2η Εργασία Βάσεις Δεδομένων – Αναφορά Παράδοσης

- Μητσάκης Νίκος : 3210122 - Παντελίδης Ιπποκράτης : 3210150

- **1ο Βήμα** : Αρχικά για την επιπεδοποίηση του αρχείου keywords.csv γράψαμε ένα python script που διαβάζει αυτό το csv αρχείο, το επεξεργάζεται και δημιουργεί δυο νέα csv αρχεία, το Keyword.csv και το hasKeyword.csv.Ειδικότερα, ανοίγουμε το αρχικό αρχείο σε 'r' mode και αφού προσπεράσαμε την επικεφαλίδα με 'next' ξεκινήσαμε ένα for loop που διαβάζει γραμμή-γραμμή το αρχείο και αρχικά χωρίζει το movie_id με την json συμβολοσειρά. Για την επεξεργασία της json συμβολοσειράς χρησιμοποιούμε python parsers και διατρέχουμε κάθε εμφωλευμένη json συμβολοσειρά με ένα for loop όπου για κάθε key παίρνουμε το αντίστοιχο value της με την μέθοδο get() και την πληροφορία που παίρνουμε την περνάμε σαν tuples στους δύο νέους πίνακες που θέλουμε να φτιάξουμε. Μετά τον χωρισμό movie_id και key_id στον αρχείο hasKeyword.csv βρήκαμε αρκετές γραμμές(1441) που είχαν movie_id, όμως είχαν null key_id και δεν τις λάβαμε υπόψη στο τελικό αρχείο.Αφού έχουν γεμίσει οι πίνακες μας, τους κάνουμε set ώστε να διαγραφούν τα διπλότυπα και έπειτα ανοίγουμε τα δύο νέα csv αρχεία σε 'w' mode και γράφουμε την επικεφαλίδα και τα δεδομένα που έχουμε συλλέξει.Τέλος παρατηρήσαμε διπλότυπα στους πίνακες : belongsTocollection, hasProductioncompany και hasGenre τα οποία αφαιρέσαμε όμοια με παραπάνω.
- **2ο Βήμα** : Στη συνέχεια χρησιμοποιήσαμε το extension SQL Server Import και πιο συγκεκριμένα την λειτουργία import wizard για την δημιουργία και την αρχικοποίηση των πινάκων μας στο Azure Data Studio από τα csv αρχεία μας. Στην δημιουργία ορισμένων πινάκων ο SQL server import δεν 'μάντεψε' σωστά τους τύπους των δεδομένων μας οπότε για να αντιμετωπίσουμε το error αναπροσαρμόσαμε τους τύπους των πεδίων σε πιο γενικούς. Επίσης σε μερικές περιπτώσεις χρειάστηκε να επιτρέψουμε NULL τιμές.
- **3ο Βήμα** : Στο τελευταίο βήμα της εργασίας, αφού αρχικοποιήσαμε και δημιουργήσαμε τους πίνακες μας, διαλέξαμε και δημιουργήσαμε τα πρωτεύοντα και τα ξένα κλειδιά των πινάκων μας. Τα κριτήρια επιλογής των πρωτευόντων κλειδιών είναι ότι αυτά είναι μοναδικά και δεν έχουν NULL τιμές, ενώ επιλέξαμε ως ξένα κλειδιά τα κλειδιά που αναφέρονται σε πρωτεύοντα κλειδιά κάποιου άλλου πίνακα(και μάλιστα σε μερικούς πίνακες βρήκαμε παραπάνω από ένα ξένο κλειδί). Αφού τα βρήκαμε, γράψαμε ένα SQL Query στο οποίο χρησιμοποιώντας την εντολή alter table δημιουργήσαμε τους περιορισμούς πρωτεύοντος και ξένου κλειδιού

σύμφωνα με τον χωρισμό που έχει κάνει η εκφώνηση.