Εργασία 2 – Αξιολόγηση κριτικών ταινιών

Συνοπτική περιγραφή

Αναπτύξτε μια εφαρμογή αξιολόγησης κριτικών για ταινίες. Η εφαρμογή συλλέγει δεδομένα από ένα μεγάλο όγκο κριτικών και τα χρησιμοποιεί για να αξιολογήσει κατά πόσο μια νέα κριτική είναι θετική, αρνητική ή ουδέτερη. Επιπλέον, παρέχει τη δυνατότητα περιορισμένης επεξεργασίας των δεδομένων που έχει συλλέξει.

Τα δεδομένα που συλλέγονται για κάθε μία κριτική είναι το όνομα της ταινίας, το όνομα του ατόμου που έγραψε την κριτική, η ίδια η κριτική και ένας βαθμός αξιολόγησης της ταινίας, ο οποίος είναι στο εύρος [0, 4] με το 0 να αντιστοιχεί σε αρνητική αξιολόγηση, το 4 σε θετική, κι ενδιάμεσες τιμές σε αντίστοιχες διαβαθμίσεις.

Για παράδειγμα, δίνονται οι παρακάτω ενδεικτικές κριτικές:

"Dune", "Brown", "An epic and spectacular sci-fi allegory with mass appeal", 3.5

"Dune", "Ewing", "Dune is a stunning sweeping epic with a cast in command of their roles", 4

"Spaceman", "Schwartz", "A dreary sci-fi drama on dealing with grief", 1

Το πρόγραμμα αποθηκεύει τις επιμέρους λέξεις κάθε κριτικής σε ένα πίνακα κατακερματισμού. Για κάθε μία λέξη καταγράφονται το συνολικό άθροισμα των βαθμών αξιολόγησης των κριτικών στις οποίες εμφανίζεται καθώς και το πλήθος εμφανίσεων της λέξης. Το πηλίκο αυτών των τιμών (μέσος όρος) αποτελεί το μέσο βαθμό αξιολόγησης της λέξης. Για παράδειγμα, με βάση τις παραπάνω κριτικές, ο μέσος βαθμός αξιολόγησης της λέξης "epic" είναι (3.5+4)/2 = 3.75 Αυτό σημαίνει ότι η λέξη epic είναι γενικά "θετική". Αντίστοιχα, για τη λέξη "with" ο μέσος βαθμός είναι (3.5+4+1)/3 = 2.83 ενώ για τη λέξη "a" είναι (4+4+1)/3 = 3.0.

Εάν ζητηθεί να αξιολογηθεί μια νέα κριτική για την οποία δεν υπάρχει βαθμός αξιολόγησης, τότε μπορεί να εκτιμηθεί ένας βαθμός γι' αυτή υπολογίζοντας το άθροισμα των μέσων βαθμών των επιμέρους λέξεων δια το πλήθος των λέξεων στην κριτική. Αν για κάποια λέξη δεν υπάρχει βαθμός, χρησιμοποιείται η τιμή 2. Για παράδειγμα, αν το κείμενο μιας νέας κριτικής είναι "Epic sci-fi film", τότε ο βαθμός της με βάση τα παραπάνω δεδομένα είναι (3.75 + 2.25 + 2.0)/3 = 2.67

Έτοιμος κώδικας

Σας δίνονται μια στατική βιβλιοθήκη libhw2.a και τα αρχεία hw2.h, htable.h, htable.c.

Το αρχείο hw2.h περιλαμβάνει:

- Τον ορισμό του τύπου struct review_t με πεδία για τα στοιχεία μίας κριτικής (όνομα ταινίας ως δυναμικά δεσμευμένη συμβολοσειρά, όνομα κριτικού ως δυναμικά δεσμευμένη συμβολοσειρά, κείμενο κριτικής ως δυναμικά δεσμευμένη συμβολοσειρά, βαθμός κριτικής).
- Την επικεφαλίδα της συνάρτησης read_review η υλοποίηση της οποίας δίνεται στη στατική βιβλιοθήκη. Η συνάρτηση παίρνει ως παράμετρο το όνομα ενός αρχείου που περιέχει πληροφορίες κριτικών. Κάθε φορά που καλείται επιστρέφει δείκτη σε ένα δυναμικά δεσμευμένο struct review_t που περιέχει τα στοιχεία της επόμενης κριτικής που βρίσκεται στο αρχείο ή NULL αν το αρχείο δεν περιέχει άλλα δεδομένα.
- Την επικεφαλίδα της συνάρτησης print_menu η υλοποίηση της οποίας δίνεται στη στατική βιβλιοθήκη. Η συνάρτηση εκτυπώνει ένα μενού επιλογών.
- Την επικεφαλίδα της συνάρτησης read_line η υλοποίηση της οποίας δίνεται στη στατική βιβλιοθήκη. Η συνάρτηση παίρνει ως παράμετρο ένα δείκτη προς συμβολοσειρά (char**), διαβάζει από το πληκτρολόγιο μία γραμμή κειμένου, δεσμεύει επαρκή μνήμη και αποθηκεύει το κείμενο στη συμβολοσειρά. Επιστρέφει 0 αν η ανάγνωση είναι επιτυχής, ή -1 σε περίπτωση λάθους.

Το αρχείο htable.h περιλαμβάνει την επικεφαλίδα της συνάρτησης κατακερματισμού hash η υλοποίηση της οποίας βρίσκεται στο htable.c. Σημειώστε πως η τιμή που επιστρέφει η συνάρτηση κατακερματισμού πρέπει να προσαρμόζεται κατάλληλα με βάση το τρέχον μέγεθος του πίνακα κατακερματισμού (modulo το μέγεθος του πίνακα) έτσι ώστε να προκύπτουν θέσεις δοχείων μέσα στα επιτρεπτά όρια του πίνακα.

Δομές δεδομένων

Οι κριτικές (όπως διαβάζονται από το αρχείο) πρέπει να αποθηκεύονται σε μια δυναμικά δεσμευμένη δομή. Η μορφή αυτής της δομής είναι δική σας επιλογή, αλλά σε κάθε περίπτωση πρέπει να είναι δυναμική με μέγεθος ανάλογο του πλήθους δεδομένων που περιέχει.

Οι λέξεις που εμφανίζονται στις κριτικές και οι πληροφορίες αξιολόγησης αυτών πρέπει να αποθηκεύονται σε έναν πίνακα κατακερματισμού ο οποίος πρέπει να υλοποιηθεί ως δυναμικός πίνακας από διπλά διασυνδεδεμένες, κυκλικές λίστες με τερματικό κόμβο (sentinel). Κάθε κόμβος λίστας πρέπει να περιέχει μια λέξη (ως δυναμικά δεσμευμένη συμβολοσειρά, με όλα τα γράμματα μικρά), το πλήθος κριτικών στις οποίες εμφανίζεται αυτή η λέξη και το άθροισμα βαθμών αξιολόγησης αυτών των κριτικών. Επιπλέον πρέπει να περιέχει ό,τι δείκτες χρειάζονται για την διασύνδεση με άλλους κόμβους. Οι λίστες διατηρούνται ταξινομημένες ως προς τις λέξεις, σε αύξουσα λεξικογραφική σειρά.

Το αρχικό μέγεθος του πίνακα κατακερματισμού είναι αποθηκευμένο στη συμβολική σταθερά INIT_HSIZE η οποία μπορεί να οριστεί κατά τη διάρκεια της μεταγλώττισης με χρήση της επιλογής -D. Εάν δεν έχει οριστεί κατά τη μεταγλώττιση, τότε η τιμή της πρέπει να είναι 1.

Ο μέγιστος επιτρεπτός βαθμός πληρότητας (load factor) του πίνακα κατακερματισμού είναι αποθηκευμένος στη συμβολική σταθερά HIGH_LOAD η οποία μπορεί να οριστεί κατά τη διάρκεια της μεταγλώττισης με χρήση της επιλογής -D. Εάν δεν έχει οριστεί κατά τη μεταγλώττιση, τότε η τιμή της πρέπει να είναι 4.0. Αντίστοιχα, ο ελάχιστος επιτρεπτός βαθμός πληρότητας είναι αποθηκευμένος στη συμβολική σταθερά LOW_LOAD η οποία πρέπει να έχει την τιμή 1.0 αν δεν έχει οριστεί κατά τη μεταγλώττιση.

