**Παρουσίαση της δεύτερης άσκησης στο μάθημα αλγοριθμικές τεχνικές για δεδομένα ευρείας κλίμακας**

Μπάρζας Γιάννης 2765

Δρίβας παναγιώτης 2690

Στον κώδικα περιέχονται επεξηγηματικά σχόλια.

**Στο ερωτημα 1:**

φορτωνονται τα καλαθια στη λιστα από λιστες userBaskets απο το αρχειο των 610 χρηστων μονο για τις γραμμες που η βαθμολιγια ξεπερνα το min\_rating που δινει ο χρηστης

**Στο ερωτημα 2:**

φορτωνονται οι ταινιες στο dataframe movies\_df απο το αρχειο.

**Στο ερωτημα 3α:**

Στην συναρτηση TriangularMatrixOfPairsCounters() δημιουργειται ένα διανυσμα ισοδυναμο με τον κατω τριγωνικο πινακα που κρατα σε σιγκεκριμενες θεσεις τους καταλληλους μετρητες για κάθε ζευγαρι.Εδω χρησιμοπουμε την βοηθητικη συναρτηση map\_movies ετσι ώστε να εχουμε movie id σε ακολουθια για την εισαγωγη στον πινακα.

Επισης χρησιμοποιειται η βοηθητικη συναρτηση create\_pairs η οποια δημιουργει όλα τα ζευγαρια από τη λιστα με τα id που δινουμε.

**Στο ερωτημα 3β:**

Εδώ στην συναρτηση HashedCountersOfPairs() κραταμε σε ένα λεξικο κλειδιου-τιμης μετρητες σαν τιμη και το ζευγαρι, σαν κλειδι μονο για τα συνχα ζευγαρια.Κραταμε τις πλειαδες στα κλειδια με ταξινομημενα id. Αυτη η μεθοδος συνφερει αν δεν εχουμε πολλα πιθανα ζευγαρια σπαταλαμε 12 bytes ανα εγγραφη αλλα κραταμε μονο τις απαραιτητες σε αντιθεση με το 3α που σπαταλαμε 4 bytes για ολους τα ζευγαρια.

**Στο ερωτημα 4:**

Εδώ στην συναρτηση myApriori() υλοποιειται ο αλγοριθμος Apriori.

Σε κάθε περασμα υπολογιζουμε πρωτα Ck=Lk-1 X L1 με την βοηθητικη

create\_kpleiades() και στην συνεχεια για κάθε καλαθι υπολογιζουμε τις συχνοτητες μονο των συνολων του Ck.Στο τελος από κάθε περασμα κανουμε ένα φιλτραρισμα στο Ck με βαση το min\_frequency ετσι υπολογιζουμε το Lk .Για κάθε καλαθι υπολογιζουμε το μεγεθως του και αναλογα αν είναι μεγαλο η μικρο για την μετρηση των συχνοτητων του Ck.Αν είναι μικρο καλαθι δημιουργουμε ολες τις πιθανες πλειαδες του καλαθιου και ελενχουμε για κάθε μια από αυτές αν είναι μεσα στα κλειδια του Ck.Αν είναι μεγαλο ελεχουμε για κάθε ένα από τα κλειδια του Ck αν είναι υποσυνολο του καλαθιου.Ετσι εξοικονομειτε αρκετος χρονος στην εκτελεση του Α-priori.Οποτε στο πρωτο περασμα υπολογιζουμε τα μονσυνολα στο δευτερο τα δυσυνολα κτλπ…Στο λεξικο prev\_freq κραταμε το Lk-1.Στο λεξικο frequencies κραταμε για κλειδια όλα τα συχνα συνολα που βρικαμε και τιμες τις σχνοτητες τους.Η

requent\_itemsets είναι μια λιστα από λιστες με τα συχνα συνολα.

**Στο ερωτημα 5:**

Εδώ φορτωνουμε στο Dataframe ratings\_stream το αρχειο των 610 ανακατεμενων χρηστων.Σε ένα περασμα χρησιμοποιωντας την βοηθητικη συναρτηση ReservoirSampling() για δειγματολυψια

Κραταμε στο λεξικο SampleOfBaskets σταθερου μεγεθους(από τον χρηστη) δειγμα καλαθιων.Με το πληκτρο y η Y σταματαει η ροη.

Μετατρεπουμε το δειγμα σε λιστα από λιστες για την εισοδο του στον

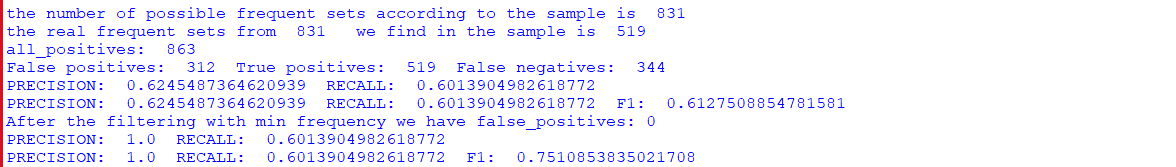
Apriori στη λιστα Apriori\_in καλειται ο Apriori myApriori() και επιστρεφεται μια λιστα από λιστες με τα συχνα συνολα Result.

