ΟΜΟΙΟΤΗΤΑ ΑΝΤΙΚΕΙΜΕΝΩΝ

**Πληροφορίες για το αρχείο:**

Το αρχείο processing.py τρέχει δίνοντας στο τερματικό το όνομα του ακολουθούμενο από το όνομα ενός αρχείου αξιολογήσεων ταινιών από τους χρήστες (πχ, το αρχείο ratings\_100users.csv για 100 χρήστες, ή το αρχείο ratings.csv για 605 χρήστες). Στη main του προγράμματος γίνεται έλεγχος που προσδιορίζει ποια συνάρτηση θα καλεστεί ανάλογα με τα αρχεία. Για το αρχείο ratings\_100users η αντίστοιχη συνάρτηση που εκτελείται είναι η test100users και για το αρχείο ratings η συνάρτηση test605users που εκτελεί τα ίδια πειράματα για μεγαλύτερο όγκο δεδομένων. Επίσης στο πρόγραμμά μας χρησιμοποιούμε τις βιβλιοθήκες sys, pandas, random, numpy, csv και matplotlib.pyplot. Επιπλέον βάλαμε σε σχόλια τον κώδικα που εμφανίζει και αποθηκεύει τις γραφικές παραστάσεις. ( Για να δημιουργηθούν οι γραφικές παραστάσεις πρέπει να αφαιρεθούν τα σχόλια στις γραμμες 69-78, 123-132, 189-198 και 243-252).

**(1α) Προεπεξεργασία:**

Η συνάρτηση calculateLists(data,filename) ανάλογα με το αρχείο που δίνεται ως είσοδος στο τερματικό (ratings\_100users.csv ή ratings.csv) δημιουργεί σύμφωνα με τις οδηγίες της εκφώνησης :

* Ένα λεξικό λιστών userList όπου κάθε γραμμή userList[userId] αντιστοιχεί και σε έναν διαφορετικό χρήστη userId, ενώ η τιμή της είναι μια λίστα με τις όλες ταινίες (σε αύξουσα σειρά των movieId πεδίων) που έχει (αξιολογήσει, άρα και) δει ο χρήστης.
* Ένα λεξικό movieMap που αντιστοιχεί τις ταινίες σε μοναδικούς αριθμούς από το {1,…,Ν}.
* Ένα λεξικό λιστών movieList, όπου κάθε γραμμή movieList[movieId] αντιστοιχεί και σε μια διαφορετική ταινία movieId, ενώ η τιμή της είναι μια λίστα με τα userId όλων των χρηστών που την έχουν παρακολουθήσει.

Τα παραπάνω κατα την εκτέλεση του προγράμματος αποθηκεύονται το καθένα σε ένα αρχείο με το κατάλληλο όνομα.

**(1β) jaccardSimilarity(movieId1,movieId2)**

Η συνάρτηση calculateJaccardSimilarity(movieList1,movieList2) δέχεται σαν όρισμα 2 στήλες από το μητρώο Μ, οι οποίες αντιστοιχούν σε 2 ταινίες και υπολογίζει την Jaccard ομοιότητα τους σύμφωνα με το τύπο που δίνεται στην εκφώνηση.

**(1γ) Δημιουργία Min-Hash Υπογραφών για τις ταινίες**

Η συνάρτηση createRandomHashFunction χρησιμοποιείται ως συνάρτηση κατακερματισμού για την παρακάτω συνάρτηση. Η συνάρτηση αυτή είναι μια τυχαία γραμμική συνάρτηση στην modulo αριθμητική.

Η συνάρτηση calculateMinHash χρησιμοποιείται για τη δημιουργία του μητρώου SIG με τις υπογραφές όλων των ταινιών. Ο υπολογισμός των υπογραφών έγινε με ένα πέρασμα του μητρώου των εγγράφων. Για να επιτευχθεί αυτό χρησιμοποιήθηκαν τόσες συναρτήσεις κατακερματισμού όσες και ο αριθμός των στηλών. Επίσης κατά την εκτέλεση, το μητρώο με τις υπογραφές αποθηκεύεται σε αρχείο με κατάλληλο όνομα.

**(1δ) Υπολογισμός Signature-Ομοιότητας μεταξύ των ταινιών**

Με τη συνάρτηση signatureSimilarity συγκρίνεται η ομοιότητα δύο διανυσμάτων υπογραφών όπως περιγράφεται στην εκφώνηση χρησιμοποιώντας την Jaccard ομοιότητα.