Η πολιτική ανακατακερματισμού (re-hashing) σε συνδυασμό με την δυναμική αυξομείωση του πίνακα βασίζεται στο βαθμό πληρότητας (load factor) του πίνακα. Συγκεκριμένα: Αν μετά την εισαγωγή νέας εγγραφής ο βαθμός πληρότητας είναι μεγαλύτερος ή ίσος της τιμής HIGH_LOAD τότε ο πίνακας διπλασιάζεται. Αν μετά τη διαγραφή μιας εγγραφής ο βαθμός πληρότητας είναι μικρότερος ή ίσος της τιμής LOW_LOAD, τότε ο πίνακας υποδιπλασιάζεται, μόνο αν αυτό δεν επιφέρει μείωση κάτω από το αρχικό μέγεθος INIT_HSIZE. Σε περίπτωση που αποτυγχάνει η δέσμευση μνήμης κατά το διπλασιασμό του πίνακα κατακερματισμού, τότε το πρόγραμμα συνεχίζει κανονικά χωρίς να γίνει ανακατακερματισμός, με τον πίνακα να παραμένει στο μέγεθος που είχε πριν την εισαγωγή της νέας εγγραφής.

Υλοποίηση

Το πρόγραμμα αρχικοποιεί κατάλληλα τις δομές δεδομένων και κατόπιν σε επανάληψη, καλεί τη συνάρτηση print_menu, διαβάζει την επιλογή του χρήστη (χωρίς να κάνει διάκριση κεφαλαίων/μικρών), και πραγματοποιεί τις απαιτούμενες ενέργειες όπως περιγράφεται παρακάτω για κάθε μια από τις επιλογές.

- Στην επιλογή 'A' εκτυπώνει το μήνυμα "Enter filename: " και διαβάζει το όνομα ενός αρχείου. Ακολούθως, διαβάζει από το αρχείο τα δεδομένα των κριτικών χρησιμοποιώντας επαναληπτικά τη συνάρτηση read_review. Κάθε κριτική αποθηκεύεται στη δομή της επιλογής σας ενώ οι επιμέρους λέξεις της κριτικής αποθηκεύονται στον πίνακα κατακερματισμού, ανανεώνοντας κάθε φορά το πλήθος εμφανίσεων και το άθροισμα βαθμών αξιολόγησης που αντιστοιχούν σε κάθε λέξη. Δε χρειάζεται να κάνετε έλεγχο για τυχόν διπλότυπες κριτικές.
- Στην επιλογή 'S' εκτυπώνει το μήνυμα "Enter word: ", διαβάζει μια λέξη από το πληκτρολόγιο, την εντοπίζει στον πίνακα κατακερματισμού κι εκτυπώνει το μήνυμα "W: Z\n" όπου W η λέξη που δόθηκε με όλα τα γράμματα μικρά και Z ο μέσος βαθμός αξιολόγησης αυτής με δύο δεκαδικά ψηφία. Αν η λέξη δεν υπάρχει στον πίνακα κατακερματισμού, εκτυπώνεται το μήνυμα "W not found.\n".

- Στην επιλογή 'I' εκτυπώνει το μήνυμα "Enter review: " και διαβάζει μια νέα κριτική. Η κριτική δεν συνοδεύεται από τίτλο ταινίας / όνομα κριτικού ούτε αποθηκεύεται στη δομή αποθήκευσης κριτικών. Το πρόγραμμα υπολογίζει ένα βαθμό αξιολόγησης για την κριτική όπως περιγράφεται παραπάνω κι εκτυπώνει το μήνυμα "S review (X).\n" όπου X είναι ο βαθμός που υπολογίστηκε με 2 δεκαδικά ψηφία και S είναι η λέξη Positive αν ο βαθμός είναι στο εύρος [3,4], Neutral αν ο βαθμός είναι στο εύρος [2, 3) ή Negative αν ο βαθμός είναι στο εύρος [0, 2).

