Διαχείριση Σύνθετων Δεδομένων
1η Σειρά Ασκήσεων

Εαρινό 2020

Πασόη Σοφία ΑΜ:2798

Υπεύθυνος Καθηγητής: κ. Μαμουλής Νικόλαος

Ταξινόμηση Αρχείων:

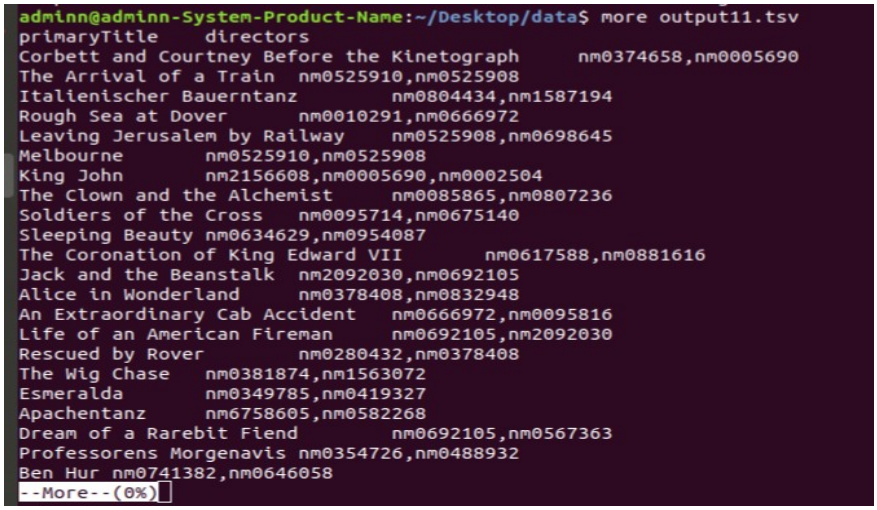
Αρχικά με τη συνάρτηση sortFile ταξινομούμε τα αρχεία με βάση το πρώτο πεδίο τους. Παίρνει ως όρισμα το όνομα του υπάρχον αρχείου και του ταξινομημένου που θέλουμε να δημιουργηθεί. Ανοίγει το υπάρχον αρχείο με δικαίωμα read και το ταξινομημένο με δικαίωμα write. Κρατάει σε 1 μεταβλητή τα στοιχεία της πρώτης σειράς, στη συνέχεια τη διαγράφει, δίοτι δε χρειάζεται ταξινόμηση. Όσο διαβάζει γραμμές, σπάει την κάθε γραμμή στο tab και ταξινομεί, με τη συνάρτηση sort. Στο αρχείο εξόδου γράφει τα στοιχεία της πρώτης γραμμής και έπειτα γράφει τις γραμμές(ταξινομημένες) ως αλφαριθμητικά.

Ερώτημα 1ο:

A: $\Pi_{primaryTitle,directors}(G_{count(title.crew.directors)>1}(\sigma_{title.basics.tconst=title.crew.tconst}(title.crew\ X\ title.basics)))$

B: Η συνάρτηση mergeSortQ11 παίρνει ως ορίσματα τα αρχεία title.basics_sorted.tsv και title.crew._sorted.tsv των οποίων τα πεδία tconst θέλουμε να συγκρίνουμε και το αρχείο στο οποίο θα γραφούν τα αποτελέσματα. Ανοίγει τα υπάρχοντα αρχεία με δικαίωμα read και το αρχείο εξόδου με δικαίωμα write. Διαβάζει την πρώτη γραμμή ταυτόχρονα και από τα 2 αρχεία και γράφει στο αρχείο εξόδου τα ονόματα των πεδίων, δηλαδή το primaryTitle και το directors από τα αρχεία title.basics_sorted.tsv και title.crew_sorted.tsv αντίστοιχα. Όσο υπάρχουν και στα 2 αρχεία γραμμές ελέγχουμε αν το πεδίο tconst είναι ίδιο και στα 2 πεδία. Εάν είναι τότε ελέγχει αν υπάρχουν παραπάνω του ενός σκηνοθέτες, και ναι τότε γράφει τα αναγνωριστικά τους στο αρχείο εξόδου. Σε διαφορετική περίπτωση προχωράει στο διάβασμα σειράς του αρχείου του οποίου το tconst είναι μικρότερο. Τέλος, κλείνει τα αρχεία.

