ΜΥΕ047 – Αλγόριθμοι για Δεδομένα Ευρείας Κλίμακας

1η Ανάθεση

Μέλη Ομάδας:

Λεωνίδας Ζαφειρίου, 2972

Αχιλλέας Νταλαγιώργος, 3049

Α) Περιεχόμενα παραδοτέου

Το παραδοτέο αρχείο αποτελείται από τα εξής:

- SOURCES
    - preprocess.py
    - main.py
    - universalHashFunctions.py (όπως μας δόθηκε)
- EXPERIMENTS
    - Παραχθέντα csv αρχεία με τα λεξικά userList, movieMap και movieList και τα μητρώα SIG για τα αρχεία δεδομένων ratings_100users.csv και ratings.csv
    - Screenshots από την εκτέλεση της πειραματικής αξιολόγησης
    - Εικόνες με τις γραφικές παραστάσεις των false-pos, false-neg και PRECISION, RECALL, F1 metrics
- report.pdf

Β) Χρήση βιβλιοθηκών

Για να γίνουν compile τα αρχεία απαιτούνται οι εξής βιβλιοθήκες:

- numpy
- matplotlib

Γ) Περιγραφή υλοποίησης

- *preprocess.py*

Το αρχείο αυτό απαντάει στο ερώτημα (1α). Ως command-line argument παίρνει ένα αρχείο δεδομένων <name>.csv και παράγει από αυτό τα λεξικά userList, movieMap και movieList τα οποία γίνονται export σε csv και αποθηκεύονται στον φάκελο EXPERIMENTS. Γι αυτό το λόγο, το preprocess.py πρέπει να εκτελεστεί πρώτο ώστε να είναι διαθέσιμα τα λεξικά σε csv για να χρησιμοποιηθούνε μετέπειτα.

- *main.py*

Το αρχείο αυτό περιέχει τη βασική λειτουργικότητα του προγράμματος. Ως command-line arguments παίρνει ένα αρχείο δεδομένων <name>.csv, το κατώφλι <s>, τον αριθμό <x> των πρώτων ταινιών που θα συγκριθούν και μία παράμετρο <i> η οποία για τιμή 1 δημιουργεί ένα καινούριο μητρώο υπογραφών SIG, ενώ για τιμή 0 χρησιμοποιεί το ήδη υπάρχον. Επίσης, στο αρχείο υλοποιούνται οι συναρτήσεις που ζητούνται στα ερωτήματα (1β), (1γ), (1δ), (1ε) και εκτελείται η πειραματική αξιολόγηση που ζητείται στο ερώτημα (1ζ).

Συγκεκριμένα, όσον αφορά τα ερωτήματα (1ζ1) και (1ζ2) αυτά υλοποιούνται με τις συναρτήσεις minHashingExperimentation( ) και LSHExperimentation( ). Οι συναρτήσεις αυτές υλοποιούν τη λειτουργικότητα που ζητείται, κάνουν print στην οθόνη τα αποτελέσματα και plot τις γραφικές παραστάσεις.

Δ) Αποτελέσματα πειραματικής αξιολόγησης και ερμηνεία

(ΣΗΜΕΙΩΣΗ: οι διακεκομμένες γραμμές που φαίνονται στις γραφικές παραστάσεις που ακολουθούν δεν σηματοδοτούν ακριβείς τιμές συνάρτησης. Ουσιαστικά πρόκειται για διάγραμμα διακριτών τιμών και οι διακεκομμένες γραμμές χρησιμοποιούνται μόνο για να δοθεί διαισθητικά μία ιδέα για το πώς συμπεριφέρονται οι τιμές)