Στην συνεχεια δημιουργουμε ένα λεξικο PossiblyFrequentMovieSet με κλειδια ολα τα συνολα που επιστρεψε ο Apriori με βαση το δειγμα

Και τιμη 0.Κανοντας ένα δευτερο περασμα της ροης κραταμε στο λεξικο

AllBackets όλα τα καλαθια του αρχειου και για κάθε ένα καλαθι ελενχουμε αν περιεχεται μεσα σε αυτό κάθε ένα από τα κλειδια του

PossiblyFrequentMovieSet και ετσι υπολογιζουμε τις πραγματικες συχνοτητες μονο για τα πιθανα συχνα συνολα που ειχαμε από το δειγμα στο τελος κανουμε ένα φιλτραρισμα με το min\_frequency και ετσι εχουμε τα true\_positive στην λιστα μετα το φιλτραρισμα και αν αφαιρεσουμε από τα πιθανως συχνα συνολα με βαση το δειγμα τα true\_positive εχουμε τα False positives.Επισης μπορουμε να υπολογιζουμε και τα false\_negatives αν όλα τα συχνα συνολα ολου του αρχειου αφαιρεσουμε τα true\_positive.



Εδώ παρατηρουμε ότι τις τιμες των precision recall,F1 ,true\_positive, ,False\_positive,false\_negative. false\_negative θεωρω τις συχνες πλειαδες του αρχειου που δεν ενφανιστηκαν από την επιλογη του δειγματος.Παρατηρειται ότι συνφωνα με τις παραμετρους που δωθηκαν ότι όλα τα συχνα συνολα του δειγματος είναι 863(σταθερα) και από το δειγμα βρισκω τα 519 (τιμη μεταβλητη αναλογα την επιλογη του δειγματος μας ) για αυτό η τιμη του Recall είναι υψηλη.Οι τιμες των

Precision και του F1 μετα το φιλτραρισμα με το min\_frequency είναι είναι αρκετα ηψηλες διοτι μηδενιζονται τα False positive.To precision είναι λογικο να είναι μοναδα καθως εχουμε μονο true\_positives στο τελικο αποτελεσμα μετα το φιλτραρισμα με min\_frequency τα Recall,F1

δηλωνουν μια πιο αξιοπηστη εικονα στην περιπτωση μας και μεταβαλονται με την επιλογη του δειγματος καθως αλλαζει το true\_positive .

**Στο ερωτημα 6:**

Εδώ στην συναρτηση AssociationRulesCreation() δημιουργουμε το Dataframe rules\_df οπου αποθηκευονται οι κανονες που παραγουμε.

Για καθε μια λιστα από τα δυσυνολα και μετα από την εξοδο του A-priori,περνουμε ένα ένα τα συνολα( itemset)και δημιουργουμε κανονες της μορφης [itemset-ena id]->[ena id] και για κάθε ένα τετοιο κανονα καλειται βοηθητικη αναδρομικη συναρτηση create\_Rules() η οποια πρωτα υπολογιζει τις τιμες confidence,lift,interest για τον κανονα που καλεστηκε ετσι ώστε οι ματαβλητες να εχουν καταλληλες τιμες.

Οι παραμετροι(requent\_itemsets,sunolo,rules\_df

,min\_confidence,MinLift,MaxLift,hypo,con,block\_con,block\_hypo)

Οπου requent\_itemsets λιστα από λιστες εξοδος A-priori

Hypo υποθεση ,con συμπερασμα και οι λιστες block\_con,block\_hypo

Που περιεχουν στοιχεια που δεν πρεπει να είναι στο συμπερασμα και στην υποθεση αντιστοιχα ετσι ώστε να μην υπολογιζουμε τους ιδιους κανονες πανω από μια φορα.Στην ουσια από την AssociationRulesCreation() και για κάθε ένα κανονα της προαναφερθησας μορφης καλειται η αναδρομικη συναρτηση create\_Rules() η οποια ελενχει μεχρι οποιο βαθος πληρουνται οι συνθηκες για να βρει κανονες.μετα την κληση create\_Rules() του κάθε κανονα της μορφης [itemset-ena id]->[ena id] μπενει στην λιστα block\_con το του συμπερασματος το id που ηταν στο συμπερασμα

Ετσι για ένα συνολο {1,2,3,4} η κληση create\_Rules() γινεται για κάθε ένα από τους 123->4 , 124->3, 134->2 , 234->1 και στους

