ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΥΕ041 - ΠΛΕ081: Διαχείριση Σύνθετων Δεδομένων (ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2019-20)

ΕΡΓΑΣΙΑ 1 – Αλγόριθμοι Αποτίμησης Ερωτήσεων (προθεσμία: 27 Μαρτίου 2020, 9μ.μ.)

Στόχος της εργασίας είναι η ανάπτυξη προγραμμάτων για αποτίμηση (evaluation) σύνθετων ερωτημάτων σε βάσεις δεδομένων. 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.
Πριν ξεκινήσετε την εργασία, **ταξινομήστε** τις γραμμές των αρχείων, ώστε οι πλειάδες σε κάθε αρχείο να είναι σε αύξουσα σειρά με βάση το πρώτο πεδίο (π.χ. tconst) και **γράψτε τα δεδομένα που προκύπτουν σε νέα αρχεία**. Από κει και πέρα θεωρήστε ότι οι πίνακες είναι τα ταξινομημένα αρχεία.

Για καθεμιά από τις παρακάτω 5 ερωτήσεις:
Α) Εκφράστε την ερώτηση με χρήση σχεσιακής άλγεβρας. Θεωρείστε ως όνομα κάθε πίνακα το όνομα του αντίστοιχου αρχείου χωρίς το tsv.gz, π.χ. title.akas, title.basics, κλπ.
Β) Γράψτε ένα πρόγραμμα το οποίο αποτιμά την ερώτηση.

Μέρος 1 (Ερωτήσεις με merging)

Τα προγράμματά σας για τις παρακάτω ερωτήσεις θα πρέπει να διαβάζουν τα αρχεία *ταυτόχρονα* και να υπολογίζουν και να γράφουν τα αποτελέσματα, **χωρίς να φορτώνουν όλα τα δεδομένα στη μνήμη**. Δηλαδή θα πρέπει τα αποτελέσματα να παράγονται, καθώς διαβάζονται οι γραμμές των αρχείων και πριν διαβαστούν όλα τα δεδομένα. Αυτό είναι δυνατόν γιατί τα αρχεία που διαβάζονται είναι ταξινομημένα με βάση το πεδίο tconst. Το κάθε πρόγραμμα **πρέπει να γράφει** τα αποτελέσματα σε ένα αρχείο εξόδου που θα έχει το όνομα της ερώτησης (π.χ. output1.1.txt).

Q1.1: Για κάθε τίτλο, ο οποίος έχει σκηνοθετηθεί από **πάνω από ένα σκηνοθέτη**, βρες το όνομα του τίτλου (πεδίο primaryTitle) και τα αναγνωριστικά των σκηνοθετών του (πεδίο directors)
Δείγμα αρχείου εξόδου:

primaryTitle	directors
Corbett and Courtney Before the Kinetograph	nm0005690,nm0374658
The Arrival of a Train	nm0525908,nm0525910
Italienischer Bauerntanz	nm1587194,nm0804434
Rough Sea at Dover	nm0010291,nm0666972
Leaving Jerusalem by Railway	nm0698645,nm0525908

Q1.2: Για κάθε τίτλο, ο οποίος είναι **το πρώτο επεισόδιο** μιας σειράς (episodeNumber=1), βρες το όνομα του τίτλου (πεδίο primaryTitle), το προσδιοριστικό της σειράς (πεδίο parentTconst) και τον αριθμό της περιόδου (seasonNumber)
Δείγμα αρχείου εξόδου:

primaryTitle	parentTconst	seasonNumber
The Fourth Anniversary Show	tt0046593	4
The Life of Henry V	tt4280482	1
Mordfall Oberhausen	tt0793361	1
Julius Caesar	tt4280486	1
Treffpunkt Bahnhof Zoo	tt0793361	2

Q1.3: Βρες τα ονόματα των τίτλων (πεδίο primaryTitle), οι οποίοι δεν έχουν rating (δηλαδή δεν εμφανίζονται στον πίνακα title.ratings)
Δείγμα αρχείου εξόδου:

primaryTitle
En classe
Le Coucher d'Yvette
Le planton du colonel
Une nuit agitée
Brittania

Μέρος 2 (Ερωτήσεις συνάθροισης)

Τα προγράμματα για τις παρακάτω ερωτήσεις πρέπει να τυπώνουν τα αποτελέσματα στην έξοδο (όχι σε αρχείο).

Q2.1: Αν οι μέσες βαθμολογίες (averageRating) των τίτλων χωρίζονται σε διαστήματα 0 έως 1, 1.1 έως 2, 2.1 έως 3,…, 9.1 έως 10, για κάθε διάστημα βρείτε τον αριθμό των τίτλων των οποίων η μέση βαθμολογία είναι μέσα σε αυτό το διάστημα. Γράψτε δύο εκδόσεις του προγράμματος (μία με sorting και μία με hashing) και συγκρίνετε τους χρόνους εκτέλεσής τους.
Δείγμα εξόδου:

```
0.1 - 1.0 :  971
1.1 - 2.0 :  4067
2.1 - 3.0 :  10831
3.1 - 4.0 :  26390
4.1 - 5.0 :  62814
```

Q2.2: Για κάθε startYear, βρείτε το μέσο όρο των βαθμολογιών των τίτλων που έχουν αυτό το startYear και έχουν βαθμολογηθεί. Για την παρακάτω ερώτηση, όπως και στις Q1.1-Q1.3 δεν θα πρέπει να διαβάσετε τα περιεχόμενα των αρχείων εισόδου στην μνήμη εξ ολοκλήρου, πριν αρχίσετε τους υπολογισμούς. Θα πρέπει να φτιάχνετε τα στατιστικά σταδιακά καθώς διαβάζετε τα αρχεία. Μέρος της ερώτησης αποτιμάται με merge-join.
Δείγμα εξόδου:

```
year: 1874 average rating: 7.0
year: 1878 average rating: 7.4
year: 1881 average rating: 5.4
year: 1883 average rating: 6.4
year: 1885 average rating: 5.4
```

Παραδοτέα: Κάντε turnin στο assignment1@mye041 τα προγράμματά σας και ένα PDF αρχείο το οποίο τεκμηριώνει τα προγράμματα και περιέχει τις ζητούμενες εκφράσεις σε σχεσιακή άλγεβρα
Προσοχή: τα προγράμματά σας πρέπει να τρέχουν στα μηχανήματα του εργαστηρίου ΠΕΠ2, όπου και θα εξεταστείτε.