- *Ερώτημα (1ζ1), ratings_100users.csv*

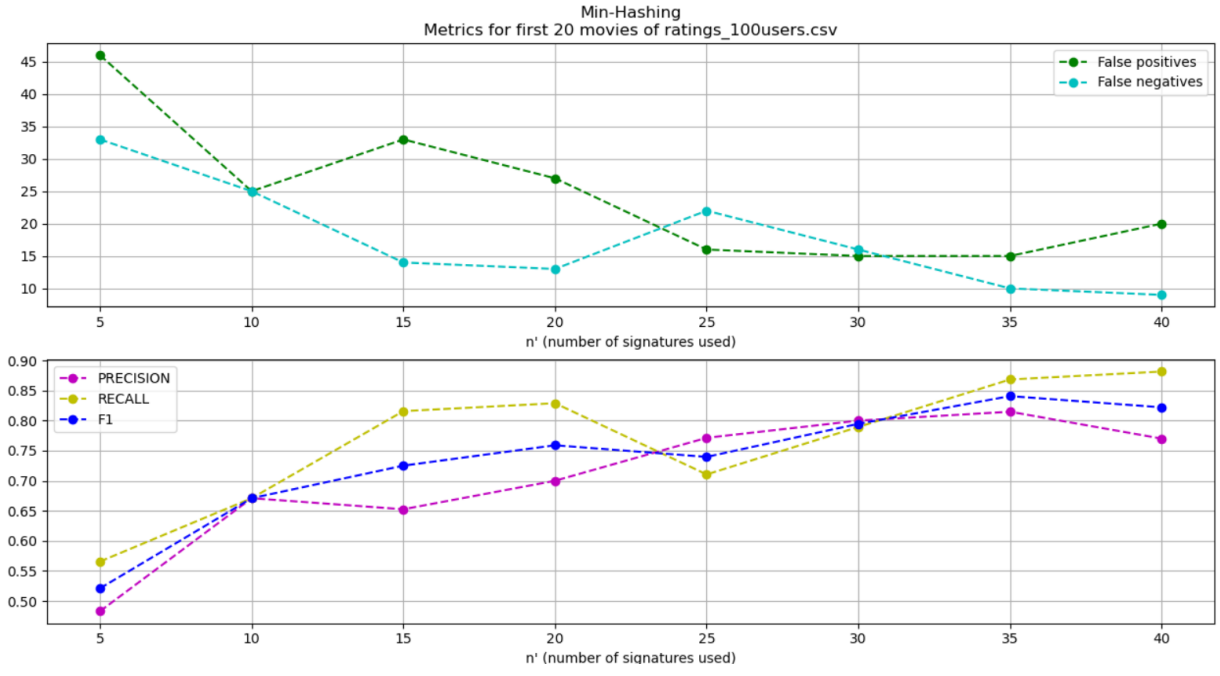
Στα σχήματα 1 και 2 που βρίσκονται παρακάτω φαίνονται τα αποτελέσματα της εκτέλεσης του αλγορίθμου minHashingExperimentation( ) για τις οκτώ (8) διαφορετικές τιμές του n'. Συγκεκριμένα, όπως φαίνεται και στη γραφική παράσταση στο Σχήμα 2, όσο μεγαλώνει ο αριθμός n' των υπογραφών που χρησιμοποιούνται από το μητρώο SIG, τόσο μειώνονται τα λάθη, δηλαδή ο αριθμός των false-negatives και false-positives και φυσικά τόσο μεγαλώνει το score F1. Αυτό, συμβαίνει γιατί όσο περισσότερες υπογραφές λαμβάνει υπόψη του ο αλγόριθμος τόσο πιο αντιπροσωπευτική είναι η σύγκριση ομοιότητας μεταξύ δύο (2) ταινιών. Οι καλύτερες τιμές του F1 σημειώνονται για τις τιμές n' = 35 και n' = 40.

