



Βάσεις Δεδομένων - Εργασία 2

Τμήμα Πληροφορικής

Εαρινό Εξάμηνο 2021-2022

Νικόλαος Δούρος (3200043)

Αθανάσιος Τριφώνης (3200298)

ΑΝΑΦΟΡΑ

Περίληψη

Στην παρούσα εργασία κατασκευάσαμε την βάση δεδομένων ταινιών MovieLens η οποία περιλαμβάνει τις πληροφορίες για ταινίες, τις αξιολογήσεις τους και τους συντελεστές τους. Δημιουργήσαμε τους πίνακες χρησιμοποιώντας τα δεδομένα των csv αρχείων που παρέχονται για την εργασία και προσθέσαμε περιορισμούς πρωτεύοντων και ξένων κλειδιών.

Κατασκευή Πινάκων

Για την κατασκευή των πινάκων χρησιμοποιήσαμε το πρόγραμμα `gen_ddl_python3.py` δίνοντας ως όρισμα κάθε φορά το όνομα του αρχείου csv για τον πίνακα που θέλαμε να κατασκευάσουμε. Το πρόγραμμα παρήγαγε ένα αρχείο sql με το όνομα του πίνακα που θα δημιουργηθεί, τις στήλες που θα περιλαμβάνει καθώς και τους τύπους των δεδομένων τους. Τα αρχεία sql τα φορτώσαμε με το query tool στο pgAdmin και τα εκτελέσαμε.

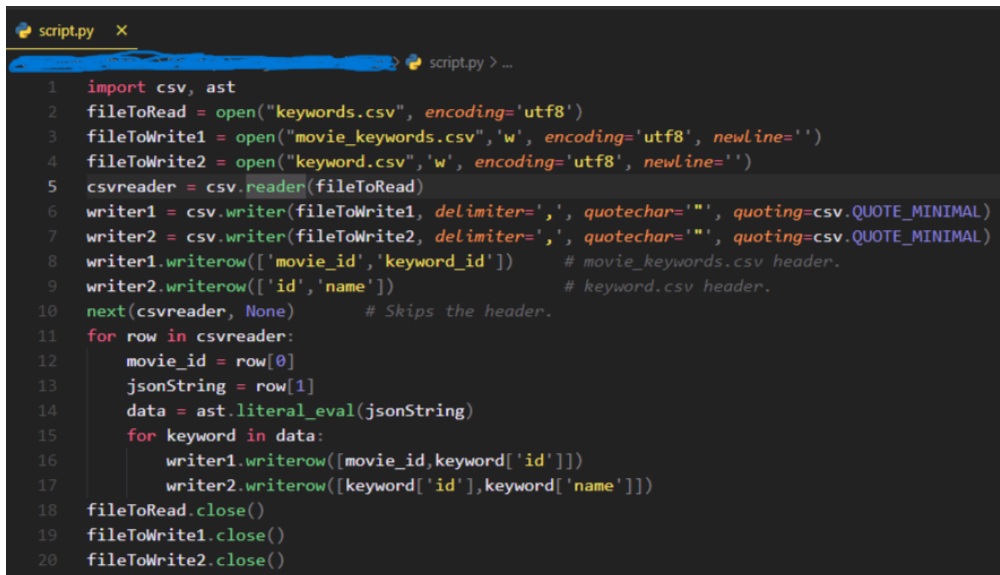
Εισαγωγή δεδομένων

Τα δεδομένα που υπάρχουν στα csv αρχεία τα κάναμε import μέσω του pgAdmin, δίνοντας τις παραμέτρους για delimiter, quote και escape καθώς και της διευκρίνησης πως υπάρχουν headers εντός του csv.

