ΑΣΚΗΣΗ 1

Παντούλας Στέφανος
ΑΜ:1115201400138

Το προγραμμά έχει υλοποιηθεί σύμφωνα με τις απαιτήσεις της εκφώνησης.Εκτώς από τις δύο παραμέτρους που μπορούν να δωθούν κατά την εκτέλεση του(docfile και k) μπορεί να δωθεί επιπλέον ένα flag {–sorted} έτσι ώστε να ταξινομείται το trie κατά την δημιουργια του(μετά από δοκιμές, η αρχικοποίηση του trie εκτελείται πιο γρήγορα χωρις ταξινόμηση, ενώ οι λειτουργείες χρειάζονται περίπου ίδιο χρόνο για να εκτελεστούν).
Ενδεικτική εκτέλεση: ./minisearch -i docfile [-sorted] [-k 5]

Εάν δε δωθεί αρχείο ή δεν υπάρχει αυτο που δώθηκε εμφανίζεται κατάλληλο μήνυμα λάθους όπως επίσης και άν δωθεί κ=0. Αν το αρχείο υπάρχει προχωράμε στον έλεγχο και αποθήκευση των documents του στην μνήμη,ενώ αν βρεθεί κάποιο λάθος ελεύθερώνεται ο δεσμευμένος χώρος και το πρόγραμμα τερματίζει με κατάλληλο μήνυμα.Κατά τον έλεγχο του αρχείου,δεχεται-προσπερνάει τις κενές γραμμές ενδιάμεσα στα κείμενα. Επίσης, μπορουν να υπάρχουν οσαδήποτε κενά πριν και μετά το id του κειμένου.Αμέσως μετά(τον έλεγχο και αποθήκευση) με τη σειρά από το πρώτο μέχρι το τελευταίο document τα χωρίζουμε σε λέξεις,τις οποίες εισάγουμε στο trie.

Το trie έχει υλοποιηθεί ως ένα δέντρο με κόμβους με στοιχεία: έναν χαρακτήρα, δύο δείκτες ίδιας δομής(next και child) και μία posting list(Γίνεται ολική απόκρυψη σε όλες τις δομές που χρησημοποιεί το πρόγραμμα).Η posting list που έχει κάθε κόμβος του δέντρου περιέχει έναν μετρητή(σε πόσα διαφορετικά έγγραφα εμφανίζεται η λέξη) καθώς και δείκτες στην αρχή και στο τέλος της πραγματικής posting list(η οποία έχει υλοποιηθεί ακριβώς όπως περιγράφεται στην εκφώνηση).Ο δείκτης στο τέλος της posting list βοηθάει αρκετά στην ενημέρωση της,διότι κατά την δημιουργία του δέντρου κάθε φορά που θα εισάγεται ίδια λέξη το id της θα είναι πάντα ≥ του τελευταίου id που υπάρχει στην posting list.Έτσι η posting list παραμένει πάντα ταξινομημένη σύμφωνα με το id και θα γίνεται άμεση ενημέρωση χωρίς αναζήτηση του id.Βέβαια,αν τυχών χρειαστεί να ενημερωθεί η posting list μετα τη δημιουργία του δέντρου με id μικρότερο του τελευταίου τότε θα γίνει κανονικά η αναζήτηση(αλλά δε θα γίνει ποτέ αυτό στην ασκηση αφού το trie δεν πειράζεται μετά την δημιουργία του).

/search ή \search:

Αποθηκεύει τις λέξεις προς αναζήτηση σε μία δομή ώστε να βρεί και να κρατήσει για όσο χρειαστεί πληροφορίες για την εμφάνιση των κειμένων.Πιό συγκεκριμένα οι πληροφορίες που χρειαζόμαστε είναι η ίδια η λέξη καθώς και ο δείκτης στον κόμβο του trie που βρίσκεται η posting list της αν υπάρχει.Μόλις τα βρούμε αυτά,πέρνουμε τα id από τις posting list κάθε λέξης και τα συγχωνεύουμε σε έναν πίνακα χωρίς διπλότυπα.Στη συνέχεια, για κάθε id βρίσκουμε το σκορ του σύμφωνα με τον τύπο BM25 και με χρήση quicksort ταξινομούμε παράλληλα τους δύο πίνακες [score] και [id] με βάση το score.Η quicksort αν και με χείριστη πολ/τα O(N^2) συνήθως αποτελεί αρκετά γρήγορο αλγόριθμο και στην πρά-ξη δίνει ευχάριστα αποτελέσματα(τον αλγόριθμο(υλοποιημένο σε c++) τον πήρα (σχεδόν) έτοιμο από δική μου υλοποίηση σε προηγούμενο μάθημα(project)).Τέλος,εμφανίζουμε τα top k ή ίσως λιγότερα αν έχουμε βρεί λιγότερα απο κ κείμενα.Ελευθερώνουμε τον χώρο που δεσμεύσαμε για τις λέξεις και περιμένουμε νέα εντολή.

/df ή \df:

Δέχεται από 0 εώς ‘απεριόριστο’ πλήθος λέξεων. Αν δωθεί σκέτη η εντολή,τότε καλείται μία αναδρομική συνάρτηση που διασχίζει κάθε κόμβο του trie,κρατάει δυναμικά ένα string που συμπληρώνει όσο κατεβαίνει το δέντρο(δείκτες child).Όταν βρεί posting list εκτυπώνει την λέξη που έχει σχηματίσει μέχρι τώρα καθώς και τον document counter που έχουμε κρατησεί αποθηκευέμο.Όταν ο δείκτης child γίνει NULL επιστρέφει και καλείται η ίδια συνάρτηση με τον next,αλλάζοντας όμως το γράμμα και όχι συμπληρωνοντας το επόμενο που κάναμε πιο πριν.Με αυτό τον τρόπο,στο τέλος θα έχουμε εκτυπώσει όλες τις λέξεις του δέντρου. Αν,όμως,δωθούν λέξεις,τότε για κάθε μία απ αυτές,διασχίζει το δέντρο αναζητώντας την και επιστρέφει τον document counter της αν υπάρχει αλλίως επιστρέφει 0.

/tf ή \tf:

Υποχρεωτικά σύνταξη όπως στην εκφώνηση.Για την λέξη που θα δωθεί, διασχίζει το δέντρο αναζητώντας την ακριβώς όπως πρίν αλλά μόλις την βρεί,αναζητάει στην posting list της το id και επιστρέφει το σχετικό frequency.Επιστρέφει 0 αν δε βρεθεί το id ή δεν υπάρχει καθόλου η λέξη.

/exit ή \exit:

Απλώς σταματάει τον βρόγχο της main που περιμένει συνέχεια εντολές απ τον χρήστη.Ελευθερώνονται όλες οι δομές που χρειαστήκαμε και τελείωνει επιτυχώς το πρόγραμμα.