- *Ερώτημα (1ζ1), ratings.csv*

Όμοια συμβαίνει και σε αυτήν την περίπτωση, όπως φαίνεται στα σχήματα 3 και 4, με την καλύτερη τιμή του F1 να σημειώνεται για την περίπτωση των n' = 40 υπογραφών.

```
                #####
Experimentation: MIN-HASHING
File: 'ratings_100users.csv'
First '20' movies are compared.
Threshold = 0.25, Number of signatures = 40
                #####

Number of relevant elements (ground truth) = 76

----- F1 scores for n' = 5: -----
True-positives = 43, False-positives = 46
True-negatives = 68, False-negatives = 33
PRECISION: 0.48314606741573035, RECALL: 0.5657894736842105, F1: 0.5212121212121212


----- F1 scores for n' = 10: -----
True-positives = 51, False-positives = 25
True-negatives = 89, False-negatives = 25
PRECISION: 0.6710526315789473, RECALL: 0.6710526315789473, F1: 0.6710526315789473


----- F1 scores for n' = 15: -----
True-positives = 62, False-positives = 33
True-negatives = 81, False-negatives = 14
PRECISION: 0.6526315789473685, RECALL: 0.8157894736842105, F1: 0.7251461988304094


----- F1 scores for n' = 20: -----
True-positives = 63, False-positives = 27
True-negatives = 87, False-negatives = 13
PRECISION: 0.7, RECALL: 0.8289473684210527, F1: 0.7590361445783133


----- F1 scores for n' = 25: -----
True-positives = 54, False-positives = 16
True-negatives = 98, False-negatives = 22
PRECISION: 0.7714285714285715, RECALL: 0.7105263157894737, F1: 0.7397260273972601


----- F1 scores for n' = 30: -----
True-positives = 60, False-positives = 15
True-negatives = 99, False-negatives = 16
PRECISION: 0.8, RECALL: 0.7894736842105263, F1: 0.794701986754967


----- F1 scores for n' = 35: -----
True-positives = 66, False-positives = 15
True-negatives = 99, False-negatives = 10
PRECISION: 0.8148148148148148, RECALL: 0.868421052631579, F1: 0.8407643312101911


----- F1 scores for n' = 40: -----
True-positives = 67, False-positives = 20
True-negatives = 94, False-negatives = 9
PRECISION: 0.7701149425287356, RECALL: 0.881578947368421, F1: 0.8220858895705521
```

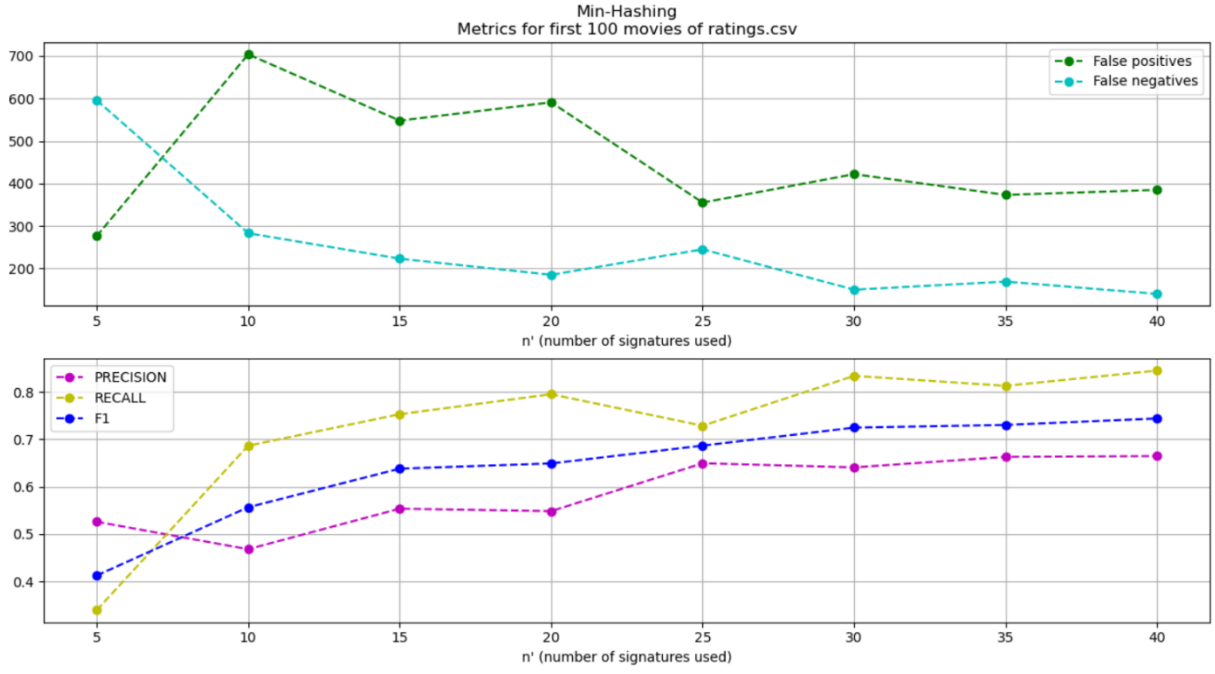
Σχήμα 1: Αποτελέσματα του αλγορίθμου minHashingExperimentation( ) για το αρχείο ratings_100users.csv