Έξοδος: Δείγμα Αρχείου Εξόδου



Ερώτημα 2ο:

A: $\Pi_{primaryTitle,parenTtconst,seasonNumber}(\sigma_{title.basics.tconst=title.episodes.parentTconst\ and\ title.episodes.episodeNumber=1}$(title.basics X title.episodes))

B: Η συνάρτηση mergeSortQ12 παίρνει ως ορίσματα τα αρχεία title.basics_sorted.tsv και title.episode_sorted.tsv των οποίων τα πεδία tconst θέλουμε να συγκρίνουμε και το αρχείο στο οποίο θα γραφούν τα αποτελέσματα. Ανοίγει τα υπάρχοντα αρχεία με δικαίωμα read και το αρχείο εξόδου με δικαίωμα write. Διαβάζει την πρώτη γραμμή ταυτόχρονα και από τα 2 αρχεία και γράφει στο αρχείο εξόδου τα ονόματα των πεδίων, δηλαδή το primaryTitle, το parentTconst και το seasonNumber από τα αρχεία title.basics_sorted.tsv και title.episode_sorted.tsv αντίστοιχα. Όσο υπάρχουν και στα 2 αρχεία γραμμές ελέγχουμε αν το πεδίο tconst είναι ίδιο και στα 2 αρχεία. Εάν είναι τότε ελέγχει αν είναι το πρώτο επεισοδιο. Αν ικανοποιείται η συνθήκη τότε γράφει στο αρχείο εξόδου τις τιμές των πεδίων primaryTitle, parentTconst και seasonNumber και προχωρά το διάβασμα. Σε διαφορετική περίπτωση προχωράει στο διάβασμα σειράς του αρχείου του οποίου το tconst είναι μικρότερο. Τέλος, κλείνει τα αρχεία.

Έξοδος: Δείγμα Αρχείου Εξόδου

```
adminn@adminn-System-Product-Name:~/Desktop/data$ more output12.tsv
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
Margin for Error        tt1845727       1
The Land of Oz  tt0051312       2
Die Zeugin im grünen Rock       tt0793361       3
Saison  tt0793361       4
Flashing Spikes tt0054513       2
The Golden Horseshoe Revue      tt0046593       9
Number Six      tt0190181       3
In jeder Stadt...       tt0793361       5
The Horse Without a Head: The 100,000,000 Franc Train Robbery   tt0046593
10
Das Haus an der Stör    tt0793361       6
Tivoli  tt0242203       3
The Diary of a Nobody: The Domestic Jottings of a City Clerk    tt0497975
1
Rehe    tt0793361       7
Boy in the Smoke        tt0969640       1
Esther's Altar  tt0212698       3
--More--(0%)
```

<u>Ερώτημα 3ο:</u>

A: $A \leftarrow \Pi_{tconst}(\text{title.basics}) - \Pi_{tconst}(\text{title.ratings})$

$\Pi_{primaryTitle}(\sigma_{A.tconst=title.basics.tconst}(A \times \text{title.basics}))$

B: Η συνάρτηση mergeSortQ13 παίρνει ως ορίσματα τα αρχεία title.basics_sorted.tsv και title.ratings_sorted.tsv και το αρχείο στο οποίο θα γραφούν τα αποτελέσματα. Ανοίγει τα υπάρχοντα αρχεία με δικαίωμα read και το αρχείο εξόδου με δικαίωμα write. Διαβάζει την πρώτη γραμμή, την αγνοεί και προχωράει το διάβασμα. Γράφει στο αρχείο εξόδου τα όνομα του πεδίου, δηλαδή το primaryTitle το αρχείο title.basics_sorted.tsv. Όσο υπάρχουν και στα 2 αρχεία γραμμές ελέγχουμε με το πεδίο tconst ποια ονόματα τίτλων δεν έχουν ratings. Εάν είναι ίσα σημαίνει ότι υπάρχει το rating, οπότε και προχωράει το διάβασμα. Αν το tconst του title.basics.tsv είναι μικρότερο αυτού του title.ratings_sorted.tsv, τότε γράφει στο αρχείο εξόδου την τιμή του πεδίου primaryTitle και προχωρά το διάβασμα στο title.basics_sorted.tsv. Αν το title.ratings_sorted.tsv έχει τελειώσει ενώ το title.basics_sorted.tsv όχι, αυτό σημαίνει ότι υπάρχουν ταινίες οι οποίες δεν έχουν ratings, συνεπώς θα εγγραφούν και αυτές στο αρχείο εξόδου. Τέλος, κλείνει τα αρχεία.

