**ΔΡΙΒΑΣ ΠΑΝΑΓΙΩΤΗΣ 2690**

**ΓΙΑΝΝΗΣ ΜΠΑΡΖΑΣ 2765**

**1η ΑΝΑΘΕΣΗ**

**Ερώτημα 1a**

Για το userList εφόσον είναι ταξινομημένο το αρχείο με βάση το user\_id , κρατάμε την μεταβλητή previous\_user την οποία αρχικοποιήσαμε στην μονάδα και για κάθε γραμμή στο αρχείο παίρνουμε το user\_id της γραμμής και το συγκρίνουμε με το previous\_user.Κάθε φορά που βλέπουμε καινούριο user\_id , κάνουμε εισαγωγή στο λεξικό την λίστα του previous\_user με κλειδί το previous\_user και ξανα αρχικοποιούμε την λίστα και βάζουμε στο previous\_user νέο id.

Για τo movieMap κάθε φορά που βλέπουμε ένα καινούριο movie\_id αν δεν υπάρχει ήδη στο movieMap το προσθέτουμε σαν κλειδί με value ένα δικό μας id(το id αυξάνεται κάθε φορά κατά ένα ).

Για το movieList ελέγχουμε αν το movie\_id υπάρχει ήδη στο movieList ,αν υπάρχει εισάγουμε στην λίστα του movie\_id με τους χρήστες τον χρήστη. Διαφορετικά αν δεν υπάρχει αρχικοποιούμε μια λίστα ,βάζουμε μέσα σε αύτη το χρήστη(user\_id ) που βλεπουμε στη γραμμη του αρχειου που ειμαστε και την εισαγουμε στο λεξικο με κλειδι το movie\_id.

**Ερώτημα 1β**

Υλοποιήσαμε την συνάρτηση jaccardsimilarity ακριβώς όπως είναι στην εκφώνηση.

**Ερώτημα 1γ**

Υλοποιήσαμε την συνάρτηση MinHash :

Αρχικοποιήσαμε τον πίνακα Sig με μεγάλες τιμές και υπολογίσαμε τα n διαφορετικά permutation και τα αποθηκεύσαμε στον δισδιάστατο πίνακα hij. Σαρώνοντας το μητρώο κατά γραμμή , αν η θεση που είμαστε έχει 1 κάνουμε μια επανάληψη για κάθε n και αν η τιμή της permutation είναι μικρότερη από τη τιμή του αντίστοιχου στοιχείου του πίνακα Sig, ενημερώνουμε τη τιμή του πίνακα Sig με την αντιστοιχη τιμη του permutation.

**Ερώτημα 1δ**

Υλοποιήσαμε SingatureSimilarity στην οποία παίρνουμε σαν όρισμα δύο movie\_id , το n\_tonos και τον πίνακα Sig.Για τις πρώτες n\_tonos γραμμές του πίνακα Sig ελέγχουμε την ομοιότητα των στηλών (ταινιών) στον πίνακα Sig (βρισκουμε ποσα κελια είναι ομοια και τα διαιρουμε με το συνολικο αριθμο).

Υλοποιούμε την συνάρτηση get\_key που δίνουμε σαν όρισμα μια τιμή και το λέξικο στο οποίο ψάχνουμε το κλειδί για την τιμή αυτη, έτσι ώστε να μπορούμε να ανακτήσουμε το movie\_id του αρχείου από το τεχνητο movie\_id που δώσαμε στο movieΜap.

**Ερώτημα 1ε**

Υλοποιήσαμε την συνάτηση two\_digits η οποία έλεγχει αν το όρισμα που της παιρνάμε είναι μικροτερο ή ίσο του 9 , και αν είναι βαζει ένα 0 μπροστά διαφορετικα επιστρέφει τον αριθμό σε string.

Υλοποιήσαμε την συνάρτηση create\_pairs η οποία χωρίζει σε ζεύγη τα στοιχεία των κάδων που έχουν παραπάνω από 2 στοιχεία και τα επιστρέφει σε λίστα.

