Εργασία 5 – Συστήματα Συστάσεων

Στόχος της εργασίας είναι η υλοποίηση και η αποτίμηση ενός συστήματος συστάσεων συνεργατικού φιλτραρίσματος αντικειμένου-αντικειμένου για ταινίες. Η υλοποίηση μπορεί να γίνει σε όποια γλώσσα προγραμματισμού προτιμάτε (συνιστάται η χρήση python). Θα χρησιμοποιήσετε τα δεδομένα του MovieLens (https://grouplens.org/datasets/movielens/latest/) και συγκεκριμένα το μικρό σύνολο με τις 100000 βαθμολογίες (απαιτείται μόνο το αρχείο με τα ratings).

1. Ως προεπεξεργασία
   - φιλτράρετε τις ταινίες ώστε να κρατήσετε μόνο αυτές που έχουν τουλάχιστον 5 βαθμολογίες.
   - φιλτράρετε τους χρήστες ώστε να κρατήσετε μόνο αυτούς που έχουν δώσει τουλάχιστον 5 βαθμολογίες.
2. Για την υλοποίηση του συστήματος ως προς την ομοιότητα θα χρησιμοποιήσετε adjusted cosine (ή pearson). Για την αποτίμηση του συστήματος, τα δεδομένα θα χωρίζονται σε δύο μέρη: σύνολο εκπαίδευσης και σύνολο ελέγχου. Το σύνολο ελέγχου θα αποτελείται από το **10% των δεδομένων** (μετά την προεπεξεργασία) και θα διατηρείται **σταθερό**.
3. Για την υλοποίηση του συστήματος ως προς την συνάρτηση πρόβλεψης θα υλοποιηθούν οι επιλογές:
   - μέσος όρος
   - σταθμισμένος μέσος όρος
   - σταθμισμένος μέσος όρος στον οποίο όμως η στάθμιση θα βασίζεται στο πλήθος των κοινών χρηστών που έχουν βαθμολογήσει τα δύο αντικείμενα και όχι στην ομοιότητα των αντικειμένων όπως υπολογίζεται από το adjusted cosine. Ορίστε συνάρτηση στάθμισης της επιλογής σας.
4. Το πρόγραμμα σας θα πρέπει να δέχεται ως παράμετρο το πλήθος των κοντινότερων γειτόνων K βάση των οποίων θα πρέπει να γίνεται η πρόβλεψη καθώς και το ποσοστό του συνόλου εκπαίδευσης.
5. Τα μέτρα αποτίμησης θα είναι το μέσο απόλυτο σφάλμα (MAE) και η ακρίβεια (precision) και ανάκληση (recall). Για τον υπολογισμό των δυαδικών μέτρων αποτίμησης θεωρείστε μια ταινία ως σχετική αν ο βαθμός της είναι >=3.

Πραγματοποιήστε τα εξής πειράματα:
--Για σταθερό σύνολο εκπαίδευσης 90% και για 5 τιμές K συγκρίνετε τις 3 συναρτήσεις πρόβλεψης.
--Χρησιμοποιώντας το καλύτερο K για κάθε μέθοδο όπως προέκυψε από το προηγούμενο πείραμα, συγκρίνετε τις 3 μεθόδους με σύνολο εκπαίδευσης: 50%, 70% και 90% του αρχικού συνόλου δεδομένων. Για τον έλεγχο χρησιμοποιούμε το ίδιο 10% πάντα.

Θα παραδώσετε:
α) τον κώδικα σας
β) μια αναφορά στην οποία θα παρουσιάζετε και θα σχολιάζετε τα αποτελέσματα σας.
Στην αναφορά σε σχέση με τον κώδικα σας, σχολιάστε όποια βελτιστοποίηση ή υπόθεση κάνατε για οτιδήποτε δεν καθορίζεται από την εκφώνηση.

Σημείωση: Αν υπολογιστικά αντιμετωπίσετε πρόβλημα με το μέγεθος των δεδομένων, μειώστε τα (αναφέρετε τα νέα μεγέθη αρχείων) Π.χ. κάνοντας προεπεξεργασία με κατώφλι μεγαλύτερο του 5.