- Στην επιλογή 'B' βρίσκει τη λέξη (ή τις λέξεις) με το μέγιστο βαθμό αξιολόγησης. Δύο βαθμοί αξιολόγησης θεωρούνται ίσοι αν δε διαφέρουν περισσότερο από 0.005. Γράψτε μια βοηθητική συνάρτηση η οποία επιστρέφει μια απλά συνδεδεμένη λίστα με τις λέξεις που έχουν το μέγιστο βαθμό αξιολόγησης (μία λέξη σε κάθε κόμβο). Η λίστα πρέπει να είναι ταξινομημένη σε αύξουσα λεξικογραφική σειρά. Το πρόγραμμα εκτυπώνει το μήνυμα "\nMost positive words (S):\n" όπου S ο μέγιστος βαθμός αξιολόγησης με δύο δεκαδικά ψηφία ή μία παύλα αν δεν έχουν βρεθεί λέξεις και ακολούθως οι λέξεις που περιέχει η λίστα (με μικρά γράμματα), με ένα κενό (space) πριν από κάθε λέξη κι ένα χαρακτήρα '\n' μετά την τελευταία.

- Στην επιλογή 'W' βρίσκει τη λέξη (ή τις λέξεις) με τον ελάχιστο βαθμό αξιολόγησης. Γράψτε μια βοηθητική συνάρτηση στο πνεύμα της παραπάνω και εκτυπώστε τα αποτελέσματα όπως περιγράφεται στην επιλογή 'B' αλλά με αρχικό μήνυμα το "\nMost negative words (S):\n".

- Στην επιλογή 'P' εκτυπώνει το μήνυμα "Enter movie title: ", και διαβάζει τον τίτλο μιας ταινίας ή έναν αστερίσκο ('*'). Αν διάβασε τον τίτλο μιας ταινίας, εκτυπώνει το μήνυμα "\nReviews for T:\n" όπου T ο τίτλος. Αν υπάρχουν κριτικές γι' αυτή την ταινία στη δομή κριτικών, τις εκτυπώνει ως εξής: Για κάθε κριτική εκτυπώνει το βαθμό αξιολόγησης της κριτικής με δύο δεκαδικά ψηφία, κόμμα, κενό, το όνομα του κριτικού, κόμμα, κενό, την κριτική και τέλος χαρακτήρα '\n'. Η σειρά εκτύπωσης κριτικών δεν έχει σημασία. Αν δεν υπάρχουν, εκτυπώνει το μήνυμα "No reviews found.\n". Αν διάβασε αστερίσκο, τότε εκτυπώνει τα στοιχεία κάθε μίας κριτικής που είναι αποθηκευμένη στη δομή κριτικών. Αυτή η λειτουργία είναι κυρίως για δική σας χρήση (debugging). Θα σας δοθούν ενδεικτικά τεστ, αλλά στο autolab θα ελεγχθούν μόνο περιπτώσεις που μία ταινία έχει καμία ή μία κριτική.

- Στην επιλογή 'R' εκτυπώνει το μήνυμα "Enter name: ", διαβάζει από το πληκτρολόγιο το όνομα ενός ατόμου και αφαιρεί από τη δομή κριτικών όλες τις κριτικές που γράφτηκαν από αυτό το άτομο, ανανεώνοντας κατάλληλα τον πίνακα κατακερματισμού. Στο τέλος, εκτυπώνει το μήνυμα "\nRemoved X reviews by N.\n", όπου X το πλήθος κριτικών που αφαιρέθηκαν και N το όνομα του ατόμου.

- Στην επιλογή 'D' εκτυπώνει τα περιεχόμενα του πίνακα κατακερματισμού ως εξής: Αρχικά εκτυπώνονται δύο χαρακτήρες '\n'. Στην επόμενη γραμμή εκτυπώνονται με κενό ανάμεσα τους το μέγεθος του πίνακα κατακερματισμού, ο συνολικός αριθμός των εγγραφών, το load factor (πλήθος εγγραφών/μέγεθος πίνακα) με δύο δεκαδικά ψηφία, και το μέγεθος του μεγαλύτερου δοχείου.