**(1ε) Δημιουργία Υποψηφίων Ζευγών με τη μέθοδο Locality-Sensitive Hashing**

Η συνάρτηση LSH τοποθετεί σε κάδους όλες τις ταινίες ανάλογα με την τιμή της υπογραφής σε κάθε μπάντα υπογραφών. Χρησιμοποιείται μια συνάρτηση κατακερματισμού για τις ταινίες. Τοποθετεί στον ίδιο κάδο δύο ταινίες αν κατακερματίζονται στον ίδιο κάδο. Αν ένας κάδος είναι αρχικά άδειος απλά προσθέτουμε την ταινία σε αυτόν. Αν έχει ήδη στοιχεία μέσα, τότε προσθέτουμε την ταινία στον κάδο μαζί με τις ήδη υπάρχουσες ταινίες. Αν ένας κάδος έχει δύο ή παραπάνω ταινίες τότε τα υποψήφια ζεύγη είναι όλες οι πιθανές δυάδες που μπορούν να προκύψουν από το σύνολο του κάδου αυτού. Στο τέλος κάθε μπάντας αδειάζουμε τους κάδους για την εισαγωγή των καινούργιων δεδομένων. Στο τέλος πριν επιστραφούν τα πιθανά ζεύγη γίνεται ένωση των διπλότυπων στοιχείων για να διαγραφούν τα διπλά, τριπλά κ.ο.κ.

**(1ζ) Πειραματική Αξιολόγηση**

**(1ζ1) Πειραματισμός για Min-Hashing**

ΓΙα το αρχείο ratings\_100users δημιουργήσαμε όλα τα πιθανά ζεύγη για

τις ταινίες με τα πρώτα 20 προσδιοριστικά. Για κάθε ένα από αυτά

υπολογίσαμε τη Jaccard ομοιότητα. Ως true-positives θεωρήσαμε όσα

ζεύγη ταινιών είχαν ομοιότητα τουλάχιστον 0.25.

Στη συνέχεια υπολογίσαμε το μητρώο υπογραφών SIG με τη συνάρτηση

MINHASH. Χρησιμοποιώντας το μητρώο υπογραφών εκτελέσαμε την

παρακάτω διαδικασία 8 φορές (με διαφορετικό πλήθος γραμμών από το

μητρώο n’= 5, 10, 15, 20, 25, 30, 35, 40):

Θεωρούμε ως false-positives τα ζεύγη των οποίων η ομοιότητα

υπογραφών ήταν τουλάχιστον s =0.25, αλλά με Jaccard-ομοιότητα είναι

μικρότερη του s. Επιπλέον θεωρούμε ως false-negatives τα ζεύγη με

Jaccard ομοιότητα τουλάχιστον s, αλλά με ομοιότητα υπογραφών

μικρότερη του s. Τέλος υπολογίσαμε τις τιμές των PRECISION, RECALL, F1

σύμφωνα με τον τύπο που δίνεται σε κάθε περίπτωση.

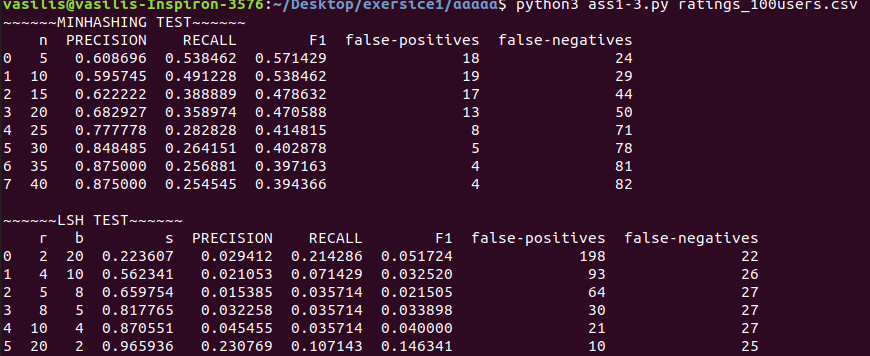
Παρακάτω παρατίθενται οι μετρικές και οι γραφικές παραστάσεις για μια

