ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ,

ΣΥΧΝΑ ΣΥΝΟΛΑ ΚΑΙ ΡΟΕΣ ΔΕΔΟΜΕΝΩΝ

**Περιγραφή Μεθόδων:**

**(1) Ρουτίνα CreateMovieBaskets :**

Η συγκεκριμένη ρουτίνα παίρνει σαν όρισμα το όνομα του αρχείου και το minScore που δίνεται σαν όρισμα από το χρήστη. Αρχικά, διαβάζει το αρχείο και αποθηκεύει σε δύο dictionaries τις ταινίες που έχει δει ο κάθε χρήστης και τις αξιολογήσεις αντίστοιχα. Τα dictionaries είναι της μορφής {χρήστης:[ταινίες]} και {χρήστης:[αξιολογήσεις]} και κρατούν τα δεδομένα ταξινομημένα όπως αναφέρεται στην εκφώνηση. Στην συνέχεια, γίνεται μετατροπή των dictionaries σε μορφή λίστα από λίστες. Τέλος, ελέγχονται και κρατούνται μόνο οι ταινίες με rating>= minScore και αποθηκεύονται σε μια τελική λίστα από λίστες όπου κάθε υπολίστα είναι και ένα καλάθι. Eπίσης, αποθηκεύεται η τελική λίστα και σε αρχείο .csv.

**(2) Ρουτίνα ReadMovies :**

Η ρουτίνα αυτή παίρνει ως όρισμα το αρχείο movies. Αρχικά,δημιουργείtαι ένα αντικείμενο dataframe από το αρχείο. Για να γίνει όμως πιο εύκολη η επεξεργασία στην συνέχεια, η ρουτίνα αυτή δημιουργεί ένα dictionary από το αρχικό dataframe της μορφής { movieId: θέση στο dataframe, title, genres}.

**(3α) Ρουτίνα TriangularMatrixOfPairsCounters :**

Η ρουτίνα αυτή παίρνει ως ορίσματα τα καλάθια, τις ταινίες και το πλήθος των ταινιών. Αρχικά, υπολογίζει το πλήθος των ζευγαριών που μπορούν να προκύψουν και δημιουργεί τον πίνακα με τους counters. Στη συνέχεια διατρέχει τη λίστα των καλαθιών, για κάθε καλάθι υπολογίζει τα ζευγάρια σε λεξικογραφική σειρά και υπολογίζει τον τύπο της θέσης κάθε ζευγαριού στο μητρώο. Τέλος, ενημερώνει τον πίνακα με τους counters.

**(3β) Ρουτίνα HashedCountersOfPairs :**

Η ρουτίνα αυτή παίρνει ως όρισμα την λίστα με τα καλάθια. Αρχικά, διατρέχει κάθε καλάθι και υπολογίζει όλα τα ζευγάρια που μπορούν να προκύψουν με την συνάρτηση combinations από την βιβλιοθήκη itertools. Στην συνέχεια, αποθηκεύει σε ένα dictionary τα ζευγάρια και τον μετρητή (αν δεν υπάρχουν ήδη στο dictionary διαφορετικά αυξάνουν τον μετρητή τους). Το dictionary είναι της μορφής {ζευγάρι:counter}. Η αποθήκευση γίνεται σε dictionary για να αποφευχθούν τα περιττά ζευγάρια, άρα καλύτερη διαχείριση της μνήμης μας.

**(4) Ρουτίνα myApriori :**

Η ρουτίνα αυτή παίρνει ως ορίσματα τα καλάθια, το min frequency και το maxlength που δίνει ο χρήστης. Αρχικά, σαρώνει όλα τα καλάθια για να βρει τη συχνότητα των αντικειμένων που υπάρχουν και τα φιλτράρει με το min frequency ώστε να κρατήσει μόνο τα συχνά εμφανιζόμενα. Στην συνέχεια, βρίσκει όλα τα επόμενα ζεύγη, τριπλέτες, κ.ο.κ. μέχρι την maxlength επανάληψη ώστε να είναι έτοιμα τα δεδομένα για την επόμενη επανάληψη. Επιπλέον τα συχνά σύνολα αποθηκεύονται σε μια λίστα όπως αναφέρεται στην εκφώνηση, μετριέται ο χρόνος εκτέλεσης και εμφανίζονται τα αποτελέσματα κάθε επανάληψης.