Έξοδος: Δείγμα Αρχείου Εξόδου

```
adminn@adminn-System-Product-Name:~/Desktop/data$ more output13.tsv
primaryTitle
En classe
Le Coucher d'Yvette
Le planton du colonel
Une nuit agitée
Brittania
Le chemin de croix
La crèche à Bethléem
The Deserter
Dorotea
Déménagement à la cloche de bois
L'entrée à Jérusalem
Les farces de Jocko
Glasgow Fire Engine
Gran corrida de toros
Indian War Council
Le jardin des oliviers
Je vous y prrrrends!
Jésus devant Pilate
Llegada de un tren a la estación de ferrocarril del Norte, de Barcelona
London Express
Saida do Paquete Duque de Braganca
Visita de Doña María Cristina y Alfonso XIII a Barcelona
Waves and Spray
The Artist and the Flower Girl
Aspectos da Praia de Cascais
Battlefield
Boat Race
Bombardment of Mafeking
Le chiffonnier
Choque de dos transatlánticos
Courte échelle
Le crucifiement
Les dangers de l'alcoolisme
La descente de croix
Le déjeuner des enfants
```

Ερώτημα 4ο:

A: $A \leftarrow G_{COUNT\ tconst}(\sigma_{averageRating>=0.1\ and\ averageRating<=1.0}(title.ratings))$

$B \leftarrow G_{COUNT\ tconst}(\sigma_{averageRating>=1.1\ and\ averageRating<=2.0}(title.ratings))$

$C \leftarrow G_{COUNT\ tconst}(\sigma_{averageRating>=2.1\ and\ averageRating<=3.0}(title.ratings))$

$D \leftarrow G_{COUNT\ tconst}(\sigma_{averageRating>=3.1\ and\ averageRating<=4.0}(title.ratings))$

$E \leftarrow G_{COUNT\ tconst}(\sigma_{averageRating>=4.1\ and\ averageRating<=5.0}(title.ratings))$

$F \leftarrow G_{COUNT\ tconst}(\sigma_{averageRating>=5.1\ and\ averageRating<=6.0}(title.ratings))$

$G \leftarrow G_{COUNT\ tconst}(\sigma_{averageRating>=6.1\ and\ averageRating<=7.0}(title.ratings))$

$H \leftarrow G_{COUNT\ tconst}(\sigma_{averageRating>=7.1\ and\ averageRating<=8.0}(title.ratings))$

$I \leftarrow G_{COUNT\ tconst}(\sigma_{averageRating>=8.1\ and\ averageRating<=9.0}(title.ratings))$

$J \leftarrow G_{COUNT\ tconst}(\sigma_{averageRating>=9.1\ and\ averageRating<=10.0}(title.ratings))$

$A \cup B \cup C \cup D \cup E \cup F \cup G \cup H \cup I \cup J$

B:

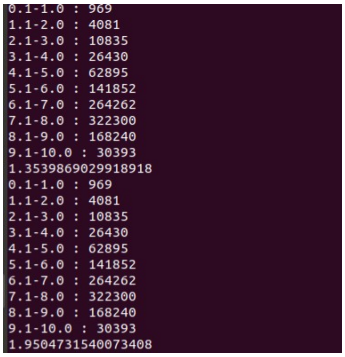
1. Hashing: Η h(x) είναι η συνάρτηση κατακερματισμού. Η h(x) υπολογίζει σε ποιο διάστημα ανήκει η βαθμολογία και μας επιστρέφει ένα αναγνωριστικό. Η συνάρτηση hashingQ21 ορίζει 10 μετρητές αρχικοποιημένους στο 0, ανοίγει το αρχείο title.ratings_sorted.tsv με δικαίωμα read και διαβάζει το αρχείο πετώντας την πρώτη γραμμή. Για κάθε γραμμή παίρνει την τιμή του πεδίου averageRating, την εκχωρεί σε 1 μεταβλητή με την οποία καλείται η h(x) και ενημερώνεται ο κατάλληλος μετρητής. Τυπώνει στην έξοδο τα αποτελέσματα σύμφωνα με τη μορφή του δοθέντος δείγματος εξόδου. Τέλος, κλείνει το αρχείο.
2. Sorting: Η συνάρτηση sortingQ21 ορίζει 10 μετρητές αρχικοποιημένους στο 0, ανοίγει το αρχείο title.ratings_sorted.tsv με δικαίωμα read και διαβάζει το αρχείο πετώντας την πρώτη γραμμή. Παίρνει όλες τις αποτιμήσεις του averageRating και τις ταξινομεί. Χρησιμοποιεί 2 τοπικές μεταβλητές τις pos και maxRating που είναι αρχικοποιημένες στο 0 και το 1 αντίστοιχα. Εκμεταλλευόμενη την ταξινόμηση, επαναληπτικά ελέγχει σε ποιο διάστημα ανήκει το avgRating. Δηλαδή, μπαίνοντας στην επανάληψη αν το avgRating<maxRating αυξάνεται ο αντίστοιχος counter. Μόλις το avgRating>maxRating θα ενημερωθούν οι τιμές των pos και maxRating. Η διαδικασία επαναλαμβάνεται εώς ότου τελειώσει το αρχείο. Τυπώνει στην έξοδο τα αποτελέσματα σύμφωνα με τη μορφή του δοθέντος δείγματος εξόδου. Τέλος, κλείνει το αρχείο.

