Θεωρία Αποφάσεων

Project 2020-2021 / Github link

Δημιουργία ενός συστήματος που προτείνει ταινίες στους χρήστες σύμφωνα με τις ταινίες που έχουν παρακολουθήσει και τις αξιολογήσεις που έχουν κάνει. Ανάλυση δεδομένων, εξαγωγή πληροφοριών και οπτικοποίηση αυτών.

Μέλη ομάδας:

Ρήγας Στέφανος - 1047065
Ζούλφος Γεώργιος - 1047141
Καψουλάκης Ευάγγελος - 1047062

User-based Collaborative filtering

Σύστημα πρότασης ταινιών

Διακρίνονται δύο περιπτώσεις χρηστών

- Χρήστης που έχει **αξιολογήσει** ταινίες από το dataset
- Νέος χρήστης για τον οποίο είναι γνωστό μόνο το **φύλο** και η **ηλικία** του

Στην πρώτη περίπτωση

- Χρησιμοποιείται ο αλγόριθμος **kNN** για καθορισμό όμοιων χρηστών, η **ομοιότητα** εξαρτάται από τις αξιολογήσεις στις ταινίες
- Το **Μέσο Απόλυτο Σφάλμα (MAE)** για εύρεση κατάλληλου αριθμού κοντινών γειτόνων
- Ως βέλτιστο **k** ορίστηκε το 300

Στη δεύτερη περίπτωση

- Χρησιμοποιείται ο αλγόριθμος **kNN** για καθορισμό όμοιων χρηστών, η **ομοιότητα** εξαρτάται από το φύλο και την ηλικία
- Δεν μπορεί να **μετρηθεί** το σφάλμα πρόβλεψης

Demographics
Εξαγωγή Πληροφοριών

Επεξεργασία των **δημογραφικών** δεδομένων των datasets (όπως επάγγελμα, ηλικία, φύλο κλπ.) για εξαγωγή **πληροφοριών**:

- Ποιά είδη ταινιών είναι περισσότερο **προτιμητέα** από άνδρες και ποιά από γυναίκες
- Ποσοστά των ηλικιών των users και πιο **δημοφιλής ταινία** για το εκάστοτε age_desc
- Δημοφιλέστερη ταινία ανάλογα το **επάγγελμα** του κάθε user
- Επεξεργασία των **timestamps** για να δούμε ποια ταινία έχει πάρει τις περισσότερες βαθμολογίες ανα χρονιά
- **Zipcodes** → Ποια ταινία είναι δημοφιλέστερη ανάλογα τον ταχυδρομικό κώδικα

Τα παραπάνω αποτελέσματα αναλύονται σε γραφήματα και σε console outputs.

Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν για **πρόταση** ταινιών στους χρήστες (έμμεσα ή άμεσα).

Showtime
Ανάλυση Δεδομένων

Στο αρχείο αυτό γίνεται **ανάλυση** των δεδομένων κυρίως των ratings και movies datasets, για να εκμαιεύσουμε **στατιστικά** και να **οπτικοποιήσουμε** δεδομένα.

Η οπτικοποίηση γίνεται με τη matplotlib και στη συνέχεια παραθέτουμε τα **συμπεράσματά** μας. Τα δεδομένα που οπτικοποιήθηκαν περιλαμβάνουν:

- **Συχνότητα** βαθμολογιών των χρηστών για όλες τις αξιολογήσεις των ταινιών στο dataset
- Πορεία **αξιολογήσεων** παραγωγών ταινιών στο πέρασμα του χρόνου
- Δημοτικότητα **ειδών** ανα τα έτη, βασισμένη στη συχνότητα αξιολογήσεων τους
- Στατιστική **δημοτικότητα** των κορυφαίων **5** ειδών ταινιών

Τέλος ένας περεταίρω υπολογισμός που έγινε, είναι η εύρεση των καλύτερων ταινιών **Δράσης** της κάθε **δεκαετίας**, οι οποίες εκτυπώνονται στο console output.

Content-Based Filtering
Προεπεξεργασία

Στον φάκελο αυτό έγινε μια προσπάθεια για την δημιουργία ενός 2ου συστήματος πρότασης ταινιών με βάση του **περιεχόμενο των ταινιών** που παρακολουθεί ένας χρήστης. Τα βήματα που υλοποιήθηκαν είναι τα εξής:

- Καθαρισμός του movies dataset και **δημιουργία** ενός νέου αρχείου που περιλαμβάνει **μόνο τα είδη** των ταινιών από το original movies dataset
- Απομόνωση των **μοναδικών** ειδών, αλλά και των μοναδικών **συνδυασμών** ειδών
- Χρήση της συνάρτησης **get_dummies** για να μετρήσουμε κάθε περίσταση που 2 είδη **συνυπάρχουν** σε μια ταινία (π.χ. Τα είδη Action με Drama βρίσκονται μαζί σε 100 ταινίες)
- Αντιστοίχιση των **συσχετίσεων** (αθροίσματα από την get_dummies) στους μοναδικούς συνδυασμούς και φθίνουσα **ταξινόμηση** των αποτελεσμάτων σε αρχείο csv

Το αποτέλεσμα των παραπάνω είναι η **γνώση των σχέσεων** μεταξύ των ειδών, που θα καθορίσει το πως θα προτείνει ταινίες ο αλγόριθμος που θα επιλέξουμε.

Ευχαριστούμε