Project - Βάσεις Δεδομένων

Προθεσμία: 14/6/2021

Σκοπός

Στο project θα ασχοληθούμε με το στήσιμο μιας βάσης δεδομένων αποτελούμενη από δεδομένα ταινιών καθώς και με μία μικρή ανάλυση και οπτικοποίηση πάνω στα δεδομένα.

Ομάδες

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

Μέρος Α

Δεδομένα:

Τα δεδομένα μπορείτε να τα βρείτε εδώ. Σύμφωνα με τα δεδομένα αυτά θα πρέπει να δημιουργήσετε το σχήμα της βάσης, τις συσχετίσεις μεταξύ των πινάκων καθώς και να εισάγετε τα δεδομένα στους πίνακες θέτοντας τους σωστούς (κατά τη γνώμη σας) τύπους δεδομένων σε όλα τα attributes.

Προεπεξεργασία δεδομένων:

Για τις ανάγκες της εργασίας θα πρέπει να διαγράψετε όλα τα διπλότυπα από τους πίνακες (εκτός του “ratings”) καθώς και να διαγράψετε δεδομένα ταινιών οι οποίες δεν υπάρχουν στον πίνακα “movies_metadata” αλλά υπάρχουν σε κάποιον από τους υπόλοιπους πίνακες.

Η προεπεξεργασία των δεδομένων πρέπει να γίνει με αυτοματοποιημένο τρόπο με όποια γλώσσα προγραμματισμού επιθυμείτε.

Παραδοτέο μέρος Α:

Για το μέρος Α θα πρέπει να παραδώσετε μέσα σε ένα φάκελο με όνομα “partA” ένα αρχείο sql που να περιέχει τις εντολές δημιουργίας των πινάκων και εντολές δημιουργίας κλειδιών και περιορισμών. Επίσης, ένα αρχείο που να περιέχει τις εντολές προεπεξεργασίας των δεδομένων, με ό,τι γλώσσα ή σύστημα επιλέξατε. Επίσης, γράψτε ένα **σύντομο** report (.pdf αρχείο) για τα βήματα που ακολουθήσατε κατά την επεξεργασία των δεδομένων σας ώστε να αποθηκευτούν στη βάση. Τέλος, θα πρέπει να παραδώσετε το ER διάγραμμα σε αρχείο εικόνας.

Bonus:

Προσθέστε δεδομένα από μια εξωτερική πηγή στη βάση σας. Αυτά θα μπορούσαν να είναι για παράδειγμα φωτογραφίες ταινιών, περιλήψεις, κλπ.

Μέρος Β

Σε αυτό το μέρος θα πρέπει να υπολογίσετε και να οπτικοποιήσετε τα παρακάτω στατιστικά με χρήση SQL και Python (μέσω σύνδεσης[1] στην βάση σας).

- Αριθμός ταινιών ανά χρόνο
- Αριθμός ταινιών ανά είδος(genre)
- Αριθμός ταινιών ανά είδος(genre) και ανά χρόνο
- Μέση βαθμολογία (rating) ανά είδος (ταινίας)
- Αριθμός από ratings ανά χρήστη
- Μέση βαθμολογία (rating) ανά χρήστη

Τέλος δημιουργήστε ένα view table και αποθηκεύστε για κάθε χρήστη τον αριθμό των ratings που έχει κάνει καθώς και τη μέση βαθμολογία που έχει βάλει και οπτικοποιήστε το με χρήση scatter plot. Παίρνουμε κάποιο insight από αυτή τη σχέση?

Παραδοτέο μέρος Β:

Για το μέρος Β θα πρέπει να παραδώσετε μέσα σε ένα φάκελο με όνομα “partB” ένα αρχείο sql που να περιέχει τις εντολές που χρησιμοποιήσατε για να υπολογίσετε όλα τα παραπάνω καθώς και την εντολή δημιουργίας του view_table. Μέσα στον φάκελο βάλτε επίσης το αρχείο Python που πραγματοποιεί την σύνδεση με την βάση σας και οπτικοποιεί τα δεδομένα. Προσθέστε επίσης και αρχεία εικόνων με τις οπτικοποιήσεις μέσα στο φάκελο.

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

[1] https://docs.microsoft.com/en-us/azure/postgresql/connect-python

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 τους δύο φακέλους για τα μέρη Α και Β σε ένα φάκελο. 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*.