Ερώτηση 3

Σε αυτή την ερώτηση θα χρησιμοποιήσετε το κοινωνικό δίκτυο μεταξύ των χρηστών του Yelp για να προβλέψετε τα ratings τους για νέες επιχειρήσεις. Για την πρόβλεψη θα υλοποιήσετε τον αλγόριθμο για value propagation τον οποίο περιγράψαμε στην τάξη.

Για την υλοποίηση σας θα ακολουθήσετε τα εξής βήματα.

Βήμα 1: Θα χρησιμοποιήσετε παρόμοια δεδομένα με αυτά που δημιουργήσατε για την Δεύτερη Σειρά Ασκήσεων, για τα συστήματα συστάσεων. Η διαφορά είναι ότι θα κάνετε πιο επιθετικό pruning. Συγκεκριμένα, στα τελικά σας δεδομένα θα έχετε ένα σύνολο από χρήστες U, και ένα σύνολο από επιχειρήσεις B, όπου ο κάθε χρήστης στο U θα έχει **τουλάχιστον 30 reviews** σε επιχειρήσεις στο B, και η κάθε επιχείρηση στο B θα έχει **τουλάχιστον 50 reviews** από χρήστες στο U. Για διευκόλυνση σας δίνονται τα δεδομένα χρήστης-επιχείρηση-rating για την πόλη της Φιλαδέλφειας τα οποία θα πρέπει να κλαδέψετε στο αρχείο philly_users_businesses_stars.csv στη σελίδα των Ασκήσεων.

Δημιουργείστε ένα γράφημα με κορυφές τους χρήστες στο σύνολο U και ακμές τις φιλίες μεταξύ των χρηστών στο U, τις οποίες θα πάρετε από το αρχείο yelp_academic_dataset_user.json . Από αυτό το γράφημα κρατήστε τη μεγαλύτερη συνεκτική συνιστώσα (θα πρέπει να είναι ίδια με όλο το γράφημα). Αυτή θα ορίσει το γράφημα G με το οποίο θα δουλέψετε. Μετά από αυτή την διαδικασία θα πρέπει να μείνετε με 1504 χρήστες και 887 επιχειρήσεις.

Βήμα 2: Δημιουργείστε τα training και test δεδομένα. Από τους χρήστες διαλέξτε τυχαία 100 χρήστες. Από την ένωση όλων των επιχειρήσεων που έχουν βαθμολογήσει οι χρήστες που επιλέξατε επιλέξτε τυχαία 100 επιχειρήσεις. Οι τριάδες χρήστης-επιχείρηση-rating με τους χρήστες και τις επιχειρήσεις που επιλέξατε θα είναι τα test δεδομένα, D_{test}. Οι υπόλοιπες τριάδες θα είναι τα train δεδομένα D_{train}. Υλοποιήστε την επιλογή των δεδομένων αλλά για τα επόμενα βήματα χρησιμοποιήστε τα δεδομένα από τα αρχεία train_data.csv και test_data.csv που σας δίνονται στη σελίδα των Ασκήσεων.

Βήμα 3: Ο στόχος είναι να προβλέψουμε τα ratings στο D_{test} πραγματοποιώντας διάχυση τιμών (value propagation) στο γράφημα G. Έστω B_{test} το σύνολο των 100 επιχειρήσεων που επιλέξατε στο Βήμα 2, και εμφανίζονται στο D_{test}. Θα πραγματοποιήσετε την διαδικασία του value propagation για κάθε επιχείρηση $b \in B_{test}$. Για κάθε τριάδα χρήστη-επιχείρηση-rating (u, b, r) που εμφανίζεται στο D_{train} ο κόμβος u στο γράφημα θα έχει σταθερή τιμή r (γίνεται απορροφητικός), και για κάθε άλλο χρήστη-κόμβο v (μη απορροφητικός), θα υπολογίσετε την τιμή $R(v, b)$ χρησιμοποιώντας τη διαδικασία του value propagation που περιγράψαμε στην τάξη. Για κάθε (μη απορροφητικό) κόμβο v, θέτουμε

$$R(v, b) = \frac{1}{|N_v|} \sum_{u \in N_v} R(u, b)$$

όπου N_v είναι οι γείτονες του κόμβου v στο γράφημα G. Ο υπολογισμός αυτός γίνεται επαναληπτικά μέχρι να συγκλίνει (δηλαδή η μέγιστη διαφορά της νέας τιμής με την παλιά να είναι κάτω από ένα threshold, π.χ. 10^{-6}). Για ένα ζευγάρι χρήστη-επιχείρηση (x, b) η πρόβλεψη σας θα είναι η τιμή $R(x, b)$. Υπολογίσετε το Root Mean Square Error (RMSE) για αυτή τη μέθοδο.

Βήμα 4: Τρέξτε τους αλγορίθμους UCF, ICF, UA, IA που υλοποιήσατε στην Δεύτερη Άσκηση για αυτό το dataset και συγκρίνετε το Root Mean Square Error (RMSE) με τη μέθοδο Propagation. Παρουσιάστε τα αποτελέσματα σας και γράψετε τις παρατηρήσεις σας.