ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2022-2023

ΚΕΧΡΙΩΤΗ ΕΛΕΝΗ – 3210078

ΤΡΙΑΝΤΑΦΥΛΛΟΣ ΕΦΡΑΙΜ ΚΙΟΣΣΕΣ – 3210079

2η Εργασία

Κώδικας Python και προεπεξεργασία του keywords.csv

Αρχικά δημιουργήσαμε των κώδικα σε python “db.py” ο οποίος κάνει import τις βιβλιοθήκες ast και csv για την επεξεργασία του αρχείου keywords.csv.

- Η μεταβλητή writer1 χρησιμοποιείται για εγγραφή στο αρχείο keywords.csv
- Η μεταβλητή writer2 χρησιμοποιείται για εγγραφή στο αρχείο hasKeywords.csv

Ανοίξαμε το αρχείο keywords.csv με κωδικοποίηση utf-8 αγνοώντας τα λάθη που μπορεί να πετάξει, όπως είναι η αποδικοποίηση χαρακτήρων πχ ‘ ” ’.
Κάνουμε χρήση της λίστας seen για την αποφυγή διπλών στοιχείων και της λίστας rows για την αποφυγή ολόκληρων διπλών γραμμών.

-Εξαγωγή JSON
Σε μία for loop για κάθε γραμμή που διαβάζει από το csv αρχείο στην μεταβλητή jsonString αποθηκεύουμε το στοιχείο της θέσης 1 από την λίστα row, στην μεταβλητή data κάνουμε την χρήση της βιβλιοθήκης ast με την εντολή ast.literal_eval όπου μετατρέπει το jsonString σε μια λίστα αποτελούμενη από dictionaries.

-Αφαίρεση Διπλοτύπων και εγγραφή στα csv
Σε ένα δεύτερο for loop για κάθε ένα dictionary της λίστας data ελέγχουμε αν υπάρχει στην λίστα seen, αν δεν υπάρχει το προσθέτουμε γιατί είναι ένα καινούριο dictionary. Στην μεταβλητή r εξάγουμε τα στοιχεία (id, name) τα οποία τα γράφουμε στο αρχείο keywords.csv με την χρήση του writer1. Και στο αρχείο hasKeywords.csv με την χρήση του writer2 γράφουμε τα στοιχεία movie_id και keyword_id.

Διαφορετικά αν υπάρχει στην λίστα seen προχωράει στο επόμενο dictionary.

Δημιουργία Πινάκων και Κλειδιά

Μετά από αυτή την επεξεργασία εγκαταστήσαμε το extension SQL Server Import και δημιουργήσαμε τους πίνακες, κάνοντας import τα αρχεία από το σύνδεσμο στο drive και αυτά που φτιάξαμε, στην βάση δεδομένων με τις κατάλληλες αλλαγές στους τύπους δεδομένων όπου χρειαζόταν.

Τέλος δημιουργήσαμε τα primary και foreign keys για τους πίνακες με ένα query με όνομα alter_tables.sql