Χρήστος Πατρινόπουλος
3200150

Αλβιόνα Μάντσο
3200098

Data Bases.

2nd Assignment

Αρχικά ασχοληθήκαμε με την «επιπεδοποίηση» των αρχείων που περιείχαν μη ατομικό τύπο ιδιοτήτων. Το μόνο τέτοιο αρχείο είναι το **keywords.csv** λόγω του ότι η δεύτερη στήλη είναι σε JSON format. Σκοπός μας ήταν να παράξουμε 2 νέα csv αρχεία τα **keywords2.csv** και **movie_keywords.csv**.

- Το αρχείο keywords2.csv αντιστοιχεί στην οντότητα keywords και περιλαμβάνει 2 στήλες, για τις ιδιότητες **keyword_id** και **name**.
- Το αρχείο movie_keywords.csv αντιστοιχεί στην συσχέτιση movie_keywords και περιλαμβάνει 2 στήλες, για τα ξένα κλειδιά **movie_id** και **keyword_id**.

Διατρέξαμε επαναληπτικά το αρχικό αρχείο keywords.csv και με τη βοήθεια των δοσμένων βιβλιοθηκών διασπάσαμε την κάθε γραμμή σε 2 δεδομένα: το movie_id (row[0]) και το JSON string (row[1]). Με χρήση της συνάρτησης literal_eval (ast) μετατρέψαμε το JSON string σε μια λίστα λεξικών (**data**) της μορφής:

{“id”: xxx..x, “name”: ---- }.

Ως προς το keywords2.csv, για την εξάλειψη των διπλοτύπων, προσθέτουμε σε ένα set (**set_of_kw**) τα id που εμφανίζονται στη λίστα data και αντιστοιχούμε κάθε id στο name του μέσω ενός dictionary (**dict_of_kw**). Μετά το τέλος του parsing του αρχικού αρχείου, διατρέχουμε το set και για κάθε k_id στο αρχείο μια νέα γραμμή με μορφή:

keyword id *name of keyword*
↘ ↘
k_id, dict_of_kw[k_id]

Όσον αφορά το movie_keywords.csv, γράφουμε στο αρχείο μια νέα γραμμή για κάθε id που εμφανίζεται στη λίστα data ως εξής:

movie id *keyword id*
↘ ↘
row[0] , d[“id”]

Σημειώνουμε ότι κατά το άνοιγμα των αρχείων για διάβασμα (το keywords.csv) και γράψιμο (τα keywords2.csv και movie_keywords.csv) θέτουμε την παράμετρο encoding='utf8' για να αποφύγουμε προβλήματα που σχετίζονται με την κωδικοποίηση.

Σχετικά σχόλια υπάρχουν και στο αρχείο **keywords_file_splitter.py**.

Τα βήματα που έγιναν στη συνέχεια είναι τα εξής:

1) Παραγωγή sql αρχείων create table από το τρέξιμο του gen_ddl_python3.py με όρισμα κάθε ένα csv.
2) Τροποποίηση των sql αρχείων που παράχθηκαν στο βήμα 2 ώστε να ορίζονται και τα πρωτεύοντα κλειδιά των αρχείων που αντιστοιχούν σε σύνολα οντοτήτων.
3) Φόρτωση και τρέξιμο κάθε αρχείου στο **pgAdmin** για να οριστούν τα σχήματα στη βάση.
4) Φόρτωση δεδομένων από τα csv αρχεία με κατάλληλο ορισμό των παραμέτρων κωδικοποίησης, διαχωριστικών κλπ.
5) Τέλος, παραγωγή sql αρχείου με εντολές alter για τον εκ των υστέρων (σύμφωνα με τα ζητούμενα της εκφώνησης) ορισμό των ξένων κλειδιών των αρχείων που αναπαριστούν συσχετίσεις μεταξύ συνόλων οντοτήτων. Τρέξιμο αυτού του αρχείου στο **pgAdmin**.