**(5)Ρουτίνα sampledApriori :**

Η ρουτίνα αυτή παίρνει ως ορίσματα το minScore, το min frequency και το maxlength που δίνει ο χρήστης. Αρχικά, διαβάζει το αρχείο με τα ανακατεμένα αρχεία και δημιουργεί τα καλάθια. Αφού, δημιουργηθούν τα καλάθια φτιάχνει μία τυχαία λίστα με καλάθια σύμφωνα με το reservoirSample όπου στην αρχή το γεμίζει με τα 50 ή 100 (ανάλογα με το αρχείο) πρώτα καλάθια και στην συνέχεια τα αλλάζει με τυχαία σειρά. Ο αλγόριθμος μπορεί να σταματήσει να επιλέγει τυχαία πατώντας τα πλήκτρα y ή Y. Στην συνέχεια υπολογίζει την καινούργια συχνότητα και καλεί τον αλγόριθμο myApriory. Αφού τελειώσει ο αλγόριθμος myApriory ζητείται από τον χρήστη αν θέλει να γίνει δεύτερο πέρασμα. Στο δεύτερο πέρασμα δημιουργείται ένα λεξικό με όλα τα στοιχεία που βρήκε ο myApriory, υπολογίζει τον αριθμό εμφάνισης των αποτελεσμάτων και κοιτάζει τα συχνά εμφανιζόμενα σύνολα.

**(6)Ρουτίνα AssociationRulesCreation :**

Η συγκεκριμένη συνάρτηση δέχεται ως είσοδο μια συλλογή των (πιθανά) συχνών συνόλων ταινιών δηλαδή την έξοδο από το βήμα 4 και τις παραμέτρους min\_confidence, MinLift, MaxLift. Επιστρέφει ως έξοδο ένα dataframe που περιέχει κανόνες συσχέτισης ταινιών. Συγκεκριμένα, για κάθε αντικείμενο από τη συλλογή υπολογίζει όλους του πιθανούς κανόνες συσχέτισης που προκύπτουν με λεξικογραφική σειρά μέσω της αναδρομικής ρουτίνας check() που υλοποιήθηκε για την εύρεση κανόνων και εκχώρησης τους στο dataframe. Υπολογίζονται οι τιμές confidence, interest και Lift για κάθε hypothesis-conclusion του κανόνα και αν τηρούν τα κριτήρια που καθορίζονται από τις παραμέτρους min\_confidence, MinLift, MaxLift τοποθετούνται στο dataframe.

**(7) Ρουτίνα presentResults:**

Η Ρουτίνα presentResults() δέχεται ως είσοδο το rules\_df που παράγεται από τους αλγορίθμους συσχέτισης, και δίνει τις δυνατότητες παρουσίασης που περιγράφονται στην εκφώνηση.

**Ερώτηση Βonus:**

Να εκφραστεί η τιμή interest(A --> B) ως συνάρτηση της εμπιστοσύνης και της κλιμάκωσης του συγκεκριμένου κανόνα.

Γνωρίζουμε άπό τους ορισμόυς των τύπων ότι:

1. interest(A --> B) = confidene(A --> B) – Pr(B)
2. lift(A --> B) = confidene(A --> B) / Pr(B)

Από τον τύπο (2) προύπτει ότι Pr(B)= confidene(A --> B) / lift(A --> B)

Αντικαθιστώντας στον τύπο (1) προκύπτει:

interest(A --> B) = confidene(A --> B) – (confidene(A --> B) / lift(A --> B))

Βγάζοντας κοινό παράγοντα το confidence(A->B) έχουμε

interest(A --> B) = confidene(A --> B)(1 – (1 / lift(A --> B))

**Πειραματική Αξιολόγηση:**

**Αξιολόγηση SampledApriori**

Για τα παρακάτω πειράματα χρησιμοποιήσαμε ως είσοδο για τον myApriori το αρχείο “ratings\_100users.csv” και για τον sampledApriori το αρχείο “ratings\_100users\_shuffled.csv”.