Σχήμα 2: Γραφικές παραστάσεις με βάση τα αποτελέσματα που φαίνονται στο Σχήμα 1

```
                #####
Experimentation: MIN-HASHING
File: 'ratings.csv'
First '100' movies are compared.
Threshold = 0.25, Number of signatures = 40
                #####

Number of relevant elements (ground truth) = 902

----- F1 scores for n' = 5: -----
True-positives = 306, False-positives = 276
True-negatives = 3772, False-negatives = 596
PRECISION: 0.5257731958762887, RECALL: 0.3392461197339246, F1: 0.41239892183288407


----- F1 scores for n' = 10: -----
True-positives = 619, False-positives = 704
True-negatives = 3344, False-negatives = 283
PRECISION: 0.46787603930461075, RECALL: 0.6862527716186253, F1: 0.5564044943820224


----- F1 scores for n' = 15: -----
True-positives = 679, False-positives = 548
True-negatives = 3500, False-negatives = 223
PRECISION: 0.5533822330888346, RECALL: 0.7527716186252772, F1: 0.6378581493658995


----- F1 scores for n' = 20: -----
True-positives = 717, False-positives = 591
True-negatives = 3457, False-negatives = 185
PRECISION: 0.5481651376146789, RECALL: 0.79490022172949, F1: 0.648868778280543


----- F1 scores for n' = 25: -----
True-positives = 657, False-positives = 355
True-negatives = 3693, False-negatives = 245
PRECISION: 0.6492094861660079, RECALL: 0.7283813747228381, F1: 0.6865203761755486


----- F1 scores for n' = 30: -----
True-positives = 752, False-positives = 422
True-negatives = 3626, False-negatives = 150
PRECISION: 0.6405451448040886, RECALL: 0.8337028824833703, F1: 0.7244701348747592


----- F1 scores for n' = 35: -----
True-positives = 733, False-positives = 373
True-negatives = 3675, False-negatives = 169
PRECISION: 0.6627486437613019, RECALL: 0.8126385809312638, F1: 0.7300796812749004


----- F1 scores for n' = 40: -----
True-positives = 762, False-positives = 385
True-negatives = 3663, False-negatives = 140
PRECISION: 0.6643417611159547, RECALL: 0.844789356984479, F1: 0.7437774524158127
```

Σχήμα 3: Αποτελέσματα του αλγορίθμου minHashingExperimentation( ) για το αρχείο ratings.csv



