

Εργασία 5 – Ανάκτηση Πληροφορίας και Μηχανές Αναζήτησης

Στόχος της εργασίας είναι η υλοποίηση και σύγκριση συστημάτων συστάσεων που βασίζονται στο συνεργατικό φιλτράρισμα (collaborative filtering). Η υλοποίηση μπορεί να γίνει σε όποια γλώσσα προγραμματισμού προτιμάτε (π.χ. python).

A. Αρχικά θα πρέπει να γίνεται αρχικοποίηση ενός πίνακα οφέλους με διαστάσεις $N \times M$, N χρήστες και M αντικείμενα. Ο πίνακας θα πρέπει να είναι γεμάτος με βαθμολογίες πραγματικών αριθμών από 1.00 έως 10.0 με ομοιόμορφη κατανομή.

B. Το πρόγραμμά σας θα πρέπει να υλοποιεί ένα σύστημα collaborative filtering αντικειμένου-αντικειμένου. Για την ομοιότητα θα πρέπει να υλοποιήσετε τέσσερις διαφορετικές συναρτήσεις ομοιότητας, Jaccard, Dice, cosine και adjusted cosine. Για την πρόβλεψη του βαθμού θα χρησιμοποιείται η τεχνική των K-κοντινότερων γειτόνων και είτε ο μέσος όρος είτε ο σταθμισμένος μέσος όρος των βαθμολογιών.

Γ. Για μια εκτέλεση το σύστημα θα επιλέγει τυχαία ένα ποσοστό $X\%$ των συνολικών βαθμολογιών ως γνωστές, και σύμφωνα με αυτές θα υπολογίζει προβλέψεις για τις υπόλοιπες $(100-X)\%$ των βαθμολογιών. Για τις προβλέψεις αυτές θα πρέπει να υπολογίζετε και το μέσο απόλυτο λάθος το οποίο θα είναι η έξοδος του προγράμματος σας. Το σύστημα θα επαναλαμβάνει για T επαναλήψεις αυτήν την εκτέλεση και θα παίρνει το μέσο λάθος από όλες.

Σχεδιάστε και εκτελέστε πειράματα που θα απαντούν στα παρακάτω:

- 1) Πώς επηρεάζει το ποσοστό των γνωστών βαθμολογιών την απόδοση (ακρίβεια) του συστήματος;
- 2) Ποια είναι η καλύτερη συνάρτηση ομοιότητας;
- 3) Πώς επηρεάζει το K την απόδοση (ακρίβεια) του συστήματος;
- 4) Ποια είναι η καλύτερη μέθοδος πρόβλεψης;

Bonus: Σχεδιάστε μια δική σας μέθοδο πρόβλεψης πέρα από τον απλό μέσο όρο και τον σταθμισμένο μέσο όρο, που πειραματικά θα δείξετε ότι έχει τουλάχιστον αντίστοιχη απόδοση με τις άλλες δύο.

Χρησιμοποιήστε ενδεικτικά (αλλά μπορείτε και να τα τροποποιήσετε αν θέλετε) 100 χρήστες και 100 αντικείμενα και 10 επαναλήψεις για κάθε πείραμα.

Παραδοτέα:

- Ο κώδικας σας, με σχόλια όπου θεωρείτε ότι χρειάζονται.
- Μια αναφορά στην οποία θα παρουσιάζετε και θα σχολιάζετε τα πειράματα που σχεδιάσατε καθώς και τα αποτελέσματα που προέκυψαν.