1η Εργασία: MovieLens Tables

Προθεσμία: 8/11/2024

Σκοπός:

Σε αυτή την εργασία θα δημιουργήσουμε την βάση δεδομένων ταινιών *Movielens* (https://movielens.org). Η συγκεκριμένη βάση δεδομένων περιέχει πληροφορίες για ταινίες, τους συντελεστές τους, και τις αξιολογήσεις τους. Για να ορίσουμε το σχήμα της βάσης θα βασιστούμε στους τύπους των δεδομένων εισόδου που βρίσκονται στα αρχεία csv. Θα δημιουργήσουμε τους πίνακες χρησιμοποιώντας SQL και θα εισάγουμε δεδομένα σε αυτούς με την εντολή `\copy`. Επίσης, θα δημιουργήσουμε περιορισμούς ξένου κλειδιού για *αναφορική ακεραιότητα* (*referential integrity*).

Περιγραφή Δεδομένων:

Τα δεδομένα της άσκησης βρίσκονται διαθέσιμα στον φάκελο της εργασίας στο αρχείο dataset.zip.

Η αρχική μορφή των δεδομένων μπορεί να βρεθεί στον ακόλουθο σύνδεσμο:
https://www.kaggle.com/rounakbanik/the-movies-dataset?select=keywords.csv

Στα δεδομένα που θα δείτε, έχει γίνει μία προεπεξεργασία των αρχικών csv αρχείων με σκοπό την απαλοιφή των JSon κελιών. Έτσι, π.χ., το αρχείο movies_metadata έχει χωριστεί σε περισσότερα του ενός csv αρχεία→ movie, movie_collection, collection, κτλ.
Τα csv αρχεία που σας δίνονται είναι τα ακόλουθα:

- movie.csv: Το αρχείο περιέχει πληροφορίες για διάφορες ταινίες, όπως π.χ. το αναγνωριστικό id μιας ταινίας, τον τίτλο της, το κόστος δημιουργίας της, τα κέρδη της και την περιγραφή της.
- genre.csv: Το αρχείο αυτό περιέχει τις διάφορες κατηγορίες ταινιών (π.χ. περιπέτεια, τρόμου, κωμωδία). Περιέχει ένα αναγνωριστικό id για κάθε είδος και το όνομα του είδους.
- productioncompany.csv: Το αρχείο αυτό περιέχει τις διάφορες εταιρίες παραγωγής. Περιέχει ένα αναγνωριστικό id για κάθε εταιρία παραγωγής και το όνομά της.
- collection.csv: Αναφέρεται σε μία συλλογή από ταινίες (π.χ. τριλογίες). Περιέχει ένα αναγνωριστικό id για κάθε συλλογή και το όνομά της.
- keyword.csv: Οι λέξεις κλειδιά που περιγράφουν μία κινηματογραφική παραγωγή.

- movie_cast: Περιέχει πληροφορίες για το cast της ταινίας (τους χαρακτήρες της, και τους ηθοποιούς που παίζουν τους χαρακτήρες). Το πεδίο cid προσδιορίζει μοναδικά κάθε επιλογή που έχει γίνει στο casting.
- movie_crew: Περιέχει πληροφορίες το crew μιας ταινίας (σκηνοθέτης, σεναριογράφος, κτλ.). Το πεδίο cid προσδιορίζει μοναδικά κάθε επιλογή που έχει γίνει για το crew μιας ταινίας.
- belongsTocollection.csv: Συσχετίζει μία ταινία με την συλλογή στην οποία ανήκει. Περιέχει τα πεδία movie_id, collection_id.
- hasGenre.csv: Συσχετίζει μία ταινία με τα είδη στα οποία ανήκει. Περιέχει τα πεδία movie_id, genre_id.
  haskeyword.csv: Η συσχέτιση μιας ταινίας με τις λέξεις κλειδιά που την περιγράφουν.
- hasproductioncompany.csv: Συσχετίζει μία ταινία με τις εταιρείες παραγωγής της. Περιέχει τα πεδία movie_id, pc_id.
- ratings.csv: Συσχετίζει έναν χρήστη με μία ταινία δίνοντας την αντίστοιχη βαθμολογία του χρήστη για την ταινία. Περιέχει τα πεδία user_id, movie_id, rating.

Τι θα φτιάξουμε:

- Τη βάση δεδομένων MovieLens σε ένα *Postgres instance*.
- Η βάση αυτή θα πρέπει να περιέχει πίνακες για τους οποίους θα ισχύουν τα εξής:
  a. Να δημιουργηθούν οι πίνακες με τους αντίστοιχους περιορισμούς **πρωτεύοντος κλειδιού**.
  b. Να εισαχθούν σε κάθε πίνακα τα αντίστοιχα δεδομένα.
  c. Να δημιουργηθούν οι **περιορισμοί πρωτεύοντος κλειδιού**.
     - Για τους πίνακες Movie, Genre, Collection, Productioncompany, movie_cast, movie_crew, keyword
  d. Να δημιουργηθούν οι **περιορισμοί ξένου κλειδιού** για τους πίνακες:
     - hasKeyword, movie_cast, movie_crew, belongsTocollection, hasGenre, hasProductioncompany

Απαραίτητα εργαλεία:

- Postgres Database
- Postgres psql client / pgAdmin

Οδηγίες:

- Το πρώτο γράμμα του ονόματος κάθε πίνακα να ξεκινάει με κεφαλαίο(π.χ. Calendar κτλ.).
- Τα ονόματα των πεδίων των πινάκων (attributes) να ξεκινάνε με μικρό (π.χ. id κτλ.).