τυχαία εκτέλεση του προγράμματος. Για κάθε εκτέλεση, λόγω της

τυχαιότητας παίρνουμε διαφορετικά αποτελέσματα αλλά με μικρές

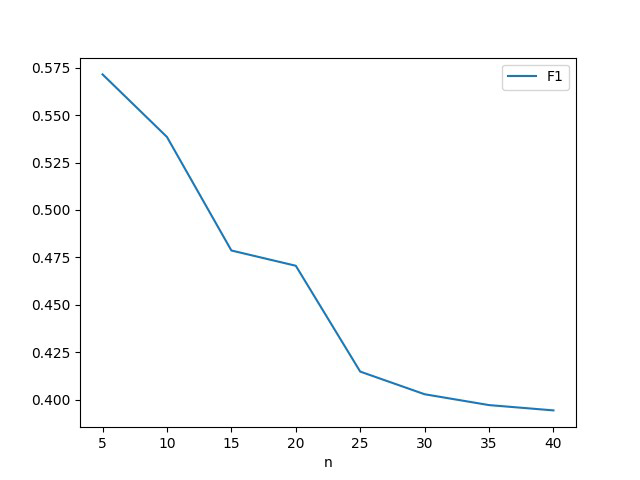
αποκλίσεις.

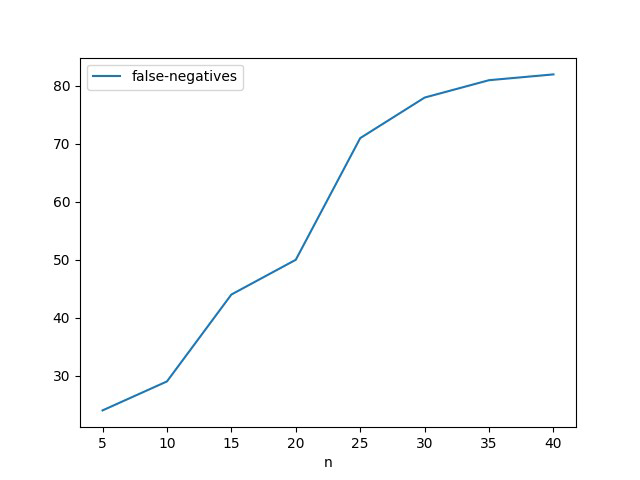
Αποτελέσματα για το αρχείο ratings\_100users:

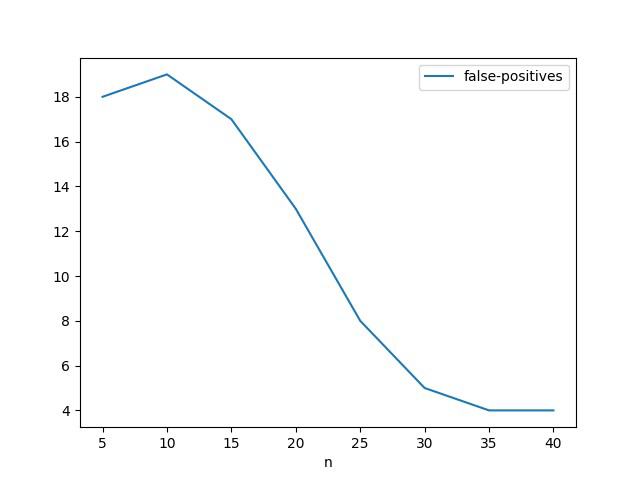


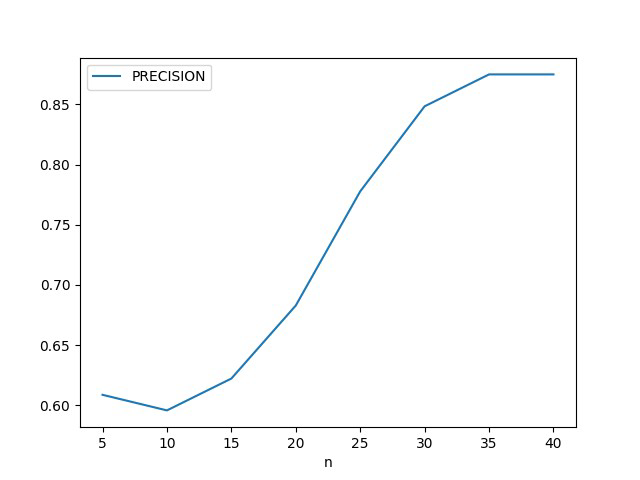
Παρατηρούμε ότι όσο περισσότερη πληροφορία έχουμε από το μητρώο υπογραφών, δηλαδή όσο αυξάνεται το πλήθος των γραμμών που χρησιμοποιούμε από το μητρώο τόσο μειώνεται το πλήθος των false-positives αλλά αυξάνεται το πλήθος των false-negatives. Συγκεκριμένα, το πλήθος των false-positives μειώνεται όσο αυξάνεται το πλήθος των γραμμών από το μητρώο υπογραφών. H ακρίβεια (PRECISION) αυξάνεται όσο αυξάνεται το πλήθος των γραμμών και παραμένει σταθερή για n=35,40 το οποίο μας δείχνει ότι δεν αλλάζει η ακρίβεια για 35 γραμμές και πάνω. Η ανάκληση μας δείχνει το πλήθος των σωστών ζευγών που ανακτώνται με επιτυχία. Παρόλο που όσο αυξάνεται το πλήθος των γραμμών μειώνεται το πλήθος των false-positives, το πλήθος των false-negative αυξάνεται όσο αυξάνεται το πλήθος των γραμμών του μητρώου. Αυτό επηρεάζει την τιμή της ανάκλησης με αποτέλεσμα να είναι καθοδική. Το αρμονικό μέσο F1 των PRECISION και RECALL είναι και αυτό καθοδικό όσο αυξάνεται το πλήθος των γραμμών του μητρώου. Επιπλέον για όλες τις μετρικές από το n=30 και πάνω οι τιμές παραμένουν παρόμοιες.

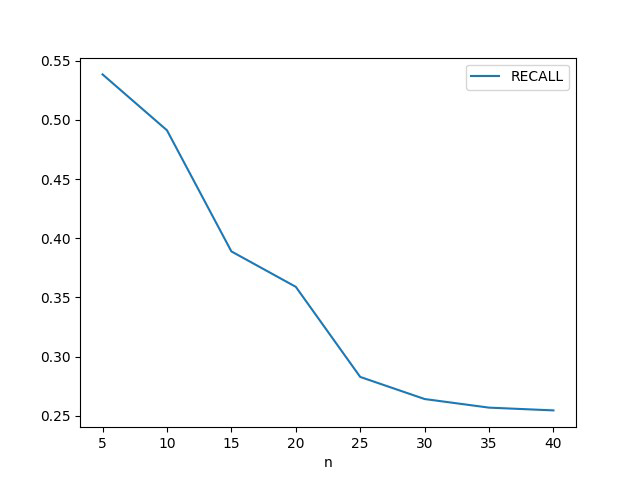
Παρακάτω παρατίθενται οι γραφικές παραστάσεις των μετρικών που περιγράφονται στις παραπάνω παρατηρήσεις.