Υλοποιήσαμε την συνάρτηση LSH που παίρνει σαν όρισμα τα n,b,r,c,fisrt\_n\_movies(οι πρώτες movies στις οποίες εφαρμόζεται ο LSH). Για κάθε μπάντα αρχικοποιούμε το λεξικό με τους κάδους(bin\_dict) και για κάθε μία από τις πρώτες fisrt\_n\_movie ταινίες και για κάθε γραμμή της μπάντας παίρνουμε την αντίστοιχη τιμή στον πίνακα Sig(στο κελι) , καλούμε τη two\_digits και η επιστρεφόμενη τιμή εισάγεται σε μια προσωρινή λίστα (mylist) .Μόλις εισαχθούν όλες οι τιμες για κάθε γραμμή της μπάντας για τη ταινία που είμαστε μετατρέπουμε τη λίστα σε ένα integer.Καλούμε την hashfunction με τιμή αυτόν τον integer.Η επιστρεφόμενη τιμή αντιπροσωπεύει το κάδο που πέφτει κάθε φορα η τανια που επεξεργαζόμαστε και ελέγχουμε αν ο κάδος υπάρχει ήδη μέσα στο λεξικό.Αν υπάρχει ενημερώνουμε την λίστα προσθέτωντας την τανία.Αν δεν υπάρχει εισάγουμε το κάδο αυτόν στο λεξικό με τιμη μια λίστα με μονοδική την ταινία αυτή που επεξεργαζόμαστε.Μόλις τελειώνει η κάθε μπάντα πριν αρχικοποιήσουμε ξανά το λεξικό με τις μπάντες ελέγχουμε αν υπάρχουν κάδοι με παραπάνω από δύο ταινίες και αν υπάρχουν καλούμε την create\_pairs για κάθε τέτοιο κάδο .Εισάγουμε τα υποψήφια ζεύγη στην λίστα same\_movies αλλιώς εισάγουμε την λίστα του κάδου όπως είναι παλι στην λίστα same\_movies.

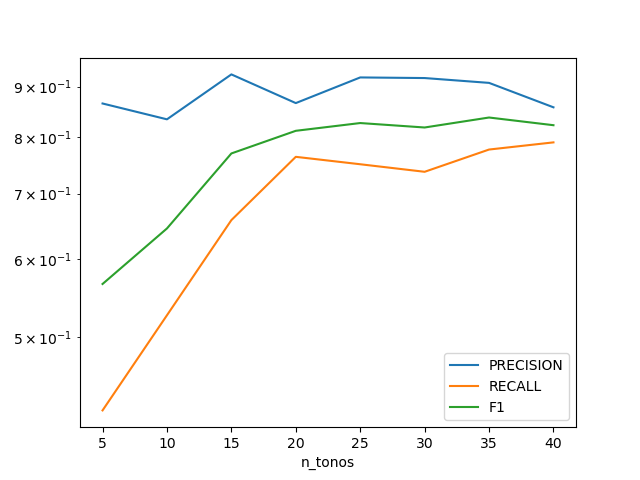
**Ερωτημα 1ζ1**

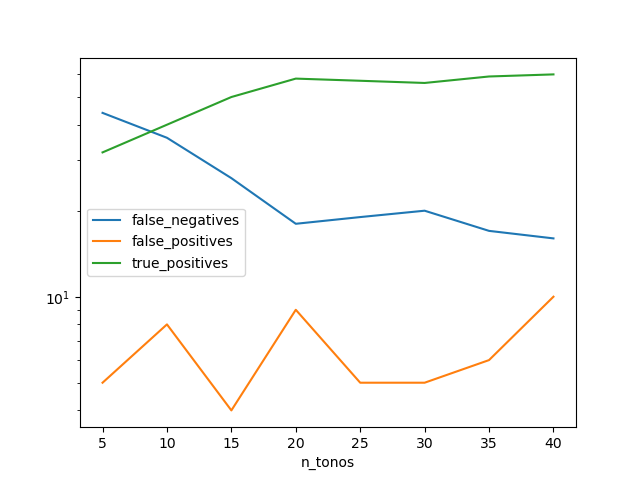
Για το αρχείο των 100 χρηστών :

Για κάθε ένα από τα 190 ζεύγη ταινιών δημιουργούμε τα λεξικά JSims\_greater\_s,

JSims\_less\_s όπου ανάλογα αν το jaccardSimilarity του ζεύγους είναι μεγαλυτερο ή μικρότερο του s το εισάγουμε στο αντίστοιχο λεξικό.Στην συνέχεια δημιουργούμε τα λεξικά SignSims\_greater\_s , SignSims\_less\_s όπου για κάθε τιμή του n και για κάθε ζεύγος ταινιών (190) για τις πρώτες 20 ταινιες, υπολόγιζουμε το signatureSimilarity του κάθε ζεύγους και ανάλογα αν είναι μεγαλύτερο ή μικροτερο του s το βάζουμε σε ένα από τα λεξικά που προαναφέραμε.Σε όλα τα παραπάνω λεξικά το κλειδί είναι ένα αλφαριθμητικό των δύο τεχνητών movie\_id χωρισμένα με ‘,’ και σαν τιμή την similarity τους.

