Πρώτο Σετ Ασκήσεων: Διαχείριση Σύνθετων Δεδομένων
Βασιλειάδης Μιλτιάδης 2944

Όλα τα προγράμματα περιέχονται σε .py αρχεία συμβατά με την version 3.5+ της python εκτελούνται με την εντολή python (ή python3 αν υπάρχει και η python2.7 στο σύστημα) <filename> έχουν hardcoded ονόματα αρχείων τα ονόματα των αρχείων που δόθηκαν ως input και παράγουν αυτό το αρχείο όπως είναι ονομασμένο στην εκφώνηση. Τα αρχεία πρέπει να είναι στον τοπικό κατάλογο και να υπάρχουν τα δικαιώματα ανάγνωσης.

Η υλοποίηση της ένωσης βάση της ισότητας του πρώτου πεδίου της κάθε πλειάδας είναι σχετικά απλή αφού οι πλειάδες δίνονται ήδη ταξινομημένες. Αρχικοποιούμε έναν άδειο buffer και ξεκινά η ανάγνωση των αρχείων. Για κάθε πλειάδα του R που διαβάζεται γίνεται έλεγχος του buffer αν περιέχει πλειάδες αν ναι ελέγχεται αν οι πλειάδες του buffer έχουν το ίδιο πεδίο σύγκρισης με το αντίστοιχο του R που μόλις διαβάστηκε αν ναι γίνεται το join γράφοντας στο αρχείο το αλφαριθμητικό που γίνεται ταύτιση tab αριθμητικό πεδίο της πλειάδας του R tab αριθμητικό πεδίο της πλειάδας του S. Αν αυτό που διαβάσαμε δεν αντιστοιχεί με αυτό που υπάρχει στο buffer τότε καθαρίζει το buffer και ξεκινά η ανάγνωση των πλειάδων του S από το αρχείο. Όπου για κάθε σύγκριση που καταλήγει σε ισότητα με το R γράφεται στο αρχείο το αποτέλεσμα του join και η πλειάδα του S εισάγεται στο buffer. Το διάβασμα του S συνεχίζεται μέχρις ότου η πλειάδα που θα διαβαστεί να ταιριάζει με την πλειάδα του R. Όταν συμβεί αυτό διαβάζεται η επόμενη πλειάδα του R γίνεται η σύγκριση αν ταιριάζει με το S αν ναι γράφεται στο αρχείο το join και συνεχίζεται η ανάγνωση του S όπως περιγράφεται επάνω αλλιώς αν το πεδίο του S είναι μεγαλύτερο από το πεδίο του R τότε ξέρουμε ότι αν διαβάσουμε επόμενο S θα είναι κι αυτό mismatch οπότε γυρνάμε στην αρχή και διαβάζουμε το επόμενο R. Η διαδικασία αυτή συνεχίζεται μέχρι να διαβαστούν οι τελευταίες πλειάδες των R και S.

Για την πράξη της ένωσης η υλοποίηση ξεκινά διαβάζοντας τις πρώτες πλειάδες από τα S και R. Γίνεται το έλεγχος μεταξύ τους αν R<S γράφουμε το R στο αρχείο, το R μπαίνει στο R_prev και διαβάζεται η επόμενη πλειάδα από το R αν S<R γράφεται το S στο αρχείο αν R=S τότε γράφεται οποιοδήποτε από τα δύο στο αρχείο και διαβάζονται νέα R και S όπως περιγράφεται επάνω. Κάθε φορά που διαβάζεται ένα νέο R ή S σε μία λούπα ελέγχεται αν αυτό που διαβάστηκε είναι ίδιο με το προηγούμενο αν ναι διαβάζεται εκ νέου μια πλειάδα μέχρι ότου να μην είναι το ίδιο με το προηγούμενο.

Τέλος αν κάποιο από τα S ή R φτάσουν στο τέλος με μία νέα λούπα διαβάζονται οι πλειάδες και γράφονται στο αρχείο αποφεύγοντας πάλι να γράψουμε διπλότυπα στην έξοδο.