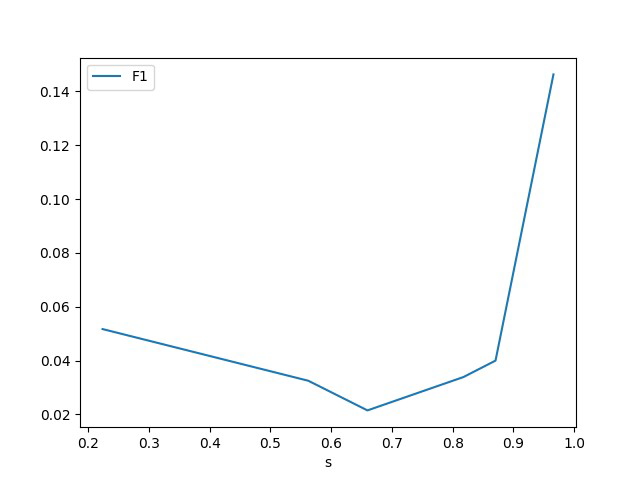


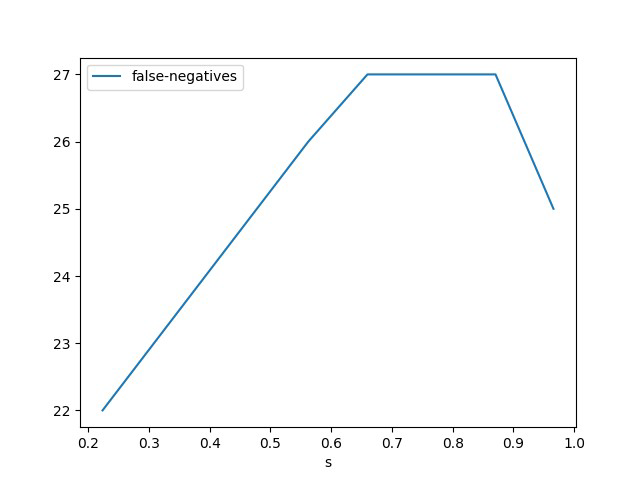
**(1ζ1) Πειραματισμός για LSH**

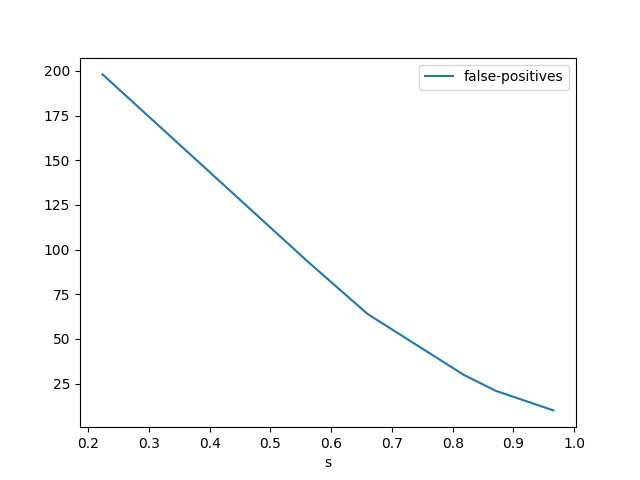
Για την περίπτωση των n’ = 40 υπογραφών, χωρίσαμε το μητρώο υπογραφών SIG σε b μπάντες, με r συναρτήσεις κατακερματισμού σε κάθε μπάντα και εκτελέσαμε τη ρουτίνα LSH. Εντοπίσουμε υποψήφια ζεύγη ταινιών με Jaccard ομοιότητα τουλάχιστον s = 0.25 (από το παραπάνω πείραμα) .

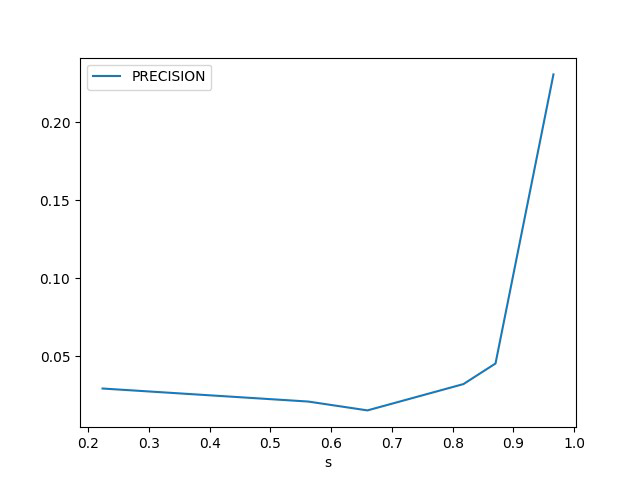
Δοκιμάσαμε τις τιμές (r=2, b = 20), (r=4, b=10), (r=5, b=8), (r = 8, b =5), (r = 10, b = 4), και (r = 20, b = 2). Βρήκαμε τα true-positives ως την τομή των ζευγαριών που επέστρεψε η ρουτίνα LSH και των true-positives από το παραπάνω πείραμα.