  Στην συνέχεια, ξεκινώντας από καινούργια γραμμή, για κάθε θέση του πίνακα, εκτυπώνονται σε μια γραμμή ο αριθμός της θέσης , χαρακτήρας ':', κενό, το πλήθος λέξεων του δοχείου και, εφόσον το δοχείο δεν είναι άδειο, στην ίδια γραμμή για κάθε λέξη του δοχείου εκτυπώνονται: κενό, χαρακτήρας '[', κενό, η λέξη, κενό, το πλήθος εμφανίσεων της, κενό, το συνολικό άθροισμα βαθμολογιών της, κενό και χαρακτήρας ']'. Μετά το τέλος των περιεχομένων του δοχείου (ανεξαρτήτως αν είναι άδειο ή όχι), εκτυπώνονται δύο χαρακτήρες '\n'.

- Στην επιλογή 'Q' αποδεσμεύει όλη τη δυναμικά δεσμευμένη μνήμη του προγράμματος και τερματίζει.

- Σε οποιαδήποτε άλλη επιλογή, το πρόγραμμα εκτυπώνει το μήνυμα "Invalid selection.\n" και συνεχίζει την επανάληψη.

Απαιτήσεις υλοποίησης

- Απαγορεύεται η χρήση global ή static μεταβλητών καθώς και η χρήση goto, gets ή fflush.
- Απαγορεύεται η χρήση variable length arrays.
- Δε γίνεται διάκριση κεφαλαίων/μικρών κατά την εισαγωγή δεδομένων ή κατά τη σύγκριση συμβολοσειρών. Θα σας φανεί χρήσιμη η συνάρτηση strcasecmp.
- Κατά την είσοδο δεδομένων από το χρήστη τα οποία δυνητικά αποτελούνται από πολλές λέξεις (π.χ. κριτική, όνομα κριτικού, τίτλος ταινίας) όλες οι λέξεις δίνονται σε μία γραμμή.
- Τα δεδομένα που εισάγονται στον πίνακα αποθήκευσης κριτικών δε χρειάζονται επεξεργασία πριν την εισαγωγή τους όσον αφορά κεφαλαία/μικρά/κενά. Τα εισάγετε (και αργότερα τα εκτυπώνετε) στη μορφή που δίνονται αρχικά. Η μόνη μετατροπή που χρειάζεται να κάνετε είναι στις λέξεις που εισάγονται στον πίνακα κατακερματισμού, οι οποίες πρέπει να είναι με όλα τα γράμματα μικρά.
- Για την απόσπαση των επιμέρους λέξεων μιας κριτικής θα σας φανεί χρήσιμη η συνάρτηση strtok. Διαβάστε προσεκτικά το manual προτού την χρησιμοποιήσετε.
- Για την υλοποίηση του πίνακα κατακερματισμού κατασκευάστε ένα struct το οποίο περιέχει ό,τι πεδία πιστεύετε πως χρειάζονται για τη λειτουργία της δομής (π.χ. δυναμικό πίνακα από λίστες, μέγεθος πίνακα, συνολικό πλήθος λέξεων, κτλ.).
- Αν το πλήθος εμφανίσεων μιας λέξης κριτικής γίνει μηδέν (λόγω αφαίρεσης όλων των κριτικών στις οποίες εμφανίζεται), τότε ο κόμβος λίστας που περιέχει τη λέξη πρέπει να αφαιρείται από το αντίστοιχο bucket του πίνακα κατακερματισμού.
- Ορίστε συναρτήσεις για συγκεκριμένες λειτουργίες. Ειδικότερα, για κάθε μία από τις επιλογές του χρήστη, φτιάξτε κατάλληλη συνάρτηση που υλοποιεί τα ζητούμενα και καλείται μέσα από την main. Η διαχείριση των επιλογών πρέπει να γίνει με κατάλληλη χρήση switch που στα cases περιέχει μόνο κλήσεις των αντίστοιχων συναρτήσεων και break, continue ή return.
- Στο hw2.c, για όλες τις λειτουργίες που αφορούν λίστες ή τον πίνακα κατακερματισμού, θα πρέπει να καλείτε κατάλληλες συναρτήσεις που ορίζονται στα αντίστοιχα αρχεία.
- Σε περίπτωση αποτυχίας δέσμευσης μνήμης (εκτός από την περίπτωση διπλασιασμού του πίνακα κατακερματισμού) εκτυπώνετε το μήνυμα "Memory error.\n" και το πρόγραμμα τερματίζει άμεσα με χρήση εντολής exit.
- Μην ξεχάσετε να διαχειριστείτε οριακές περιπτώσεις όπου χρειάζεται, π.χ. αν επιλεγεί 'B' και ο πίνακας κατακερματισμού είναι άδειος, τότε θα επιστραφεί άδεια λίστα και δε θα εκτυπωθούν λέξεις.
- Φροντίστε να αποδεσμεύετε πάντα σωστά τη δυναμικά δεσμευμένη μνήμη. Πρέπει να αποφύγετε τα memory leaks καθώς και την αποδέσμευση μνήμης που έχει ήδη αποδεσμευθεί (double free).
- Το πρόγραμμα πρέπει να ακολουθεί τους κανόνες μορφοποίησης, σχολιασμού, κτλ. που έχουν αναρτηθεί στο eclass, καθ' όλη τη διάρκεια της ανάπτυξης του.

