# ΑΛΓΟΡΙΘΜΟΙ ΓΙΑ ΔΕΔΟΜΕΝΑ ΕΥΡΟΙΑΣ ΚΛΙΜΑΚΑΣ

2η ΑΝΑΘΕΣΗ ΓΙΑ ΤΟ ΣΠΙΤΙ

## Τεχνικές Μέτρησης Ζευγαριών

#### Κάτω Τριγωνικό Μητρώο Μέτρησης Ζευγαριών

Η υλοποίηση του κάτω τριγωνικού μητρώου μέτρησης ζευγαριών βρίσκεται στο αρχείο pair\_finder.py, στην ρουτίνα TriangularMatrixOfPairCounters(line: 17). Αρχικά φτιάχνω μία λίστα μεγέθους n\*(n-1)/2 (n=πλήθος ταινιών της συλλογή) η οποία είναι μία γραμμική αναπαράσταση του κάτω τριγωνικού πίνακα. Στην συνέχεια διατρέχω τα καλάθια των χρηστών και δημιουργώ μία λίστα με όλα τα πιθανά ζευγάρια που μπορούν να παραχθούν για το συγκεκριμένο καλάθι. Διατρέχω την κάθε λίστα με τα ζεύγη και υπολογίζω τα indexes της κάθε ταινίας με βάση το movieID (Πχ η λίστα με τα ζευγάρια για ένα καλάθι ‘χ’ είναι [[3,2], [6,5]]. Οι δύο τιμές του κάθε ζευγαριού είναι movieIDs εμείς όμως για να μπορέσουμε να βρούμε την θέση που αντιστοιχεί στον μητρώο χρειαζόμαστε το index της κάθε ταινίας) και τα εισάγω στον τύπο που διδαχθήκαμε στο μάθημα για να βρω την θέση του μετρητή του ζευγαριού στο κάτω τριγωνικό μητρώο.

#### Λεξικό Μέτρησης Ζευγαριών

Η υλοποίηση του λεξικού μέτρησης ζευγαριών βρίσκεται στο αρχείο pair\_finder.py, στην ρουτίνα HashCountersOfPairs(line: 50). Η ρουτίνα δουλεύει ως εξής. Διατρέχω ένα ένα τα καλάθια των χρηστών. Για κάθε καλάθι δημιουργώ όλα τα πιθανά ζευγάρια και τα επιστρέφω ως λίστα. Στην συνέχει διατρέχω την λίστα με τα ζευγάρια και ελέγχω αν υπάρχει ήδη το ζευγάρι στο λεξικό. Αν υπάρχει απλά αυξάνω τον μετρητή κατά ένα, αλλιώς αρχικοποιώ νέο μετρητή στο λεξικό με κλειδί το τρέχον ζευγάρι.

#### Σύγκριση Τεχνικών

Είναι ξεκάθαρο από τα αποτελέσματα που φαίνονται παρακάτω ότι το λεξικό είναι πιο αποδοτικό.

Εικόνα που περιέχει μέτρο

Περιγραφή που δημιουργήθηκε αυτόματα\*\*Η σύγκριση έγινε για κριτικές με MinScore = 4 στο αρχείο με τους 100 χρήστες.

Ο τριγωνικός πίνακας είναι πολύ αργός για τους παρακάτω λόγους:

1. Ο υπολογισμός της θέσης του ζευγαριού στη λίστα σειριακή αναπαράστασης του κάτω τριγωνικού πίνακα περιέχει ακριβές πράξεις (πολλαπλασιασμός, διαίρεση).
2. Πρέπει να υπολογίζουμε κάθε φορά τις θέσεις των δύο ταινιών στο movie dataframe. (Αυτή καθυστέρηση θα μπορούσε να αντιμετωπιστεί αν περνούσαμε τα δεδομένα μας από προεργασία ώστε το movieId να είναι ένας αύξων αριθμός από το 0 έως το ν-1.)
3. Ο πίνακας δεσμεύει χώρο ακόμα και για τα ζευγάρια που δεν συναντάμε.

Καταλήγουμε λοιπόν ότι ο πίνακας κατακερματισμού (λεξικό) είναι σημαντικά πιο αποδοτικός από άποψη χρόνου αλλά και από άποψη χώρου.

## Tεχνικές Εύρεσης Κανόνων Συσχέτισης

#### Αλγόριθμος A Priori (Χωρίς δειγματοληψία)

#### Αλγόριθμος A Priori (Με δειγματοληψία)

#### Απόδοση A Priori με δειγματοληψία

#### Σχηματισμός Κανόνων Συσχέτισης