Για κάθε πείραμα αποθηκεύσαμε το σύνολο των αντικειμένων που παρήχθησαν από τους δύο αλγορίθμους στα αρχεία με τη μορφή myApN.csv και samApN.csv όπου Ν ο αριθμός του εκάστοστε πειράματος. (Π.χ Στο πείραμα 1 αντιστοιχούν τα αρχεία myAp1.csv και samAp1.csv) και υπάρχουν στον αντίστοιχο φάκελο expN.

Για τον αλγόριθμο Sampled APRIORI εκτιμήσαμε την ακρίβεια της προσέγγισης (ως προς το ground-truth που παράγει ο APRIORI) σύμφωνα με τις μετρικές που χρησιμοποιήθηκαν και στην πρώτη εργασία (precision,recall,f1-score).

* Ως True-Positives θεωρήσαμε το πλήθος των αντικειμένων που ανήκουν στα σύνολα και των δύο αλγορίθμων.
* Ως False-Positives θεωρήσαμε το πλήθος των αντικειμένων που ανήκουν στο σύνολο του αλγόριθμου Sampled APRIORI και δεν ανήκουν στον APRIORI
* Ως False-Negatives θεωρήσαμε το πλήθος των αντικειμένων που ανήκουν στον αλγόριθμο APRIORI και δεν ανήκουν στον Sampled APRIORI

***Πείραμα 1ο:***

Επειδή ο αλγόριθμος sampled Apriorι επιλέγει τυχαία καλάθια με τη χρήση του reservoir sampling για να υπολογίσουμε τις παρακάτω μετρικές υπολογίσαμε το μέσο όρο των τιμών για 4 επαναλήψεις.

minscore:4, minfreq:0.2, maxlength:3 με διακοπή ροής στον Sampled APRIORI και χρήση second pass

Precision = 0.295

Recall = 0.61

F1-measure = 0.37

Επειδή ο sampled apriory αξιοποιεί μόνο ένα δείγμα των καλαθιών και επιπλέον το αποτέλεσμα βασίζεται στην τυχαιότητα του reservoir sampling, περιμένουμε τα αποτελέσματα του συγκριτικά με τον myApriorι να μην είναι τόσο καλά.

Επιπλέον επειδή διακόπτεται πρόωρα η ανάγνωση της ροής, τα καλάθια όλων των χρηστών είναι μικρότερα (γιατί οι αξιολογήσεις εμφανίζονται στη ροή “ανακατεμένες”). Αυτό σημαίνει ότι όλες οι συχνότητες συνόλων ταινιών στο δείγμα θα είναι μικρότερες σε σχέση με τις συχνότητές τους σε ολόκληρη τη ροή, αλλά όχι απαραίτητα αναλόγως μικρότερες. Όσο πιο κοντά φτάνει η διαδικασία ανάγνωσης της ροής, τόσο πιο κοντά στις πραγματικές συχνότητες θα είναι οι συχνότητες που μετρώνται εντός του δείγματος.

Από την τιμή του Recall καταλαβαίνουμε ότι ο sampled apriori βρίσκει σύνολα που δεν ανήκουν στα αποτελέσματα του myApriori. Όμως η τιμή του recall υποδεικνύει ότι έχει βρει σωστά ένα ποσοστό των αποτελεσμάτων του myApriori. Το αρμονικό μέσο (f1) δείχνει ότι δεν είναι τόσο αποδοτικός όσο ο myApriori.

**Αξιολόγηση myApriori:**

**Αποτύπωση κανόνων ως προς εμπιστοσύνη:**

Για το παρακάτω διάγραμμα τρέξαμε τον αλγόριθμο myApriori με τιμές για τις παραμέτρους

Min Score=4, Max Length=3, Min Confidence= 0.2, και αγνοήσαμε τις παραμέτρους Min Lift και Max Lift βάζοντας και στις δύο τιμές -1.

Παρατηρούμε ότι όσο αυξάνεται η τιμή της παραμέτρου Min Frequency το πλήθος των αποτελεσμάτων μειώνεται. Αυτό είναι λογικό διότι αυξάνοντας το κατώφλι μειώνεται το πλήθος των ταινιών που λαμβάνει υπόψιν του αλγόριθμος. Επίσης παρατηρούμε ότι όσο αυξάνεται η συγκεκριμένη παράμετρος μειώνονται οι διακυμάνσεις στο σχήμα, δηλαδή έχουμε περίπου το ίδιο πλήθος ταινιών για κάθε τιμή του confidence.

