------	-----------------	--------------

Υπάρχουν έγγραφα που ταιριάζουν με τα ακόλουθα ερωτήματα φράσεων;

(α) “μην τρέχεις πολύ”

(β) “μην τρέχεις πολύ” AND “υπάρχει κίνδυνος να σκοτωθείς”

5. Σε ένα ερώτημα έχουμε δυο όρους με τις παρακάτω postings lists:

term1 → [4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180]

term2 → [47]

Υπολογίστε τον αριθμό των συγκρίσεων που απαιτούνται για την τομή των δυο λιστών ανάλογα με την ακολουθούμενη στρατηγική:

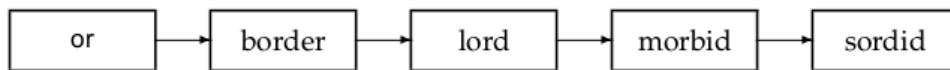
(α) Με χρήση των postings lists ως έχουν

(β) Χρησιμοποιώντας skip pointers με μήκος \sqrt{p} , όπου p είναι το μήκος της postings list.

6. Σε ένα permuterm index κάθε όρος δείχνει στον αρχικό όρο του λεξιλογίου από τον οποίο προέκυψε. Πόσοι τέτοιοι αρχικοί όροι μπορεί να υπάρχουν στην postings list ενός permuterm όρου;

7. Υπολογίστε την edit distance ανάμεσα στους όρους **paris** και **alice**. Δώστε τον πλήρη 5x5 πίνακα με τις αποστάσεις ανάμεσα σε όλα τα προθήματα των όρων που υπολογίζει ο αλγόριθμος και εξηγήστε ποιες ακριβώς είναι οι πράξεις μετάβασης από τον έναν όρο στον άλλο.

8. Υπολογίστε τους Jaccard coefficients ανάμεσα στο ερώτημα **bord** και τους όρους που περιέχουν το bigram **or**.



9. Βρείτε δυο τελείως διαφορετικά κύρια ονόματα τα οποία έχουν τον ίδιο soundex code.

10. Βρείτε δυο κύρια ονόματα που φωνητικά είναι όμοια αλλά έχουν διαφορετικό soundex code.