Στην συνέχεια δημιουργούμε το λεξικό metrics όπου περιέχει τα κλειδία false\_positives , false\_negatives, true\_positives, PRECISION, RECALL, F1 με τιμές μια λίστα για το καθένα 8 τιμών όσες είναι και οι διαφορετικές τιμές του n\_tonos.Για κάθε μια τιμή του n\_tonos υπολογίζουμε τις παραπάνω τιμές και τις εισάγουμε στην κατάλληλη λίστα του metrics .Τέλος εκτυπώνουμε τα αποτελέσματα στην οθόνη με τη χρηση toυ DataFrame metrics\_dataframe και φτιάχνουμε τα διαγράμματα.





Παρατηρούμε στο πρωτο διάγραμμα ότι για τα RECALL και F1 καθώς αυξάνεται η τιμή του n\_tonos

αυξάνονται και αυτά και για το PRECISION στην αρχή παρατηρούμε αυξομειώσεις που σταθεροποιείται κοντα στο 0,9. ΟΙ παραπάνω παρατηρήσεις είναι λογικές καθώς το RECALL είναι το ποσοστό των όμοιων ζευγαριών αυτών που πήραμε σε σχέση με τον αριθμο όλων των όμοιων που υπάρχουν συνολικα ενώ το PRECISION το ποσοστό των όμοιων ζευγαριών που πήραμε σε σχέση με των αριθμο αυτων που πηραμε .Καθώς αυξάνεται το n\_tonos αυξάνεται το true\_positive γιατί στη συνάρτηση SignatureSimilarity αυξάνεται κάθε φορά και το πλήθος των υπογραφών που ελέγχουμε για να δούμε αν δύο ταινίες είναι ίδιες , με συνέπεια να έχουμε μεγαλύτερο ποσοστο επιτυχίας.

Στο δεύτερο διάγραμμα καθώς το n\_tonos αυξάνεται ,παρατηρούμε ότι το true\_positive αυξάνεται , μειώνονται τα false\_negatives και αυξoμειώνονται τα false\_positives αλλα διατηρούνται σε χαμηλές τιμές. Αυτές οι μεταβολές είναι λογικές και η αιτία είναι το γεγονός ότι καθώς αυξάνεται το n\_tonos αυξάνεται το accurancy SignatureSimilarity και πλησιαζει ολο και περισσοτερο το accurancy του jaccardSimilarity .Με απλά λόγια οσο περισσότερες υπογραφές ελένχουμε τοσο το αποτέλεσμα του SignatureSimilarity πλησιαζει το jaccardSimilarity και αρα αυξανετε η επίδοση του αλγορίθμου.Το ιδιο ισχυει και για το πρώτο διαγραμμα καθως ολες αυτές οι μετρικες είναι αλληλοεξαρτόμενες.

**Ερωτημα 1ζ2**

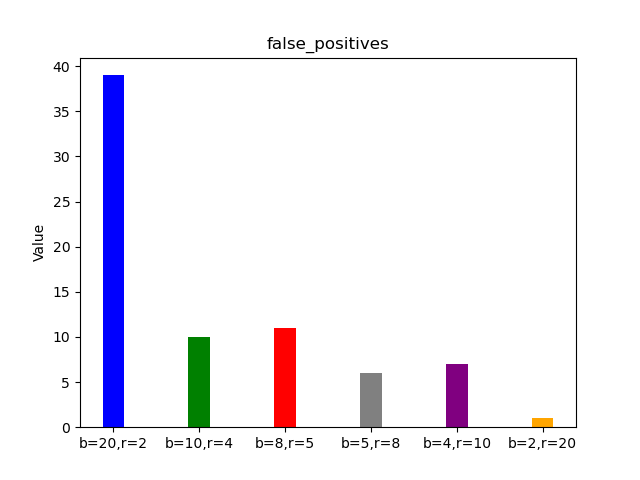
Αρχικα για κάθε συνδιασμο b,r καλουμε την LSH και υπολογιζουμε τα κλειδία false\_positives , false\_negatives, true\_positives, PRECISION, RECALL, F1.

