ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΥΕ041 - ΠΛΕ081: Διαχείριση Σύνθετων Δεδομένων (ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2018-19)

ΕΡΓΑΣΙΑ 1 – Αλγόριθμοι Αποτίμησης Συνενώσεων (προθεσμία: 30 Μαρτίου 2019, 9μ.μ.)

Στόχος της εργασίας είναι η ανάπτυξη και ο έλεγχος αλγορίθμων για αποτίμηση (evaluation) σύνθετων ερωτημάτων σε βάσεις δεδομένων τα οποία περιλαμβάνουν συνενώσεις (joins). Στην εργασία θα χρησιμοποιήσουμε δεδομένα από τη βάση της IMDB (Internet Movie Database). Τα δεδομένα και την περιγραφή τους μπορείτε να τα βρείτε εδώ:
https://www.imdb.com/interfaces/
Προσοχή: τα αρχεία είναι πάνω από 2GB, φροντίστε να έχετε αρκετό χώρο στο δίσκο σας. Μελετήστε με προσοχή το σχήμα της βάσης και δείτε τις πρώτες γραμμές των αρχείων ώστε να κατανοήσετε τα περιεχόμενά τους και το πως σχετίζονται μεταξύ τους τα αρχεία.

Μέρος 1 (40%, merge-join combined with grouping)

Στο πρώτο μέρος της εργασίας αυτής θα υλοποιήσετε έναν merge join αλγόριθμο ο οποίος θα διαβάσει τα δεδομένα από τα title.basics.tsv και title.akas.tsv. Το πρώτο αρχείο περιέχει γενικές πληροφορίες για κάθε τίτλο (π.χ. ταινία, σειρά, επεισόδιο) στη βάση. Το δεύτερο αρχείο περιέχει εναλλακτικές ονομασίες για τίτλους (ενδεχομένως σε διάφορες γλώσσες). Και τα δύο είναι αρχεία κειμένου, όπου σε κάθε γραμμή (εκτός από την πρώτη) υπάρχουν πλειάδες τιμών χωρισμένες με tab (tab separated files). Στόχος μας είναι να συνενώσουμε τα αρχεία με βάση το πρώτο πεδίο τους το οποίο είναι ένας μοναδικός identifier για κάθε τίτλο. Για κάθε τίτλο που εμφανίζεται και στα δύο αρχεία εισόδου, θα πρέπει να τυπώνουμε τον identifier, τον primaryTitle, ο οποίος βρίσκεται στο title.basics.tsv και για κάθε εναλλακτικό τίτλο του στο title.akas.tsv, τον εναλλακτικό τίτλο και σε παρένθεση τις περιοχές στις οποίες χρησιμοποιείται ο εναλλακτικός τίτλος. Το αποτέλεσμα θα είναι ένα tab separated αρχείο εξόδου. Παρακάτω βλέπετε τις πρώτες 5 γραμμές του αρχείου εξόδου.

titleId	primaryTitle	title (regions)				
tt0000001	Carmencita	Carmencita - spanyol tánc (HU)	Καρμενσίτα (GR)	Карменсита (RU)	Carmencita (US,\N)	
tt0000002	Le clown et ses chiens	Le clown et ses chiens (\N,FR)	A bohóc és kutyái (HU)	Clovnul si cainii sai (RO)	Клоун и его собаки (RU)	The Clown and His Dogs (US)
tt0000003	Pauvre Pierrot	Sarmanul Pierrot (RO)	Szegény Pierrot (HU)	Бедный Пьеро (RU)	Pauvre Pierrot (\N,FR)	Poor Pierrot (\N)
tt0000004	Un bon bock	Un bon bock (\N,FR)	Un ţap de bere (RO)	Полная кружка пива (RU)	A Good Beer (\N)	Egy jó pohár sör (HU)

Γράψτε ένα πρόγραμμα το οποίο θα διαβάζει τα 2 αρχεία title.basics.tsv και title.akas.tsv και θα δημιουργεί το αρχείο εξόδου με το επιθυμητό αποτέλεσμα. Το πρόγραμμά σας θα πρέπει να διαβάζει τα αρχεία *ταυτόχρονα* και να υπολογίζει και να γράφει το αποτέλεσμα, **χωρίς να φορτώσει όλα τα δεδομένα στη μνήμη**. Δηλαδή θα πρέπει να παράγει αποτελέσματα, καθώς διαβάζει τις γραμμές των αρχείων και πριν φτάσει στο τέλος τους. Συγκεκριμένα, για κάθε τίτλο στο title.basics.tsv πρέπει να παράγεται το αποτέλεσμα γι αυτό τον τίτλο πριν διαβάσουμε την επόμενη γραμμή.
Προσοχή: υπάρχουν τίτλοι στο title.basics.tsv που δεν εμφανίζονται στο title.akas.tsv και το ανάποδο.