Διάβασμα και εγγραφή csv, διάβασμα json

Για το αρχείο keywords.csv το οποίο είχε στο εσωτερικό του στην στήλη keywords εμφωλευμένες πληροφορίες στη μορφή JSON συμβολοσειρών κατασκευάσαμε ένα πρόγραμμα στην python με όνομα `script.py`. Αυτό το πρόγραμμα αρχικά κάνει εισαγωγή την βιβλιοθήκη csv και ast. Έπειτα ανοίγουμε το αρχείο που θέλουμε να διαβάσουμε keywords.csv με παράμετρο `encoding = 'utf8'` και δύο επιπλέον αρχεία στα οποία θέλουμε να γράψουμε, `movie_keywords.csv` και `keyword.csv`, ορίζοντας τις παραμέτρους `'w'`, `encoding = 'utf8'` και `newline = ''`. Στην συνέχεια καλούμε την συνάρτηση `reader` της csv βιβλιοθήκης με όρισμα το αρχείο που θα διαβάσουμε, καθώς και την συνάρτηση `writer` της ίδιας βιβλιοθήκης δύο φορές, μία για κάθε αρχείο που θέλουμε να γράψουμε. Στα αρχεία που θέλουμε να γράψουμε καλούμε την μέθοδο `writerow` με όρισμα μία λίστα η οποία έχει τα ονόματα των στηλών για το κάθε αρχείο. Για να προσπεράσουμε το header του αρχείου που διαβάζουμε καλούμε την συνάρτηση `next` και ύστερα τρέχουμε όλες τις γραμμές

του αρχείου με μία for, κρατάμε σε μεταβλητές το movie_id και το jsonString από το πρώτο και το δεύτερο στοιχείο της γραμμής αντίστοιχα, καλούμε την συνάρτηση literal_eval της ast με όρισμα την μεταβλητή jsonString και τέλος χρησιμοποιούμε μια ακόμα for για να τρέξουμε κάθε δεδομένο της λίστας που παράγεται σε κάθε γραμμή από το jsonString που καλέσαμε πάνω του το literal_eval. Τα δεδομένα της λίστας είναι λεξικά με κλειδί το όνομα της στήλης οπότε στο αρχείο movie_keywords γράφουμε με το writerow το movie_id και το id του keyword της ταινίας, ενώ στο αρχείο keyword.csv γράφουμε το id του keyword και το όνομά του. Τέλος κλείνουμε και τα τρία αρχεία που είχαμε ανοίξει.



```
1 import csv, ast
2 fileToRead = open("keywords.csv", encoding='utf8')
3 fileToWrite1 = open("movie_keywords.csv", 'w', encoding='utf8', newline='')
4 fileToWrite2 = open("keyword.csv", 'w', encoding='utf8', newline='')
5 csvreader = csv.reader(fileToRead)
6 writer1 = csv.writer(fileToWrite1, delimiter=',', quotechar='"', quoting=csv.QUOTE_MINIMAL)
7 writer2 = csv.writer(fileToWrite2, delimiter=',', quotechar='"', quoting=csv.QUOTE_MINIMAL)
8 writer1.writerow(['movie_id', 'keyword_id']) # movie_keywords.csv header.
9 writer2.writerow(['id', 'name']) # keyword.csv header.
10 next(csvreader, None) # Skips the header.
11 for row in csvreader:
12     movie_id = row[0]
13     jsonString = row[1]
14     data = ast.literal_eval(jsonString)
15     for keyword in data:
16         writer1.writerow([movie_id, keyword['id']])
17         writer2.writerow([keyword['id'], keyword['name']])
18 fileToRead.close()
19 fileToWrite1.close()
20 fileToWrite2.close()
```

Με τα δύο καινούργια csv που παράχθηκαν με το παραπάνω πρόγραμμα τρέξαμε το πρόγραμμα που κάνει generate την sql για την δημιουργία πινάκων, φορτώσαμε την sql και την εκτελέσαμε στο pgAdmin και έπειτα ανεβάσαμε τα δεδομένα των csv με το import όπως με τους προηγούμενους πίνακες.

Αφαίρεση διπλότυπων

Στον πίνακα Keyword που δημιουργήσαμε υπήρχαν εξ ορισμού κάποια διπλότυπα καθώς ήταν βέβαιο πως κάποιες λέξεις κλειδιά θα αντιστοιχούσαν σε περισσότερες από μία ταινίες με αποτέλεσμα να γίνουν διπλότυπες εγγραφές. Αυτό το αντιμετωπίσαμε χρησιμοποιώντας έναν επιπλέον πίνακα ο οποίος έχει την κάθε λέξη κλειδί μία μόνο φορά, με χρήση της εντολής distinct στην sql (το αρχείο περιλαμβάνεται στο συνημμένα).

Περιορισμοί πρωτεύοντος και ξένου κλειδιού

Για να προσθέσουμε primary και foreign keys χρησιμοποιήσαμε εντολές alter table στην sql στο αντίστοιχο sql αρχείο alter_tables.