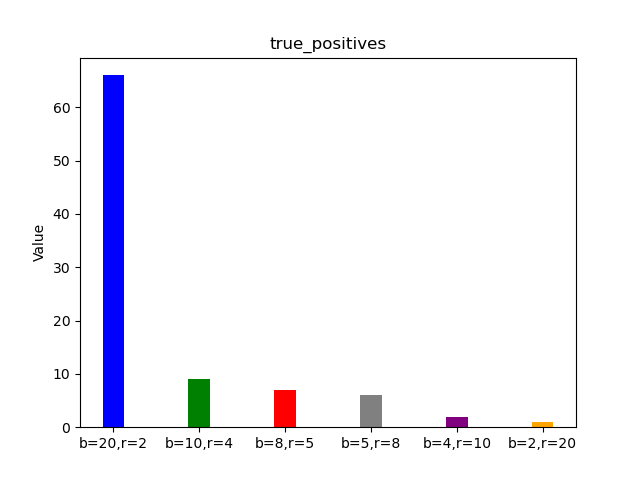
Εδώ δημιουργούμε το λεξικό metrics2 όπου περιέχει τα κλειδία false\_positives , false\_negatives, true\_positives, PRECISION, RECALL, F1 με τιμές μια λίστα για το καθένα 8 τιμών όσες είναι και οι διαφορετικές τιμές του n\_tonos.Για κάθε μια τιμή του n\_tonos υπολογίζουμε τις παραπάνω τιμές και τις εισάγουμε στην κατάλληλη λίστα του metrics .Τέλος εκτυπώνουμε τα αποτελέσματα στην οθόνη με τη χρηση toυ DataFrame metrics\_dataframe2 και φτιάχνουμε τα διαγράμματα.

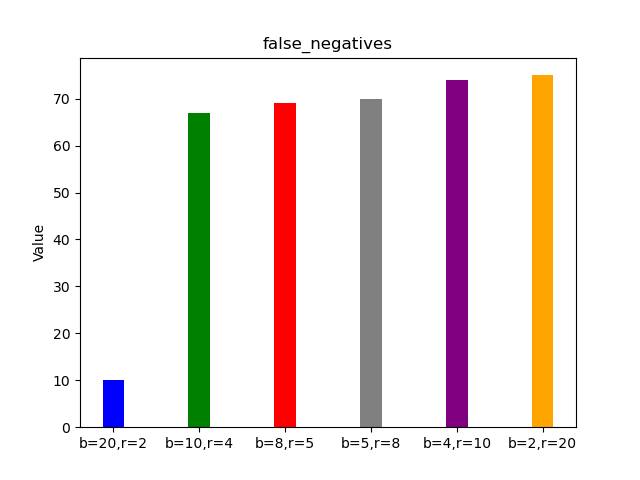
Στο αρχειο final\_results(100).csv: αποθηκευουμε τους πινακες με τις τιμες και των 2 πειραματικων.

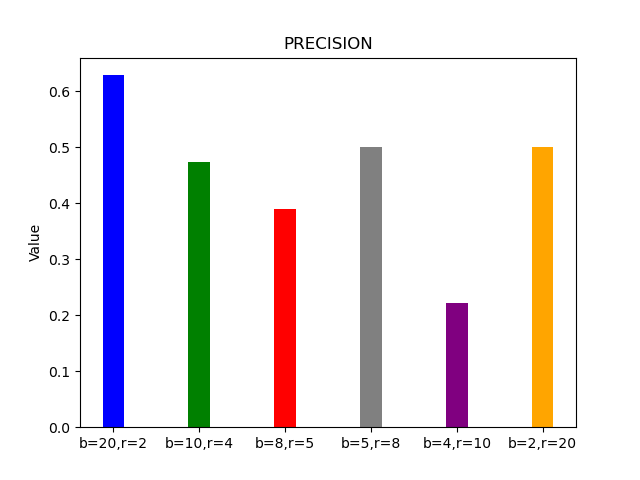
Στο αρχειο sig\_table(100).csv: αποθηκευουμε το μητρωο υπογραφών.

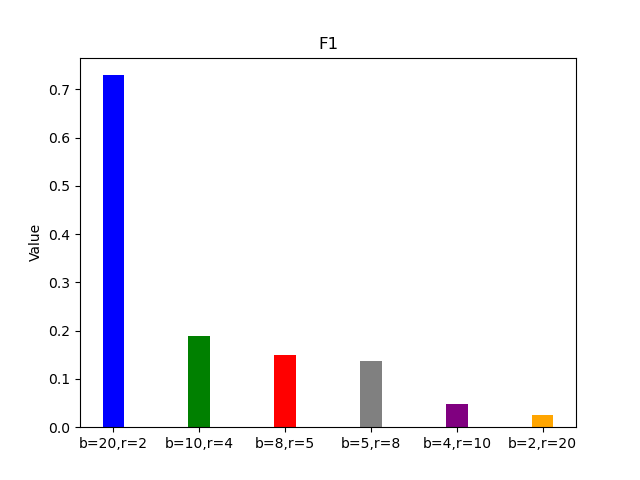
Στο αρχειο mitrwa(userList,movieList,movieMap)(100 users,s=0.25).csv: αποθηκευουμε τα λεξικα userList,movieList,movieMap.