Εικόνα που περιέχει στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματα

**Αποτύπωση κανόνων ως προς κλιμάκωση**

Όμοια με το παραπάνω διάγραμμα τρέξαμε τον αλγόριθμο myApriori με τιμές για τις παραμέτρους

Min Score=4, Max Length=3, Min Confidence= 0.2, και αγνοήσαμε τις παραμέτρους Min Lift και Max Lift βάζοντας και στις δύο τιμές -1.

Παρατηρούμε ότι όσο αυξάνεται η τιμή της παραμέτρου Min Frequency το πλήθος των αποτελεσμάτων μειώνεται. Αυτό είναι λογικό διότι αυξάνοντας το κατώφλι μειώνεται το πλήθος των ταινιών που λαμβάνει υπόψιν του αλγόριθμος άρα υπάρχουν λιγότερες συσχετίσεις μεταξύ αυτών και μειώνεται η σπουδαιότητα της συσχέτισης. Επιπλέον μπορούμε να δούμε ότι για τη μικρότερη τιμή της παραμέτρου min frequency (0.1) οι περισσότερες τιμές συγκεντρώνονται γύρω από την τιμή 0. Αυτό μας δείχνει ότι ο αλγόριθμός βρήκε κανόνες που δεν είναι σημαντικοί.

Εικόνα που περιέχει στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματα

**Πειράματα:**

Για τα παρακάτω πειράματα χρησιμοποιήσαμε ως είσοδο για τον myApriori το αρχείο “ratings.csv”.

Για κάθε πείραμα αποθηκεύσαμε το σύνολο των αντικειμένων που παρήχθησαν σε φάκελο με τη μορφή expN όπου Ν ο αριθμός του εκάστοστε πειράματος.

Παρακάτω παραθέτουμε μερικά από τα πειράματα που δοκιμάσαμε. Δεν έχουμε παραθέσει σε όλα τα πειράματα κυκλικά διαγράμματα γιατί περιείχαν μεγάλο όγκο ταινιών και κανόνων και δημιουργούσαν σύνθετα διαγράμματα.

***Πείραμα 2ο:***

Για το παρακάτω πείραμα τρέξαμε τον αλγόριθμο myApriori με τιμές για τις παραμέτρους

minscore:5, min frequency:0.05, max length:4, min confidence:0.2, min Lift=-1, max Lift=-1. Όπως φαίνεται στις εικόνες παρακάτω, ο αλγόριθμος επέστρεψε ένα σύνολο αντικειμένων που περιείχε ζευγάρια και τριπλέτες. Ο αλγόριθμος βρήκε πιθανές τετραπλέτες που δεν περνούσαν το κατώφλι του min confidence και τις απέρριψε.

Στην δεύτερη εικόνα φαίνεται το σύνολο των κανόνων που δημιουργήθηκαν από το βήμα 6 του αλγορίθμου, δηλαδή από τη ρουτίνα AssociationRulesCreation(). Όμοια και για τα υπόλοιπα πειράματα.

Εικόνα που περιέχει πίνακας, υπολογιστής, μεγάλος, νερό

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει πίνακας, μεγάλος, υπολογιστής, νερό

Περιγραφή που δημιουργήθηκε αυτόματα

**Οπτικοποίηση Αποτελεσμάτων**

Από το παρακάτω διάγραμμα βλέπουμε ότι για το συγκεκριμένο παράδειγμα η σχέση του confidence με το lift είναι σχεδόν ανάλογη γιατί η ευθεία που δημιουργούν τα δεδομένα τείνει να είναι ανοδική. Βλέπουμε επίσης ότι όσο αυξάνεται η τιμή του lift τόσο αυξάνεται η διασπορά τον δεδομένων.***Εικόνα που περιέχει χάρτης

Περιγραφή που δημιουργήθηκε αυτόματα***

Στο παρακάτω ιστόγραμμα βλέπουμε ότι οι περισσότεροι κανόνες έχουν τιμή confidence στο διάστημα (0.2,0.3) και (0.4,0.55). Αυτό υποδεικνύει ότι οι περισσότεροι κανόνες δεν είναι τόσο σημαντικοί.