Σχήμα 4: Γραφικές παραστάσεις με βάση τα αποτελέσματα που φαίνονται στο Σχήμα 3

- *Ερώτημα (1ζ2), ratings_100users.csv*

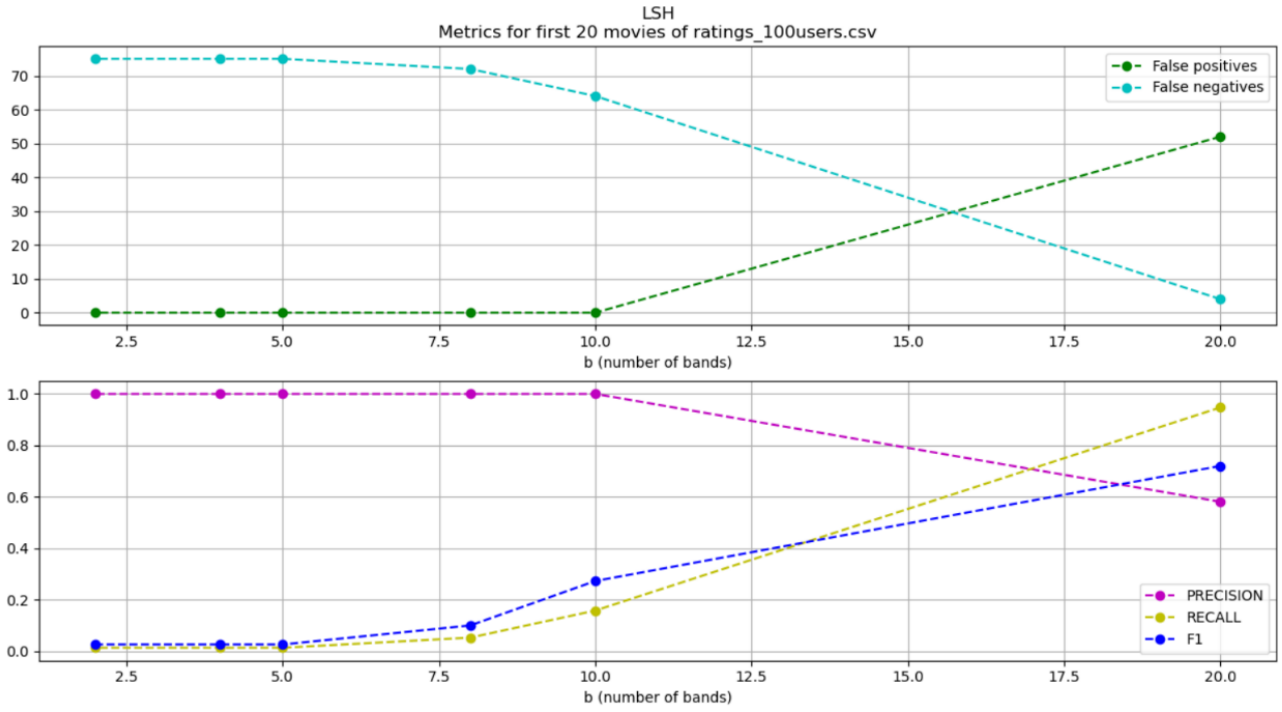
Στα σχήματα 5 και 6 που βρίσκονται παρακάτω φαίνονται τα αποτελέσματα της εκτέλεσης του αλγορίθμου LSHExperimentation( ) για τα 6 διαφορετικά ζευγάρια τιμών b, r.

Το κατώφλι που επιθυμούμε να πετύχουμε είναι s = 0.25. Επομένως, σύμφωνα με το ότι ισχύει s ~ (1/b)^(1/r), για n = 40 υπογραφές που έχει το μητρώο SIG, η τιμή του b για το συγκεκριμένο κατώφλι είναι b ~ 18,87. Γι' αυτό το λόγο, όπως βλέπουμε και στη γραφική παράσταση, η καλύτερη τιμή του F1 αντιστοιχεί στις 20 μπάντες (άρα στο συνδυασμό (r, b) = (2, 20)).