- Τοποθετήστε τις εντολές ***`create table`*** σε ένα αρχείο. Για παράδειγμα create_tables.sql.
- Τοποθετήστε όλες τις εντολές **`alter table`** σε ένα αρχείο, alter_tables.sql.

Συμβουλές για την υλοποίηση:

- Συνδεθείτε στη βάση σας στο χρησιμοποιώντας το PgAdmin για να δημιουργήσετε τους πίνακες που ζητά η άσκηση.
- Επειδή μερικές εντολές create table έχουν πολλά πεδία, μπορείτε να τρέξετε το python πρόγραμμα gen_ddl_python3.py (θα το βρείτε στα έγγραφα του μαθήματος στο eclass) αν έχετε python 3 στον υπολογιστή σας ή το gen_ddl_python2.py (θα το βρείτε στα έγγραφα του μαθήματος στο eclass) αν έχετε python 2, το οποίο παίρνει ως παράμετρο το .csv αρχείο των δεδομένων, π.χ. ratings.csv για τον πίνακα Ratings, και παράγει ένα αρχείο .sql με την εντολή create table για τον αντίστοιχο πίνακα. **Ελέγξτε την παραγόμενη εντολή και προσθέστε τους περιορισμούς για πρωτεύοντα κλειδιά**.
- Χρησιμοποιήστε την εντολή **`\i <filename>`** στην psql για να εκτελέσετε τον κώδικα SQL που έχετε αποθηκεύσει σε ένα αρχείο. Για παράδειγμα **`\i create_tables.sql.`** Εναλλακτικά, στο pgAdmin επιλέξτε τη βάση, πατήστε το query tool (κεραυνός πάνω αριστερά) και τρέξτε ένα sql script ως εξής: πατήστε το “Open file” (εικονίδιο φακέλου πάνω αριστερά στο query tool), επιλέξτε το sql script από τον υπολογιστή σας και πατήστε τον “κεραυνό” στη μπάρα του query tool.
- Χρησιμοποιήστε την εντολή **`copy`** στην psql ή τη λειτουργία **`Import/Export`** στο pgAdmin για να εισάγετε τα δεδομένα.
    - Λάβετε υπόψη σας ότι η πρώτη γραμμή στα .csv αρχεία είναι ο header. Δε θέλουμε να εισάγουμε τον header στον πίνακα. Θέστε την κατάλληλη παράμετρο είτε στο **`\copy`** είτε στην λειτουργία **`Import/Export`** ώστε να προσπεράσετε αυτή τη γραμμή. Παράδειγμα εντολής \copy:
      `\copy movie FROM 'movie.csv' DELIMITER ',' QUOTE '"' ESCAPE '"' CSV HEADER;`
    - Επίσης κατά την εισαγωγή των δεδομένων να χρησιμοποιηθούν οι ακόλουθοι παράμετροι

delimiter	,
quote	"
escape	"

- Πριν εκτελέσετε την εντολή **`copy`**, τρέξτε την εντολή **`set client_encoding to 'utf8';`** στην psql για να αποφύγετε προβλήματα με την κωδικοποίηση των χαρακτήρων.
- Προσθέστε τους περιορισμούς **ξένου κλειδιού μετά την εισαγωγή των δεδομένων στους πίνακες**.

Χρήσιμα links:

Εντολή create table:
https://www.postgresql.org/docs/9.6/sql-createtable.html

Εντολή copy:
https://www.postgresql.org/docs/9.6/sql-copy.html

Εντολή alter table:
https://www.postgresql.org/docs/9.6/sql-altertable.html

Postgres meta commands, όπως η \copy:
https://www.postgresql.org/docs/9.2/app-psql.html

pgAdmin query tool:
https://www.pgadmin.org/docs/pgadmin4/dev/query_tool.html

pgAdmin import:
https://www.pgadmin.org/docs/pgadmin4/dev/import_export_data.html

Παραδοτέα:

- Βάλτε σε ένα φάκελο
    a. τα **.sql αρχεία**,
    b. τον **python/java** κώδικα για την επεξεργασία του αρχείου keywords.csv,
    c. καθώς και μία **συνοπτική αναφορά** (~ 10 γραμμές) για το τι κάνατε σε κάθε βήμα της άσκησης.
- Ανεβάστε το .zip αρχείο στο eclass στην ενότητα *Εργασίες / 1η Εργασία*.