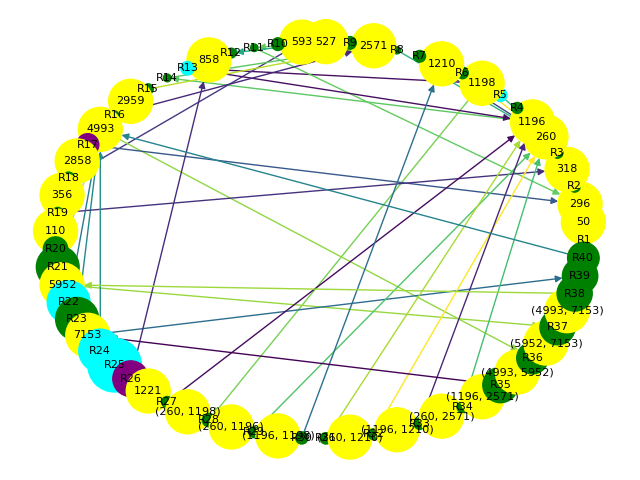
***Εικόνα που περιέχει στιγμιότυπο οθόνης

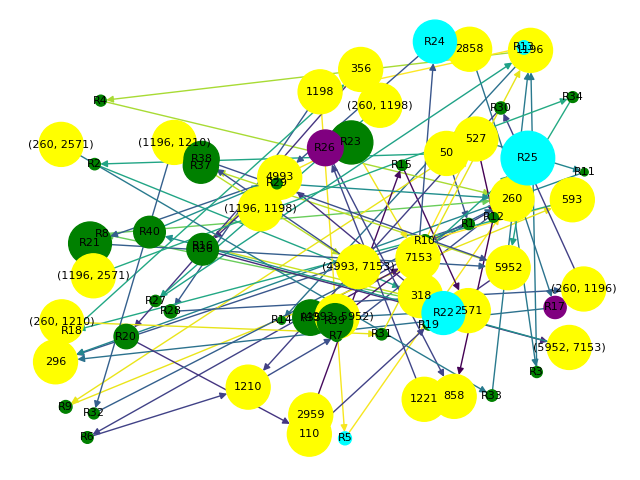
Περιγραφή που δημιουργήθηκε αυτόματα***

Παρακάτω βλέπουμε ότι οι περισσότεροι κανόνες είναι συγκεντρωμένοι κοντά στην τιμή 0 για το Lift το οποίο δείχνει ότι οι περισσότεροι κανόνες δεν παρουσιάζουν ενδιαφέρον.

***Εικόνα που περιέχει στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματα***

******

******

***Εικόνα που περιέχει κείμενο, χάρτης

Περιγραφή που δημιουργήθηκε αυτόματα***

***Παρατηρήσεις για τα κυκλικά διαγράμματα:***

Τα διαγράμματα απεικονίζουν τους κανόνες και τις ταινίες των αποτελεσμάτων. Μέσω αυτών μπορούμε να κατανοήσουμε τις συσχετίσεις που υπάρχει μεταξύ ταινιών και κανόνων αλλά και των κανόνων μεταξύ τους. Επιπλέον, το μέγεθος κάθε κόμβου-κανόνα σχετίζεται με την τιμή της κλιμάκωσης, ενώ το χρώμα του σχετίζεται με την τιμή της εμπιστοσύνης του.

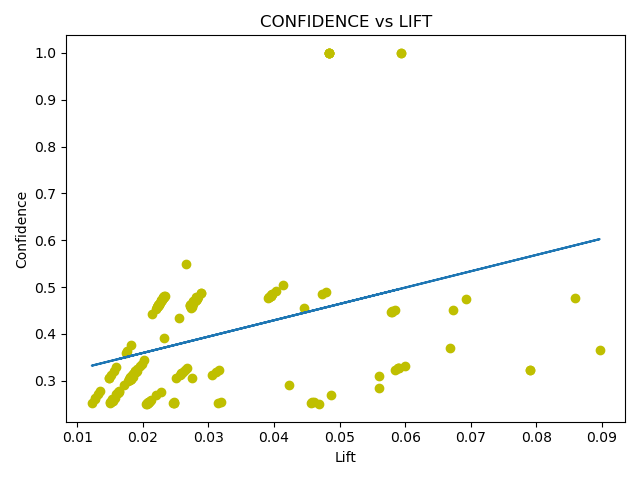
***Πείραμα 3ο:***

minscore:3, min frequency:0.15, max length:3, min confidence:0.25, min Lift=-1, max Lift=0.1