Μία ακόμη σημαντική παρατήρηση είναι ότι όσο περισσότερες γραμμές έχει μία μπάντα (και άρα όσο λιγότερες μπάντες έχουμε) τόσο περισσότερα είναι τα false-negatives και τόσο λιγότερα τα false-positives. Αυτό συμβαίνει γιατί, όσο μεγαλύτερες είναι οι υπογραφές της μπάντας, τόσο μικρότερη πιθανότητα έχουν δύο ταινίες να θεωρηθούν όμοιες περιορίζοντας έτσι τα false-positives. Αντίστοιχα, όσο λιγότερες γραμμές έχει μία μπάντα (και άρα όσο περισσότερες μπάντες έχουμε) τόσο αυξάνονται τα false-positives και μειώνονται τα false-negatives. Αυτά συμβαίνουν γιατί με μικρότερες υπογραφές δύο μη-όμοιες ταινίες έχουν μικρότερη πιθανότητα να κατακερματιστούν στον ίδιο κάδο. Τα παραπάνω γίνονται ξεκάθαρα παρατηρώντας το Σχήμα 6.

```
                #####
Experimentation: LSH
File: 'ratings_100users.csv'
First '20' movies are compared.
Threshold = 0.25, Number of signatures = 40
                #####

Number of relevant elements (ground truth) = 76

----- r = 2, b = 20 -----
Number of candidate pairs found = 124
True-positives = 72, False-positives = 52
True-negatives = 62, False-negatives = 4
PRECISION = 0.5806451612903226, RECALL = 0.9473684210526315, F1 = 0.72

----- r = 4, b = 10 -----
Number of candidate pairs found = 12
True-positives = 12, False-positives = 0
True-negatives = 114, False-negatives = 64
PRECISION = 1.0, RECALL = 0.15789473684210525, F1 = 0.2727272727272727

----- r = 5, b = 8 -----
Number of candidate pairs found = 4
True-positives = 4, False-positives = 0
True-negatives = 114, False-negatives = 72
PRECISION = 1.0, RECALL = 0.05263157894736842, F1 = 0.1

----- r = 8, b = 5 -----
Number of candidate pairs found = 1
True-positives = 1, False-positives = 0
True-negatives = 114, False-negatives = 75
PRECISION = 1.0, RECALL = 0.013157894736842105, F1 = 0.025974025974025976

----- r = 10, b = 4 -----
Number of candidate pairs found = 1
True-positives = 1, False-positives = 0
True-negatives = 114, False-negatives = 75
PRECISION = 1.0, RECALL = 0.013157894736842105, F1 = 0.025974025974025976

----- r = 20, b = 2 -----
Number of candidate pairs found = 1
True-positives = 1, False-positives = 0
True-negatives = 114, False-negatives = 75
PRECISION = 1.0, RECALL = 0.013157894736842105, F1 = 0.025974025974025976
```

Σχήμα 5: Αποτελέσματα του αλγορίθμου LSHExperimentation( ) για το αρχείο ratings_100users.csv



