Project - Βάσεις Δεδομένων

Προθεσμία: 20/6/2022

Σκοπός

Στο project θα ασχοληθούμε με την ανάλυση και οπτικοποίηση των δεδομένων της βάσης Movielens (https://movielens.org) την οποία και εξετάσαμε στα πλαίσια των προηγούμενων εργασιών. Η συγκεκριμένη βάση δεδομένων περιέχει πληροφορίες για ταινίες, τους συντελεστές τους, και τις αξιολογήσεις των ταινιών από χρήστες.

Ομάδες

Για την ολοκλήρωση του project θα πρέπει να σχηματίσετε ομάδες των **2 ή 3 ατόμων** και όχι παραπάνω ή λιγότερα άτομα ανά ομάδα.

Δεδομένα

Μπορείτε να φορτώσετε τα δεδομένα της άσκησης κάνοντας χρήση του Backup που βρίσκεται στον ακόλουθο σύνδεσμο ή να χρησιμοποιήσετε την βάση που φτιάξατε στις προηγούμενες ασκήσεις.

Για όσους δημιουργήσουν την βάση από τον σύνδεσμο, το restore της βάσης γίνεται στα ακόλουθα βήματα:

- Επιλογή της βάσης (database) στην οποία θα δημιουργηθουν οι νέοι πίνακες.
- Δεξί click και επιλογή restore.
- Στο πεδίο format "Custom or tar" στο πεδίο Filename το όνομα του αρχείου που κατεβάσατε.

Οπτικοποίηση Στατιστικών

Σε αυτό το μέρος θα πρέπει να υπολογίσετε και να οπτικοποιήσετε τα παρακάτω στατιστικά με χρήση SQL και Python (μέσω σύνδεσης στην βάση σας).

Στατιστικά ταινιών

1. Αριθμός ταινιών ανά έτος.
2. Αριθμός ταινιών ανά είδος (genre).
3. Αριθμός ταινιών ανά είδος (genre) και ανά έτος.
4. Το υψηλότερο budget ταινίας ανά έτος (δεν μας ενδιαφέρει για ποια ταινία).

Στατιστικά ηθοποιών

5. Για τον αγαπημένο σας ηθοποιό, το σύνολο των εσόδων (revenue) για τις ταινίες στις οποίες έχει συμμετάσχει **ανά έτος**.

Βαθμολογίες Χρήστη

6. Μέση βαθμολογία (rating) ανά χρήστη (scatter plot).

7. Αριθμός από βαθμολογίες ανά χρήστη (scatter plot).
8. Scatter plot το οποίο θα έχει ένα σημείο για κάθε χρήστη που στον **x άξονα** φαίνεται ο αριθμός των αξιολογήσεων του χρήστη και στον **y άξονα** η μέση βαθμολογία του.
9. Μέση βαθμολογία (rating) ανά είδος ταινίας.

Σημείωση: Επειδή η βάση μας δεν περιέχει τον πλήρη κατάλογο ταινιών, θεωρούμε ότι κάποια από τα στατιστικά δεν είναι ακριβή.

Παραδοτέο

Θα πρέπει να παραδώσετε τα ακόλουθα:

1. Ένα αρχείο sql που να περιέχει τις εντολές που χρησιμοποιήσατε για να υπολογίσετε όλα τα παραπάνω.
2. Το αρχείο Python που πραγματοποιεί την σύνδεση με την βάση (με κωδικούς examiner) και οπτικοποιεί τα δεδομένα.
3. Ένα pdf το οποίο περιέχει τις οπτικοποιήσεις μαζί με μία σύντομη επεξήγηση της πληροφορίας που απεικονίζεται. Σε περίπτωση που θεωρείτε ότι για κάποιο ερώτημα υπάρχει κάποια εναλλακτική απεικόνιση της πληροφορίας, μπορείτε να την προσθέσετε μαζί με αντίστοιχη επεξήγηση.

Χρήσιμα links:

Python

- Azure Connection: https://docs.microsoft.com/en-us/azure/postgresql/connect-python

- Matplotlib: https://matplotlib.org/

- Transpose Matrix: https://numpy.org/doc/stable/reference/generated/numpy.transpose.html

Postgres

- Extract from Date: https://www.postgresql.org/docs/current/functions-datetime.html

- Order by: https://www.postgresql.org/docs/current/queries-order.html

- Aggregate: https://www.postgresql.org/docs/current/tutorial-agg.html

- Changing a Column's Data Type: https://www.postgresql.org/docs/current/ddl-alter.html#id-1.5.4.8.10

Τελικά Παραδοτέα

- Δημιουργήστε ένα .txt αρχείο στο οποίο θα αναγράφονται το endpoint του Azure instance σας (Server name στο Overview tab του Azure), το όνομα της βάσης σας και το username και το password ενός χρήστη με read-only δικαιώματα, ώστε να μπορούμε να δούμε τους πίνακες της βάσης σας. Το .txt αρχείο θα πρέπει να έχει την παρακάτω μορφή:

  Endpoint: <name_of_the_endpoint>

  Username: <username>

Password: <password>

Database: <name_of_the_database>

- Βάλτε σε ένα φάκελο το txt αρχείο καθώς και τα παραδοτέα που σχετίζονται με sql ερωτήματα, τον κώδικα σε python, και τις απεικονίσεις μαζί με τις επεξηγήσεις τους. Το όνομα του φακέλου πρέπει να αποτελείται από τους αριθμούς μητρώου σας χωρισμένους με παύλα, δηλαδή *αριθμός_μητρώου_1-αριθμός_μητρώου_2-αριθμός_μητρώου_3*. Δημιουργήστε ένα .zip αρχείο αυτού του φακέλου, το οποίο θα έχει το ίδιο όνομα με τον φάκελο.
- Κάντε υποβολή το .zip αρχείο στο eclass στην ενότητα *Εργασίες / Project*.