Τα αποτελέσματα παρουσιάζουν μεγάλη ομοιότητα με το πείραμα 2. Στο διάγραμμα confidence-lift βλέπουμε ότι τα δεδομένα είναι πιο πυκνά για μικρές τιμές του confidence. Οι τιμές στα ιστογράμματα είναι όμοιες με παραπάνω που δείχνουν ότι οι περισσότεροι κανόνες δεν είναι σημαντικοί.

***Εικόνα που περιέχει καθιστός, πίνακας, μαύρο

Περιγραφή που δημιουργήθηκε αυτόματα***

Εικόνα που περιέχει στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματαΕικόνα που περιέχει στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματα

***Πείραμα 4ο:***

minscore:4, min frequency:0.1, max length:4, min confidence:0.15, min Lift=-1, max Lift=0.4

Στο συγκεκριμένο πείραμα βλέπουμε ότι τα δεδομένα στο διάγραμμα confidence-lift είναι πυκνά και τείνουν να σχηματίσουν μια ευθεία. Επιπλέον στα ιστογράμματα μπορούμε να δούμε ότι οι τιμές διασκορπίζονται στο εύρος τιμών του Confidence και Lift αντίστοιχα.

Εικόνα που περιέχει καθιστός, μαύρο

Περιγραφή που δημιουργήθηκε αυτόματα

***Εικόνα που περιέχει κείμενο, χάρτης

Περιγραφή που δημιουργήθηκε αυτόματα***

**Εικόνα που περιέχει στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματα**

Εικόνα που περιέχει στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματα

***Πείραμα 5ο:***

minscore:4, min frequency:0.17, max length:3, min confidnce:0.2, min Lift=-1, max Lift=-1

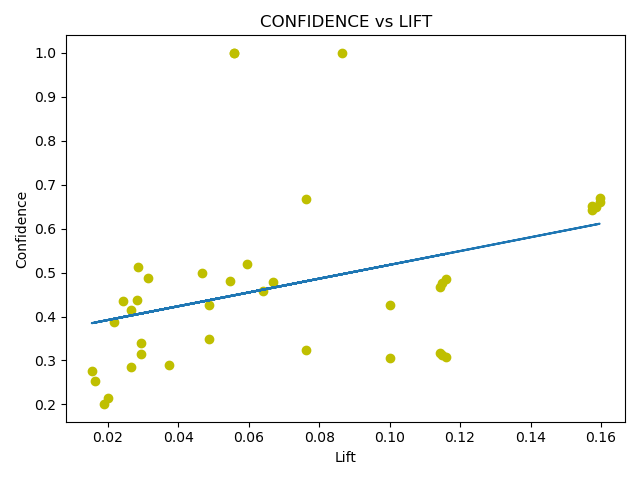
Όμοια με το πείραμα 4 βλέπουμε ότι τα δεδομένα στα ιστογράμματα διασκορπίζονται στο εύρος τιμών του Confidence και Lift αντίστοιχα. Επίσης επειδή οι κανόνες που προέκυψαν είναι λίγοι παρατηρούνται καλύτερα οι συσχετίσεις στα κυκλικά διαγράμματα.

