Ανάκτηση Πληροφορίας και Μηχανές Αναζήτησης – Εργασία 3

Θα υλοποιήσετε ένα σύστημα συστάσεων (recommender system). Η υλοποίηση μπορεί να γίνει σε όποια γλώσσα προγραμματισμού προτιμάτε. Συγκεκριμένα, το πρόγραμμα σας θα πρέπει να παρέχει τις παρακάτω λειτουργίες.

A. Τυχαία παραγωγή πίνακα οφέλους, με διαστάσεις N×M (N χρήστες × M αντικείμενα). Ο πίνακας θα πρέπει να είναι κατά X% γεμάτος με βαθμολογίες ακεραίων από 1 έως 5. Θεωρείστε ομοιόμορφη κατανομή για την κατανομή των βαθμών στον πίνακα.
B. Θα υλοποιεί σύστημα συστάσεων χρήστη-χρήστη και αντικειμένου-αντικειμένου.
C. Συνάρτηση ομοιότητας Jaccard.
D. Συνάρτηση ομοιότητας συνημιτόνου (cosine).
E. Συνάρτηση ομοιότητας Pearson.
F. Υλοποίηση πρόβλεψης βαθμού με βάση K κοντινότερους γείτονες και χρήση ζυγισμένου αθροίσματος (στον παρονομαστή το άθροισμα των ομοιοτήτων).
G. Υπολογισμός μέσου απόλυτου λάθους.

Το πρόγραμμα θα πρέπει να διαβάζει από ένα configuration αρχείο τις παραμέτρους εκτέλεσης, οι οποίες θα περιλαμβάνουν (με τη σειρά):

T: ο αριθμός επαναλήψεων της εκτέλεσης
N: το πλήθος των χρηστών του συστήματος
M: το πλήθος των αντικειμένων του συστήματος
X: το ποσοστό που δείχνει πόσο γεμάτος είναι ο πίνακας
K: το πλήθος των κοντινότερων γειτόνων για την πρόβλεψη βαθμολογίας

Στη συνέχεια, επαναληπτικά για T επαναλήψεις:

1) Θα παράγει έναν νέο τυχαίο πίνακα οφέλους σύμφωνα με τις παραμέτρους εισόδου.
2) Θα υλοποιεί ένα σύστημα χρήστη-χρήστη, χρησιμοποιώντας μία από τις ομοιότητες κάθε φορά (άρα τρείς φορές) και θα γεμίζει τον πίνακα οφέλους με προβλέψεις για όλες τις βαθμολογίες που λείπουν λαμβάνοντας υπόψη K κοντινότερους γείτονες.
3) Θα αποτιμά το μέσο απόλυτο λάθος ανά χρήστη, αλλά και συνολικά για τον πίνακα για κάθε μέτρο ομοιότητας.
4) Θα γράφει τον πίνακα αλλά και τις μετρήσεις λάθους σε ξεχωριστά αρχεία (με χαρακτηριστικά ονόματα).
5) Θα επαναλαμβάνει τα βήματα 2 έως 4 για ένα σύστημα που χρησιμοποιεί ομοιότητα αντικειμένου-αντικειμένου. Στο βήμα 3 σε αυτήν την περίπτωση βγάλτε το λάθος ανά αντικείμενο και φυσικά το συνολικό όπως και πριν.

Θα εκτελέσετε τα εξής πειράματα:
Με T=10, N=50, M=50.
A) Με σταθερό ποσοστό X=75% γεμάτο πίνακα, τρέξτε το πρόγραμμα σας για K=3, K=5 και K=10 και σχεδιάστε γραφική παράσταση που δείχνει το μέσο απόλυτο λάθος (πάρτε το μέσο όρο λάθους από τις 10 επαναλήψεις για κάθε διαμόρφωση) για κάθε διαφορετική ομοιότητα τόσο για σύστημα με ομοιότητα χρήστη-χρήστη, όσο και αντικειμένου-αντικειμένου. Δηλαδή συνολικά έχετε 6 συνδυασμούς για κάθε τιμή του K.
B) Με σταθερό K=5, τρέξτε το πρόγραμμα σας για X=80%, X=70%, X=50% και X=30%, και σχεδιάστε γραφική παράσταση που δείχνει το μέσο απόλυτο λάθος (πάρτε το μέσο όρο λάθους από τις 10 επαναλήψεις για κάθε διαμόρφωση) για κάθε διαφορετική ομοιότητα τόσο για σύστημα με ομοιότητα χρήστη-χρήστη, όσο και αντικειμένου-αντικειμένου. Δηλαδή συνολικά έχετε 6 συνδυασμούς για κάθε τιμή του X.

Παραδοτέα:

- Ο κώδικας του προγράμματος σας καθώς και ένα αρχείο configuration (ως δείγμα για να δω τη μορφή του).
- Τα αρχεία εξόδου που παράγονται συγκεκριμένα για T=10, N=50, M=50, X=75, K=5.
- Τις δύο γραφικές παραστάσεις που παρουσιάζουν τα αποτελέσματα των δύο πειραμάτων.
- Ένα αρχείο κειμένου με τις οδηγίες εκτέλεσης ή όποιο άλλο σχόλιο ή παρατήρηση θέλετε.

Συμπιέστε όλα τα παραπάνω και στείλτε το αρχείο που προκύπτει.