Οργάνωση αρχείων και μεταγλώττιση

Ο κώδικας του προγράμματος σας πρέπει να περιέχεται στα παρακάτω αρχεία:

- slist.h: περιέχει ορισμούς struct και επικεφαλίδες συναρτήσεων που αφορούν λειτουργίες για απλά διασυνδεδεμένες λίστες.
- slist.c: περιέχει τις υλοποιήσεις των συναρτήσεων που ορίζονται στο slist.h
- dlist.h: περιέχει ορισμούς struct και επικεφαλίδες συναρτήσεων που αφορούν λειτουργίες για διπλά διασυνδεδεμένες, κυκλικές λίστες με τερματικό κόμβο.
- dlist.c: περιέχει τις υλοποιήσεις των συναρτήσεων που ορίζονται στο dlist.h
- htable.h: περιέχει ορισμούς struct και επικεφαλίδες συναρτήσεων που αφορούν λειτουργίες του πίνακα κατακερματισμού.
- htable.c: περιέχει τις υλοποιήσεις των συναρτήσεων που ορίζονται στο htable.h.
- hw2.h: περιέχει ορισμούς struct και επικεφαλίδες συναρτήσεων που αφορούν γενικές λειτουργίες της εργασίας.
- hw2.c: περιέχει τη main και τις υλοποιήσεις των βοηθητικών συναρτήσεων που θα γράψετε για την εργασία (που δεν αφορούν τις υλοποιήσεις των λιστών και του πίνακα κατακερματισμού).

Για να μεταγλωττίσετε τον κώδικα σας γράψτε:

```
gcc -Wall -g slist.c dlist.c htable.c hw2.c -lhw2 -L. -o hw2
```

ή, αν ο κατάλογος δεν περιλαμβάνει άλλα C αρχεία εκτός των παραπάνω,

```
gcc -Wall -g *.c -lhw2 -L. -o hw2
```

Αργότερα θα μάθετε ένα πιο ορθό τρόπο μεταγλώττισης και διασύνδεσης πολλαπλών αρχείων.

Υποβολή κώδικα

Κατασκευάστε ένα αρχείο zip που περιέχει τα αρχεία c και h της εφαρμογής σας ως εξής:

- (μέσω τερματικού) Μεταβείτε στον κατάλογο που βρίσκονται τα αρχεία και γράψτε την εντολή `zip -r hw2.zip ?list.[ch] htable.[ch] hw2.[ch]`

  ή, αν δεν υπάρχουν άλλα αρχεία c και h στον κατάλογο, γράψτε πιο απλά:

  `zip -r hw2.zip *.[ch]`
- (ή μέσω GUI) Μεταβείτε στον κατάλογο που βρίσκονται τα αρχεία, επιλέξτε με το ποντίκι τα slist.c, slist.h, dlist.c, dlist.h, htable.c, htable.h, hw2.c, hw2.h, μετά δεξί click και επιλέξτε από το μενού Compress here as ZIP.

Υποβάλετε το hw2.zip στο autolab του μαθήματος. Ο ελάχιστος βαθμός που πρέπει να πετύχετε είναι 70.

Το μέγιστο πλήθος υποβολών χωρίς βαθμολογική ποινή είναι 30. Κάθε υποβολή πλέον των 30 έχει ποινή ίση με 1% του τελικού βαθμού.

Παράδοση

Κυριακή 21/4/2024, 21:00

Δεν υπάρχει δυνατότητα καθυστερημένης υποβολής.