Εικόνα που περιέχει στιγμιότυπο οθόνης, μαύρο, καθιστός, οθόνη

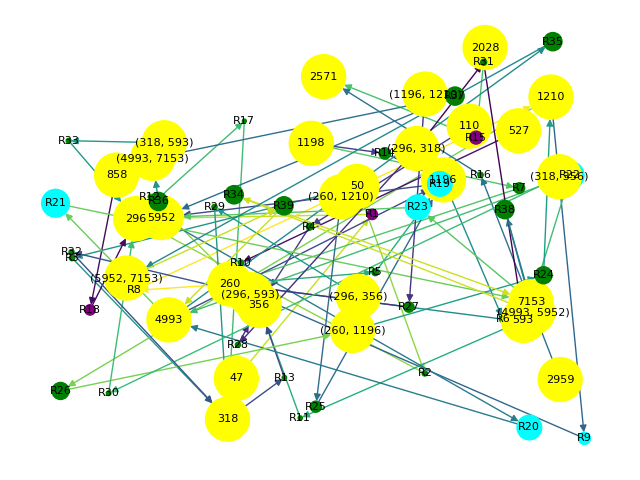
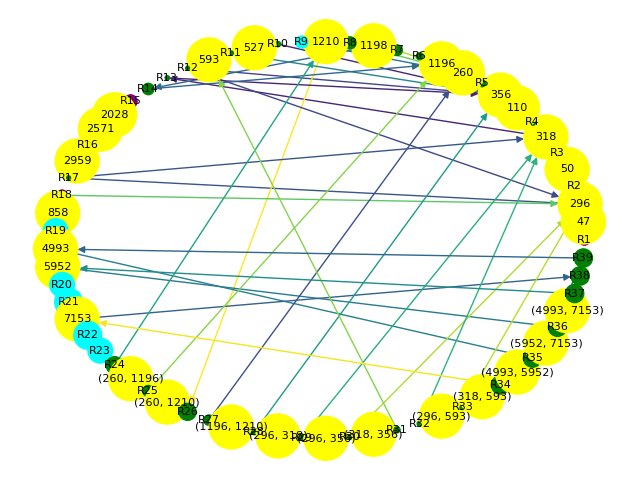
Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει υπολογιστής, μεγάλος, πίνακας, μαύρο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματαΕικόνα που περιέχει στιγμιότυπο οθόνης

Περιγραφή που δημιουργήθηκε αυτόματαΕικόνα που περιέχει κείμενο, χάρτης

Περιγραφή που δημιουργήθηκε αυτόματα

**Γενικές Παρατηρήσεις**

1. Κατά την διάρκεια των πειραμάτων παρατηρήσαμε ότι μια τιμή-ανώφλι του min frequency είναι η τιμή 0.2 γιατί για αυτή την τιμή και πάνω τα αποτελέσματα μειώνονταν σημαντικά, ειδικά σε συνδυασμό με μεγάλο Min Score.
2. Αν κρατήσουμε σταθερό το min frequency και μειώσουμε σταδιακά το Min Score τότε τα αποτελέσματα πληθαίνουν κάτι που είναι λογικό αφού τα καλάθια με τις ταινίες μεγαλώνουν.
3. Η χρήση association rules ως βασικού συστατικού σε ένα recommendation system είναι σημαντική καθώς μας δίνει πληροφορίες για το πως συνδέονται διαφορετικά αντικείμενα μεταξύ τους. Πολλά αντικείμενα με μη προφανή συσχέτιση μπορεί να αποδειχθούν μεγάλης σημασίας.
4. Υπάρχουν ενδιαφέροντες κανόνες που δίνουν απρόβλεπτες συστάσεις αλλά κατά πλειοψηφία υπερτερούν οι κανόνες που σχετίζουν ταινίες που είναι συνέχεια ή μία της άλλης ή ανήκουν στην ίδια κατηγορία.
5. Οι κανόνες που παράγονται δεν απλώνονται πολύ αφού καθώς διαπερνάμε τα δεδομένα παρατηρούμε ότι όλο και περισσότερα στοιχεία γίνονται μη συχνά με αποτέλεσμα να κόβονται στο φιλτράρισμα του myApriori. Οπότε με τα διαδοχικά περάσματα οι κανόνες εγκλωβίζονται σε συγκεκριμένες ταινίες.
6. Οι ταινίες που ανήκουν στην μακριά ουρά είναι οι ταινίες με το μεγαλύτερο support οπότε για να τις δείχνουμε στοχευμένα μπορούμε να τις παρουσιάσουμε ταξινομημένες με βάση το support ή το frequency.

Παρακάτω φαίνεται ότι λειτουργούν επίσης σωστά οι υπόλοιπες λειτουργίες από το μενού επιλογών της ρουτίνας presentResults. Στη συνέχεια παρατίθονται κάποια ενδεικτικά παραδείγματα για κάθε μία από τις λειτουργίες (απευθύνονται στο πείραμα 5).