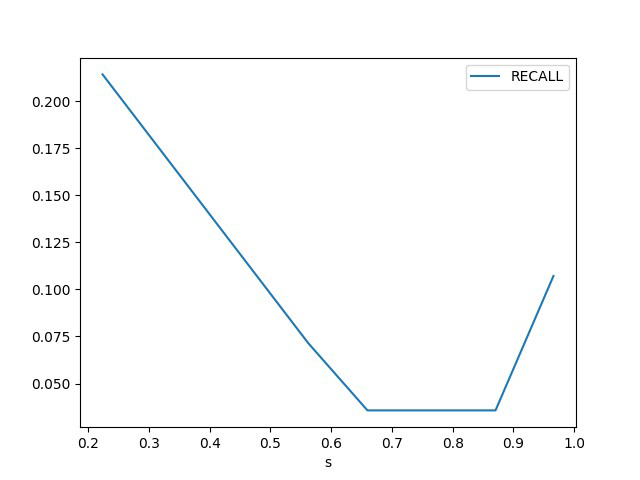
Για το πείραμα με τον LSH παίρνουμε καλύτερα αποτελέσματα. Παρατηρούμε και εδώ πτώση των τιμών των false-positive αλλα και των τιμών των false-negatives είτε να μειώνονται είται να παραμένουν σταθερές. Οι τιμές των PRECISION, RECALL και F1 δεν έχουν καθαρά ανοδική ή καθοδική κλίση. To RECALL μειώνεται όσο αυξάνεται το πλήθος των false-negatives και όμοια επηρεάζεται η τιμή της μετρικής F1. Παρατηρούμε ότι ο συνδυασμός r=10, b=4 δίνει τα καλύτερα αποτελέσματα γιατί το threshold s> ⅘ δηλαδή μεγαλύτερο από 80% ομοιότητα σε αντίθεση με τις άλλες τιμές των r, b όπου το s έχει μικρότερη ομοιότητα από 80%. Όποτε τα ζευγάρια που θα βρεθούν θα είναι πιο πιθανόν να ανήκουν σε true-positives αφού το κατώφλι είναι μεγάλο και άρα πιο πιθανό να ισχύει.











**(1ζ2) Πειραματισμός για Min-Hashing**

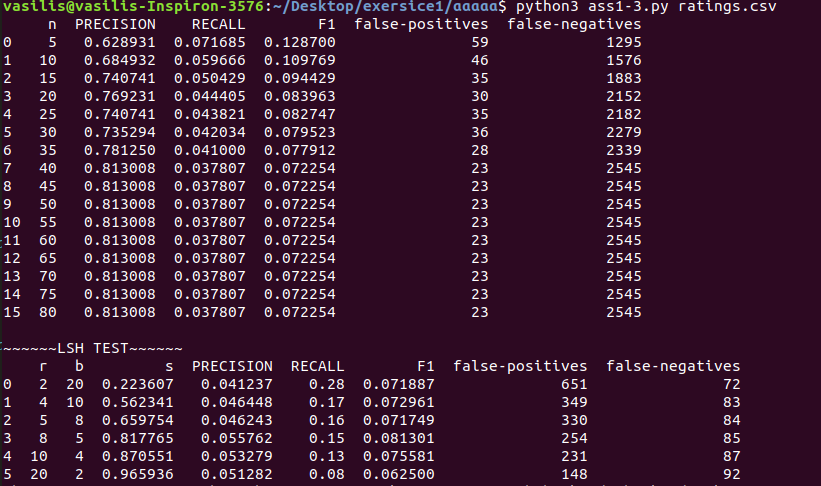
ΓΙα το αρχείο ratings όμοια δημιουργήσαμε όλα τα πιθανά ζεύγη για τις ταινίες με τα πρώτα 100 προσδιοριστικά ταινιών. Για κάθε ένα από αυτά υπολογίσαμε τη Jaccard ομοιότητα. Ως true-positives θεωρήσαμε όσα ζεύγη ταινιών είχαν ομοιότητα τουλάχιστον 0.25.

Στη συνέχεια υπολογίσαμε το μητρώο υπογραφών SIG με τη συνάρτηση MINHASH. Χρησιμοποιώντας το μητρώο υπογραφών εκτελέσαμε την παρακάτω διαδικασία 8 φορές (με διαφορετικό πλήθος γραμμών από το μητρώο n’= 5, 10, 15, 20, 25, 30, 35, 40, 50, 55, 60, 65, 70, 75, 80):

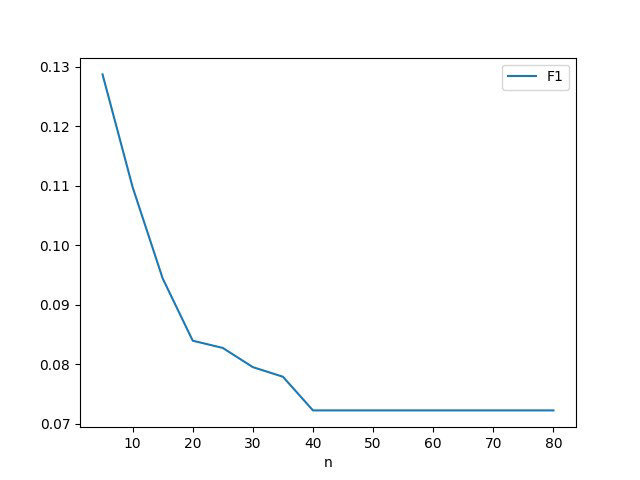
Θεωρούμε ως false-positives τα ζεύγη των οποίων η ομοιότητα υπογραφών ήταν τουλάχιστον s=0.25, αλλά με Jaccard-ομοιότητα είναι μικρότερη του s. Επιπλέον θεωρούμε ως false negatives τα ζεύγη με Jaccard ομοιότητα τουλάχιστον s, αλλά με ομοιότητα υπογραφών μικρότερη του s. Τέλος υπολογίσαμε τις τιμές των PRECISION, RECALL, F1 σύμφωνα με τον τύπο που δίνεται σε κάθε περίπτωση.