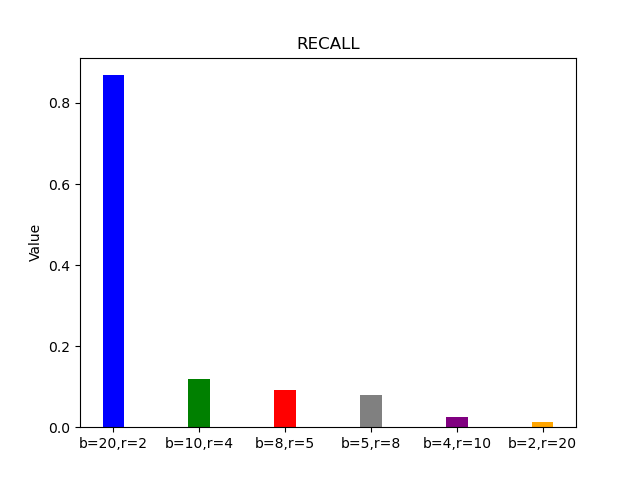












false\_positives , false\_negatives, true\_positives, PRECISION, RECALL, F1

Το διάγραμμα του false\_positives εχει φθήνουσα πορεια καθως μικραίνει ο αριθμός των μπαντών αυτό ευθύνεται στο γεγονός ότι ο αλγόριθμος διαλέγει λιγοτερα ζευγάρια καθώς ολο και μικραίνει η πιθανοτητα 2 ταινιες να συμπεσουν στον ιδιο καδο.Αυτη είναι μια θετικη επιδραση που εχουμε.

Στο διαγραμμα του true\_positives φθήνει γιατι παλι λογω της μείωσης των μπαντών o LSH επιστρέφει λιγοτερες τιμες και αρα βρίσκει λιγοτερα true\_positives καθως και παλι μικραινει η πιθανοτητα 2 ταινιες να συμπεσουν στον ιδιο καδο και αρα και το true\_positives.

Στο διαγραμμα του false\_negatives παρατηρουμε μια αυξηση καθως μειωνεται το b.Αυτο συμβαινει γιατι καθως επιλεγουμε λιγοτερα αποτελεσματα και αρα μικραινει το true\_positive ,αυξανονται τα ζευγαρια που ηταν ομοια και δεν τα πηραμε.Τα ομοια ζευγη(relevant) είναι παντα ιδια και ισα με τα true\_positives + false\_negatives. Αφου τα ομοια ζευγη παραμενουν ιδια και τα true\_positives μειωνωνται τοτε είναι λογικο τα false\_negatives να αυξανονται .

Στο διαγραμμα του RECALL παρατηρειται ότι υπαρχει μειωση καθως μειώνετε το b

Παλι αυτό συμβαινει γιατι το RECALL είναι το ποσοστό των όμοιων ζευγαριών που πήραμε σε σχέση με τον αριθμο όλων των όμοιων που υπάρχουν συνολικα . Αρα είναι λογικο που μειωνεται γιατι μειωνονται και τα true\_positives που είναι αυτά που πηραμε και είναι ομοια ενώ δεν μειωνετε ο αριθμος όλων των όμοιων που υπάρχουν συνολικα.

Στο διαγραμμα του PRECISION εχει καποιες αυξομειωσεις καθως μικραινει το b αλλα παρολα αυτά οι τιμες του είναι ψηλες.Το PRECISION είναι το ποσοστό των όμοιων ζευγαριών που πήραμε σε σχέση με τον συνολικο αριθμο των ζευγαριων που πηραμε.Αυτο σημαινει ότι παρολου που μικραινει το true\_positives μικραινει και ο παρανομαστης του PRECISION ( δηλαδη ο αριθμος των ζευγαριων που επιλεγουμε) Ετσι το ποσοστο των ζευγαριων που περνουμε και είναι ομοια σε σχεση με το συνολικο αριθμο των ζευγαριων που περνουμε κρατιεται σε ψηλες τιμες (το κλασμα).

Στο διαγραμμα του F1 μειώνεται καθως μειωνεται το b λογω του ότι μειωνεται το RECALL και λογω του πολλαπλασιασμου στον αριθμητη του τυπου ο αριθμητης μικραινει πιο γρηγορα από τον παρανομαστη και ετσι εχουμε μειωση.Ειναι λογικο γιατι όταν μειωνεται ο αριθμος των ζευγαριων που επιλεγουμε μειωνεται και το recall το οποιο είναι αναλογο του F1.