Πρώτο Σετ Ασκήσεων: Διαχείριση Σύνθετων Δεδομένων
Βασιλειάδης Μιλτιάδης 2944

Για την πράξη της τομής διαβάζουμε τις πλειάδες από τα αρχεία. Για κάθε πλειάδα που διαβάζεται αν R == S γράφουμε την πλειάδα στο αρχείο. Το S και το R γράφονται σε temp μεταβλητές. Διαβάζονται εκ νέου πλειάδες από τα αρχεία. Για να αποφευχθούν τα διπλότυπα αν αυτό που διαβάστηκε είναι το ίδιο με το προηγούμενο διαβάζεται η επόμενη μεταβλητή από το αρχείο μέχρι ότου να μην ισχύει η ισότητα. Αν R>S διαβάζεται νέα πλειάδα από το S μέχρι να έχουμε πάλι ισότητα ώστε να διαβαστούν εκ νέου νέες πλειάδες. Αν R<S τότε διαβάζεται νέα πλειάδα από το R και η διαδικασία ξεκινά από την αρχή.

Για την διαφορά των συνόλων των πλειάδων ακολουθούμε παρόμοια διαδικασία με την τομή. Με την διαφορά ότι γράφουμε την πλειάδα του R όταν το S που διαβάστηκε είναι μεγαλύτερο του R. Αν το R == S ξαναδιαβάζουμε και από τα 2 αρχεία πλειάδες. Αν η πλειάδα του R είναι μικρότερη της πλειάδας του S την γράφουμε στο αρχείο και διαβάζουμε την επόμενη πλειάδα από το R και ξαναξεκινούν οι έλεγχοι Αν η πλειάδα του S είναι μικρότερη από αυτή του R διαβάζεται εκ νέου νέο S και ξαναξεκινούν οι παραπάνω έλεγχοι (έως ότου δηλαδή διαβαστεί παλι s == r ή R<S. Τέλος αν το S έχει λιγότερες πλειάδες από το R με ένα ξεχωριστό loop περνιούνται οι εναπομείναντες πλειάδες του R στο αρχείο με το αποτέλεσμα. Η ιδέα είναι ότι εφ’ όσον τα δύο αρχεία είναι ταξινομημένα αν το τρέχον R είναι μικρότερο του S τότε σημαίνει ότι όσα S προηγούμενα δεν θα είναι σίγουρα στο R. αν το R είναι μεγαλύτερο διαβάζω το S και όποιες πλειάδες είναι ίσες τις πετάω.

Ο έλεγχος για τα διπλότιμα γίνεται όπως και στις προηγούμενες υλοποιήσεις με επαναληπτική ανάγνωση μέχρις ότου αυτό που διαβάστηκε δεν είναι ίδιο με το προηγούμενο.

Για την πράξη της συνάθροισης χρειάζεται να ταξινομηθούν οι πλειάδες του R. Διαβάζουμε το R σε έναν πίνακα για την υλοποίηση χρησιμοποιείται η merge sort. Σπάει τον πίνακα σε υποπίνακες στην μέση μεχρι να φτάσουν σε μέγεθος μιας πλειάδας γίνονται οι συγκρίσεις κλπ. Στο τέλος της ταξινόμησης διατρέχεται άλλη μια φορά ο ταξινομημένος πλέον πίνακας (λίστα). Βάζουμε το πρώτο στοιχείο σε μία νέα λίστα που θα περιέχει τα groups. Ξεκινώντας τώρα από το δεύτερο στοιχείο του ταξινομημένου πίνακα και μέχρι το τέλος για κάθε στοιχείο του R αν το αλφαριθμητικό πεδίο ταιριάζει με το αλφαριθμητικό στην κεφαλή της λίστας των groups τότε προστίθεται στο αριθμητικό πεδίο του group τον αριθμό από την πλειάδα του R. Αν τα αλφαριθμητικά δεν είναι ίδια τότε προστίθεται η πλειάδα του R στην λίστα των group δημιουργώντας ένα νέο group και επαναλαμβάνεται η παραπάνω διαδικασία. Τέλος η λίστα με τα groups και τα αθροίσματα γράφεται σε ένα αρχείο tsv

Προσπάθησα να κάνω την βελτιστοποιημένη μορφή κάνοντας aggregate μεταξύ των runs αλλά δεν τα κατάφερα. Παρ όλα αυτά συμπεριλαμβάνεται στο αρχείο ως συνάρτηση mergeAggregate(table)