Σύγκριση: Για τον υπολογισμό του χρόνου χρησιμοποιήθηκε η βιβλιοθήκη της Python. Πριν την κλήση των συναρτήσεων κρατήθηκε ο χρόνος έναρξης και μετά ο χρόνος λήξης. Τέλος ο χρόνος εκτέλεσης υπολογίζεται ως: χρόνος λήξης- χρόνος έναρξης.

Hashing Time: 1.3539869029918918, Sorting Time: 1.9504731540073408

Το Sorting Time είναι μεγαλύτερο του Hashing Time. Το αποτέλεσμα ήταν αναμενόμενο διότι γνωρίζουμε πως ο κατακερματισμός παρόλο που δεν παρέχει εγγύησεις, όταν τα δεδομένα δεν είναι ταξινομημένα είναι ο καλύτερος τρόπος.

Έξοδος:



Ερώτημα 5ο:

A: startYear $G_{AVERAGE\ averageRating}(\sigma_{title.basics.tconst=titlie.ratings.tconst}(title.basics\ X\ title.ratings))$

B: Η συνάρτηση mergeSortQ13 παίρνει ως ορίσματα τα αρχεία title.basics_sorted.tsv και title.ratings_sorted.tsv. Ανοίγει τα υπάρχοντα αρχεία με δικαίωμα read, διαβάζει την πρώτη γραμμή, την αγνοεί και προχωράει το διάβασμα. Φτιάχνει ένα λεξικό το οποίο για κάθε startYear θα κρατάει το άθροισμα των μέσων βαθμολογιών και το πλήθος των ταινιών. Όσο υπάρχουν και στα 2 αρχεία γραμμές ελέγχουμε αν τα πεδία tconst είναι ίσα. Εάν είναι ίσα σημαίνει ότι υπάρχει το avgRating για αυτό το startYear. Βάζει σε 1 μεταβλητή την αποτίμηση του πεδίου startYear, στη συνέχεια ορίζει ένα default key και σε 1 δεύτερη την αποτίμηση του avgRating. Αν η τιμή του key είναι ίση με τη default, τότε ξέρει ότι βρήκε την πρώτη ταινία για το συγκεκριμένο startYear. Ενημερώνει το λεξικό βάζοντας στο key το avgRating που διάβασε και το πλήθος ίσο με ένα. Αλλιώς προσθέτει το avgRating που διάβασε στο ήδη υπάρχον άθροισμα, αυξάνει το πλήθος κατά 1 και ενημερώνει το λεξικό.Αν το tconst του title.basics_sorted.tsv είναι μικρότερο αυτού του title.ratings_sorted.tsv, τότε το διάβασμα στο title.basics_sorted.tsv, αλλιώς προχωρά το διάβασμα του title.ratings_sorted.tsv. Για να τυπώσει τα στατιστικά, αρχικά βάζει σε μια λίστα τα keys του λεξικού και την ταξινομεί. Επαναλαπτικά, βάζει σε μεταβλητή τη χρονολογία και τραβάει με αυτήτν από το λεξικό τις τιμές του key της. Υπολογίζει τα στατιστικά και τυπώνει στην έξοδο τα αποτελέσματα στη μορφή του δοθέντος δείγματος. Τέλος, κλείνει τα αρχεία.

Έξοδος: Δείγμα Αρχείου 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
year: 1887 average rating: 4.928888888888889
year: 1888 average rating: 6.24
year: 1889 average rating: 5.4
year: 1890 average rating: 5.2
year: 1891 average rating: 4.909999999999999
year: 1892 average rating: 5.3
year: 1893 average rating: 4.833333333333333
year: 1894 average rating: 4.928865979381443
year: 1895 average rating: 4.912068965517243
year: 1896 average rating: 4.853597650513951
```

- Τρέχουμε τα αρχεία πληκτρολογώντας στο τερματικό την εντολή: python3 "filename".
- Μπορούμε να δούμε το περιχόμενο των αρχείων στο τερματικό πληκτρολογώντας την εντολή: more "filename".
- Κάθε αρχείο αντιστοιχεί και σε ένα ερώτημα. Υπάρχει και ένα έκτο το οποίο κάνει την αρχική ταξινόμηση.