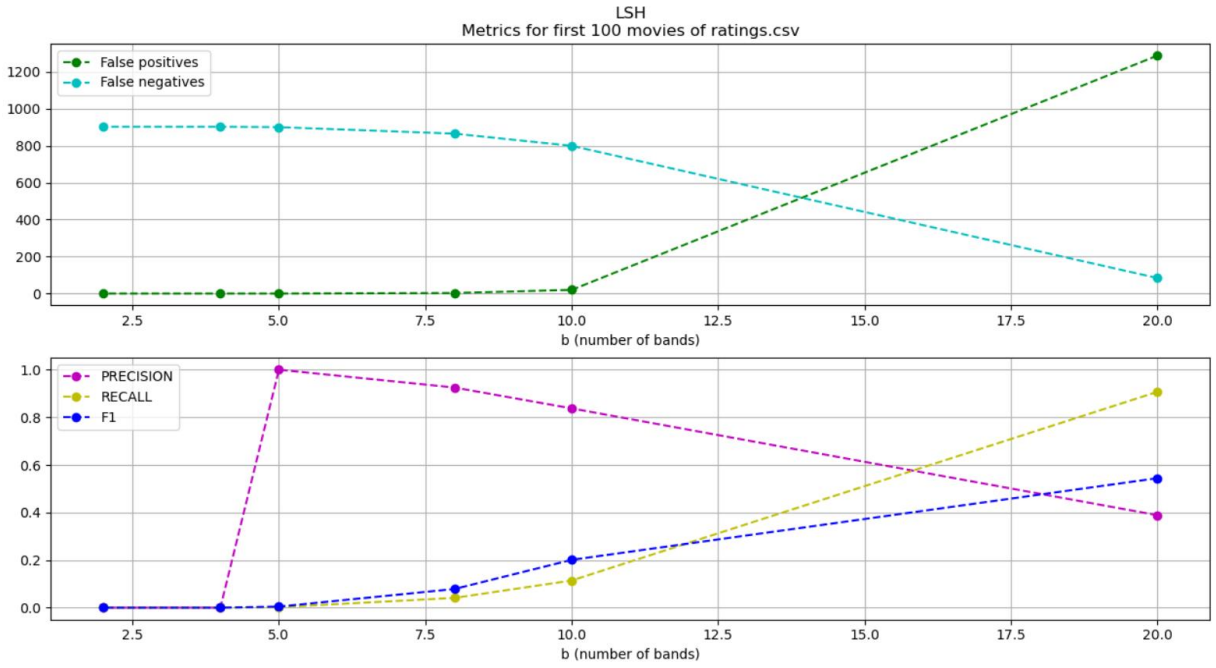
Σχήμα 6: Γραφικές παραστάσεις με βάση τα αποτελέσματα που φαίνονται στο Σχήμα 5

- *Ερώτημα (1ζ2), ratings.csv*

Όμοια συμπεριφέρονται τα αποτελέσματα και σε αυτήν την περίπτωση, όπως φαίνεται και στα σχήματα 7 και 8. Και εδώ η καλύτερη τιμή σημειώνεται για το συνδυασμό (r, b) = (2, 20) σύμφωνα με την τιμή του F1.

```
                #####
Experimentation: LSH
File: 'ratings.csv'
First '100' movies are compared.
Threshold = 0.25, Number of signatures = 40
                #####

Number of relevant elements (ground truth) = 902

----- r = 2, b = 20 -----
Number of candidate pairs found = 2105
True-positives = 818, False-positives = 1287
True-negatives = 2761, False-negatives = 84
PRECISION = 0.3885985748218527, RECALL = 0.9068736141906873, F1 = 0.5440638510143

----- r = 4, b = 10 -----
Number of candidate pairs found = 123
True-positives = 103, False-positives = 20
True-negatives = 4028, False-negatives = 799
PRECISION = 0.8373983739837398, RECALL = 0.11419068736141907, F1 = 0.20097560975609757

----- r = 5, b = 8 -----
Number of candidate pairs found = 40
True-positives = 37, False-positives = 3
True-negatives = 4045, False-negatives = 865
PRECISION = 0.925, RECALL = 0.041019955654102, F1 = 0.07855626326963908

----- r = 8, b = 5 -----
Number of candidate pairs found = 2
True-positives = 2, False-positives = 0
True-negatives = 4048, False-negatives = 900
PRECISION = 1.0, RECALL = 0.0022172949002217295, F1 = 0.004424778761061947

----- r = 10, b = 4 -----
Number of candidate pairs found = 0
True-positives = 0, False-positives = 0
True-negatives = 4048, False-negatives = 902
PRECISION = 0, RECALL = 0.0, F1 = 0

----- r = 20, b = 2 -----
Number of candidate pairs found = 0
True-positives = 0, False-positives = 0
True-negatives = 4048, False-negatives = 902
PRECISION = 0, RECALL = 0.0, F1 = 0
```

Σχήμα 7: Αποτελέσματα του αλγορίθμου LSHExperimentation( ) για το αρχείο ratings.csv



Σχήμα 8: Γραφικές παραστάσεις με βάση τα αποτελέσματα που φαίνονται στο Σχήμα 7