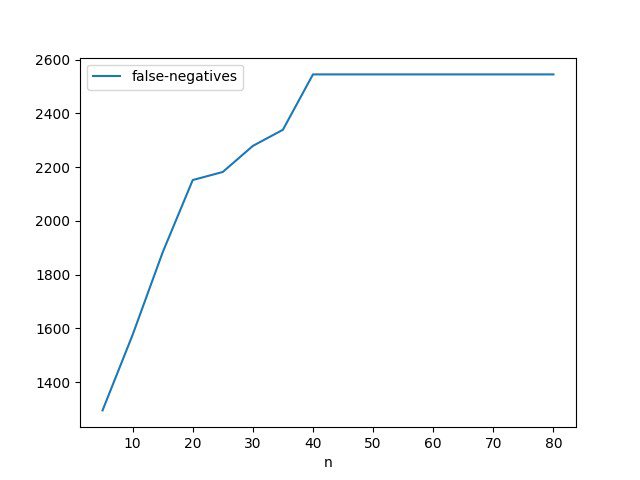
Παρακάτω παρατίθενται οι μετρικές και οι γραφικές παραστάσεις για μια τυχαία εκτέλεση του προγράμματος. Για κάθε εκτέλεση, λόγω της τυχαιότητας παίρνουμε διαφορετικά αποτελέσματα αλλά με μικρές αποκλίσεις.

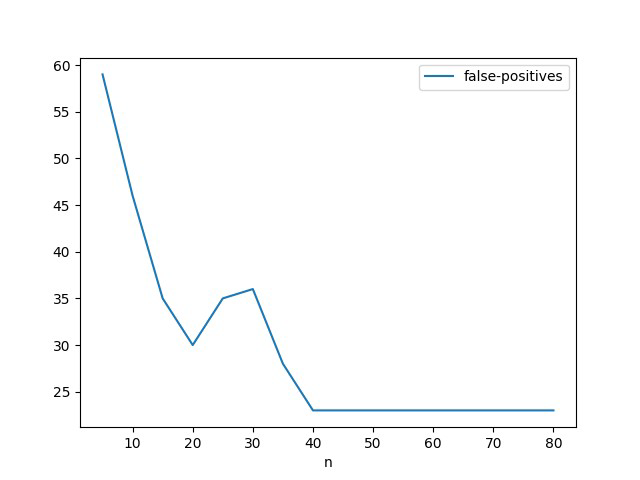
Αποτελέσματα για το αρχείο ratings:

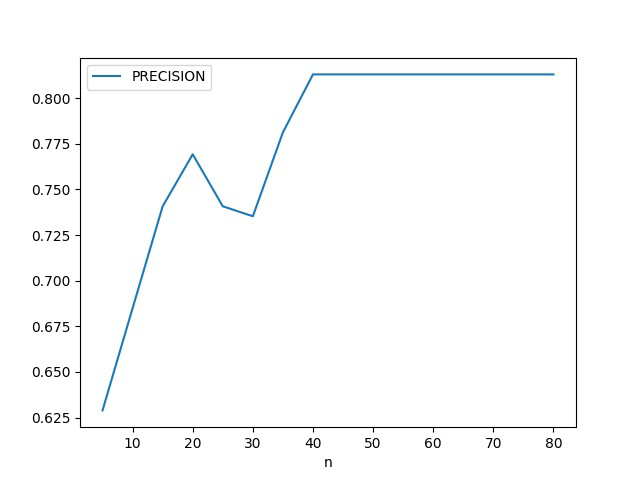


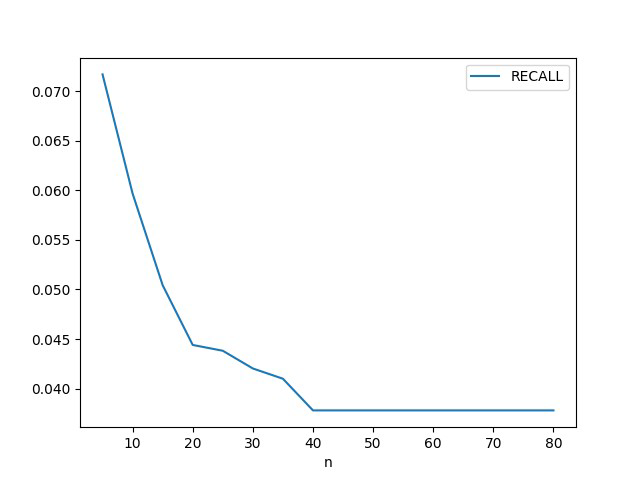
Παρατηρούμε ότι όσο περισσότερη πληροφορία έχουμε από το μητρώο υπογραφών, δηλαδή όσο αυξάνεται το πλήθος των γραμμών που χρησιμοποιούμε από το μητρώο τόσο μειώνεται το πλήθος των false-positives αλλά αυξάνεται το πλήθος των false-negatives. Συγκεκριμένα, το πλήθος των false-positives μειώνεται όσο αυξάνεται το πλήθος των γραμμών από το μητρώο υπογραφών. H ακρίβεια (PRECISION) αυξάνεται όσο αυξάνεται το πλήθος των γραμμών. Η ανάκληση μας δείχνει το πλήθος των σωστών ζευγών που ανακτώνται με επιτυχία. Παρόλο που όσο αυξάνεται το πλήθος των γραμμών μειώνεται το πλήθος των false-positives, το πλήθος των false-negative αυξάνεται όσο αυξάνεται το πλήθος των γραμμών του μητρώου. Αυτό επηρεάζει την τιμή της ανάκλησης με αποτέλεσμα να είναι καθοδική. Το αρμονικό μέσο F1 των PRECISION και RECALL είναι και αυτό καθοδικό όσο αυξάνεται το πλήθος των γραμμών του μητρώου. Επιπλέον για όλες τις μετρικές από το n=40 και πάνω οι τιμές παραμένουν παρόμοιες το οποίο μας δείχνει ότι από τις 40 γραμμές και πάνω δεν επηρεάζονται τα αποτελέσματα.











**(1ζ2) Πειραματισμός για LSH**

Για την περίπτωση των n’ = 40 υπογραφών, χωρίσαμε το μητρώο υπογραφών SIG σε b μπάντες, με r συναρτήσεις κατακερματισμού σε κάθε μπάντα και εκτελέσαμε τη ρουτίνα LSH. Εντοπίσουμε υποψήφια ζεύγη ταινιών με Jaccard ομοιότητα τουλάχιστον s = 0.25 (από το παραπάνω πείραμα) .

Δοκιμάσαμε τις τιμές (r=2, b = 20), (r=4, b=10), (r=5, b=8), (r = 8, b = 5), (r = 10, b = 4), και (r = 20, b = 2). Βρήκαμε τα true-positives ως την τομή των ζευγαριών που επέστρεψε η ρουτίνα LSH και των true-positives από το παραπάνω πείραμα.

Για το πείραμα με τον LSH παίρνουμε καλύτερα αποτελέσματα. Παρατηρούμε και εδώ πτώση των τιμών των false-positive αλλα και των τιμών των false-negatives είτε να μειώνονται είται να παραμένουν σταθερές. Επίσης οι τιμές των μετρικών PRECISION, RECALL και F1 παραμένουν και αυτές σχετικά ίδιες χωρίς μεγάλες αποκλίσεις.

Παρατηρούμε ότι ο συνδυασμός r=10, b=4 δίνει τα καλύτερα αποτελέσματα γιατί το threshold s> ⅘ δηλαδή μεγαλύτερη από 80% ομοιότητα σε αντίθεση με τις άλλες τιμές των r, b όπου το s έχει μικρότερη ομοιότητα από 80%. Όποτε τα ζευγάρια που θα βρεθούν θα είναι πιο πιθανόν να ανήκουν σε true-positives αφού το κατώφλι είναι μεγάλο και άρα πιο πιθανό να ισχύει.

