Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Τμήμα Πληροφορικής & Τηλεπικοινωνιών
Προγραμματισμός Συστήματος

Γιώργος Κατσογιάννης-Μεϊμαράκης
sdi1400065@di.uoa.gr
Εαρινό Εξάμηνο 2017-18

Minisearch

1 Μεταγλώττιση & Εκτέλεση

- Η μεταγλώττιση του προγράμματος γίνεται χρησιμοποιώντας το Makefile με την εντολή make.
- Το πρόγραμμα εκτελείται ως εξής: `./minisearch -k <K> -i <inputFile>`

1.1 Εντολές Προγράμματος

Το πρόγραμμα υλοποιεί όλες τις ζητούμενες λειτουργίες της άσκησης:

- `search <q1> [q2 ... q10]`
- `tf <index> <word>`
- `df [word]`

Και για τον τερματισμό του προγράμματος χρησιμοποιείται η εντολή:

- `exit`

2 Δομή Προγράμματος

Το πρόγραμμα είναι χωρισμένο σε διαφορετικά αρχεία, καθένα από τα οποία υλοποιεί κάποια διαφορετική λειτουργία και δομή δεδομένων. Συγκεκριμένα, τα αρχεία:

- `minisearch.c`
- `textIndex.c/.h`
- `trie.c/.h`
- `postingList.c/.h`

3 Text Index

Η δομή text index είναι υπεύθυνη για το διάβασμα του αρχείου, την αποθήκευση του και την προσπέλαση των κειμένων και άλλων στατιστικών σχετικά με το αρχείο. Τέτοια στατιστικά είναι το πλήθος κειμένων του αρχείου και το συνολικό πλήθος λέξεων όλων τον κειμένων. Ακόμα μπορεί να υπολογίσει πόσες λέξεις έχει κάθε κείμενο. Όλα αυτά είναι απαραίτητα για τον υπολογισμό τους σκορ στην εντολή search.

4 Trie

Η δομή trie δημιουργείται με βάση ένα text index. Αποθηκεύει κάθε λέξη των κειμένων του και για κάθε λέξη δημιουργεί μια posting list η οποία περιέχει τις εμφανίσεις της λέξης.

4.1 Trie Node

Το trie αποτελείται από nodes, καθένα από τα οποία έχει κάποια τιμή (π.χ. ένα γράμμα), ένα δείκτη σε posting list αν το μονοπάτι από τη ρίζα μέχρι τον κόμβο σχηματίζει μια λέξη του κειμένου (αλλιώς NULL) και δείκτες σε άλλα nodes. Συγκεκριμένα έχει δείκτες στο γονιό του (προηγούμενο γράμμα λέξης), στο παιδί του με τη μικρότερη τιμή (επόμενο γράμμα λέξης) και στον προηγούμενο (με μικρότερή τιμή) και επόμενο (με μεγαλύτερη τιμή) αδερφό του (άλλα πιθανά γράμματα λέξης στη συγκεκριμένη θέση).

4.2 Λειτουργίες Trie

Η δομή trie υλοποιεί ουσιαστικά όλες τις εντολές του προγράμματος:

4.2.1 Εντολή search

Βρίσκουμε την posting list (αν υπάρχει) κάθε όρου της αναζήτησης και υπολογίζουμε και αποθηκεύουμε σε έναν πίνακα το σκορ κάθε κειμένου που εμφανίζεται στις posting lists των όρων. Ταξινομούμε τα k μεγαλύτερα σκορ και τα εκτυπώνουμε.

4.2.2 Εντολή tf

Βρίσκουμε την posting list της λέξης στο trie και ψάχνουμε τον κόμβο της που αναφέρεται στο index του κειμένου που δόθηκε. Αν βρεθεί εκτυπώνουμε τον αντίστοιχο αριθμό εμφανίσεων σε αυτό το κείμενο. Αλλιώς εκτυπώνουμε 0.

4.2.3 Εντολή df

Εάν δοθεί όρισμα λέξης στην εντολή, αναζητείται η συγκεκριμένη λέξη στο trie και εκτυπώνεται το μήκος της posting list της, δηλαδή σε πόσα κείμενα εμφανίζεται.
Εάν δε δοθεί λέξη, διασχίζεται όλο το trie πρώτα κατά βάθος (άρα οι λέξεις συναντιούνται σε αύξουσα σειρά) και όποτε συναντάμε κόμβο με posting list (δηλαδή μονοπάτι που εκφράζει λέξη), εκτυπώνεται η λέξη και το πλήθος κειμένων που περιέχεται.

5 Posting List

Η posting list είναι μια μονά συνδεδεμένη λίστα, της οποίας ο κάθε κόμβος περιέχει το index του αντίστοιχου κειμένου του αρχείου στο οποίο εμφανίζεται η λέξη στην οποία αναφέρεται η λίστα και το πόσες φορές εμφανίζεται σε αυτό το κελίμενο.

5.1 Ταξινομημένη σειρά

Η εισαγωγή νέων στοιχείων γίνεται έτσι ώστε οι κόμβοι της λίστας να είναι ταξινομημένοι ως προς το index του κειμένου τους σε αύξουσα σειρά.