Μέρος 2 (40%, implementation of merge-join with pipelining)
Υλοποιήστε το merge-join σαν iterator, ο οποίος μπορεί να χρησιμοποιηθεί για pipelining. Συγκεκριμένα, πρέπει να υλοποιήσετε μια συνάρτηση, η οποία θα υπολογίζει και θα επιστρέφει κάθε φορά το **επόμενο** αποτέλεσμα. Θεωρείστε ότι οι δύο είσοδοι του τελεστή είναι επίσης iterators και ότι κάθε φορά επιστρέφουν το επόμενο αποτέλεσμα. Θεωρείστε ότι οι δύο είσοδοι δίνουν τις πλειάδες τους σε μορφή tab delimeted text, όπου το πρώτο πεδίο είναι το πεδίο συνένωσης (join attribute) και ότι οι πλειάδες δίνονται ταξινομημένες ως προς το πεδίο αυτό. Ο τελεστής merge join παίρνει σαν είσοδο τους δύο τελεστές τα αποτελέσματά των οποίων συνενώνονται και τις θέσεις των άλλων πεδίων που επιλέγονται μαζί με το πεδίο συνένωσης.
Στα πλαίσια της εργασίας θα υλοποιήσετε μαζί με τον τελεστή mj που κάνει το merge-join και ένα τελεστή scan, ο οποίος διαβάζει και επιστρέφει μία-μία τις γραμμές από ένα αρχείο. Για παράδειγμα η mj(scan('title.basics.tsv'),scan('title.ratings.tsv'),(2),(1,2)) δημιουργεί τα εξής αποτελέσματα:

tt0000001	Carmencita	5.8	1467
tt0000002	Le clown et ses chiens	6.4	176
tt0000003	Pauvre Pierrot	6.6	1094
tt0000004	Un bon bock	6.5	105
…	…	…	…

γιατί ζητείται η συνένωση των title.basics και title.ratings με βάση το πρώτο τους πεδίο, ενώ εκτός από αυτό θέλουμε και το πεδίο 2 της title.basics (δηλ. το primaryTitle) και τα πεδία 1 και 2 της title.ratings (δηλ. τα averageRating και numVotes).

Αφού υλοποιήσετε τον τελεστή κάντε χρήση του για να δώσει στην έξοδό του τα εξής joins:
(α) mj(scan('title.basics.tsv'),scan('title.ratings.tsv'),(2),(1,2))
Δίνει στην έξοδο για κάθε τίτλο, τον identifier, το όνομα του τίτλου (primaryTitle) τη μέση βαθμολογία (averageRating) και τον αριθμό των ψήφων (numVotes).
(β) mj(scan('title.basics.tsv'),scan('title.principals.tsv'),(2),(2))
Δίνει στην έξοδο πλειάδες του τύπου (identifier τίτλου, όνομα τίτλου, identifier συντελεστή). Προσοχή: για κάθε τίτλο μπορεί να υπάρχουν πολλοί συντελεστές, άρα στην έξοδο μπορεί να έχουμε πολλές πλειάδες για κάθε τίτλο.
(γ) mj(mj(scan('title.basics.tsv'),scan('title.principals.tsv'),(2),(2)),scan('title.ratings.tsv'),(1,2),(1))
Δίνει στην έξοδο πλειάδες του τύπου (identifier τίτλου, όνομα τίτλου, identifier συντελεστή, βαθμολογία). Προσοχή: στην έξοδο μπορεί να έχουμε πολλές πλειάδες για κάθε τίτλο.

Υπόδειξη: Για την υλοποίηση μπορείτε να χρησιμοποιήσετε generators
https://en.wikipedia.org/wiki/Generator_(computer_programming)

Μέρος 3 (20%, γραπτό)
Έστω ότι θέλουμε να εμφανίσουμε για κάθε τίτλο που είναι επεισόδιο, το όνομα του επεισοδίου (primaryTitle), το όνομα της σειράς στην οποία ανήκει το επεισόδιο (primaryTitle), την περίοδο (seasonNumber) και τον αριθμό επεισοδίου (episodeNumber). Η πληροφορία βρίσκεται τους πίνακες title.basics και title.episode.
(α) Εκφράστε την ερώτηση σε σχεσιακή άλγεβρα.
(β) Με δεδομένο ότι και οι 2 πίνακες εισόδου είναι ταξινομημένοι με βάση το πεδίο του join (identifier των τίτλων), προτείνετε ένα κατάλληλο πλάνο εκτέλεσης για την ερώτηση. Αν ένα B+-tree είναι διαθέσιμο για τον πίνακα title.basics στο πρώτο πεδίο του πίνακα, μπορούμε να το εκμεταλλευτούμε;
(γ) Η έξοδος μπορεί να έχει διπλοεγγραφές; Αν ναι προτείνετε ένα τρόπο απαλοιφής τους.

Παραδοτέα: Κάντε turnin στο assignment1@mye041 τα προγράμματά σας και ένα PDF αρχείο το οποίο τεκμηριώνει τα προγράμματα στα μέρη 1 και 2, και περιέχει την απάντησή σας στο μέρος 3. Προσοχή: τα προγράμματά σας πρέπει να τρέχουν στα μηχανήματα του εργαστηρίου ΠΕΠ2, όπου και θα εξεταστείτε.