, 124->3, 134->2 , 234->1

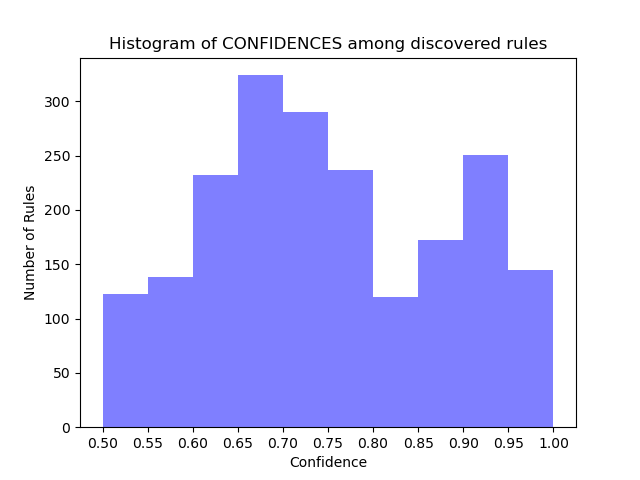
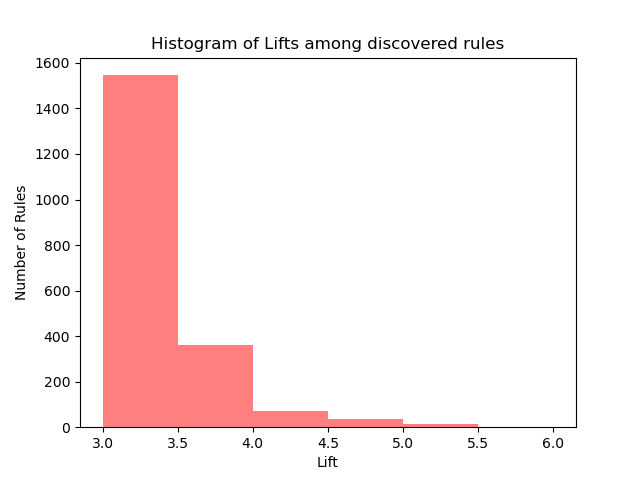
δεν υπολογιζει ξανα κανονες που στο συμπερασμα εχουν 4 στους 134->2 , 234->1 δεν υπολογιζουμε κανονες που στο συμπερασμα εχουν 4 ,3 κτλπ Επισης μεσα στην create\_Rules() ενηπερωνεται καταλληλα και

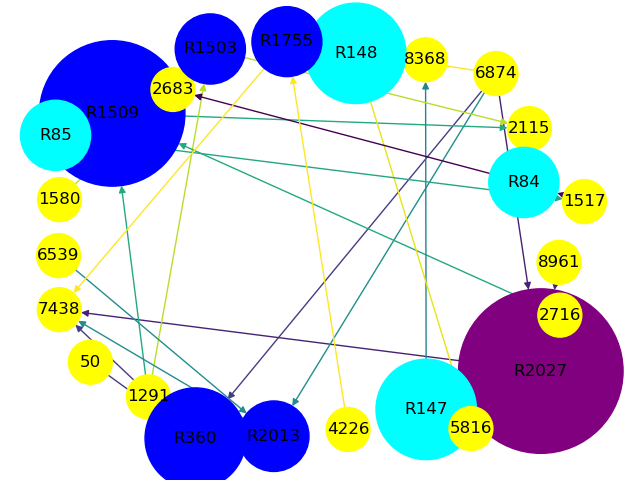
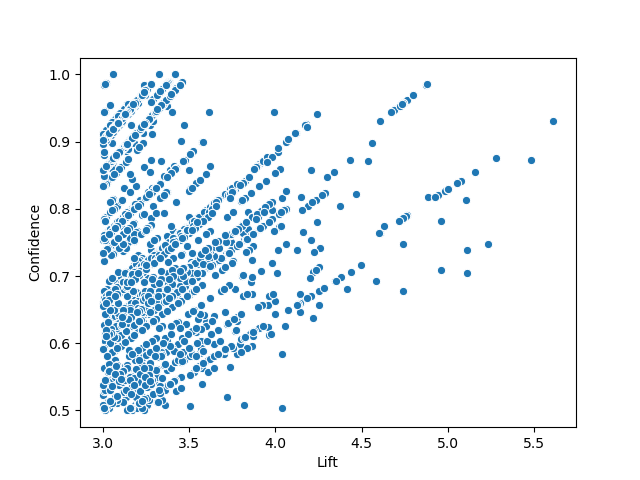
η ,block\_hypo λιστα για να μην υπολογιζει ξανα κανονες του ιδου itemset που περιεχουν στην υποθεση τα ιδια μονοσυνολα.

**Πειραματα:**

**Για minfrequency=0.1\_minconf=0.5\_minlift=3\_maxlift=1\_maxlength=3**

**\_minrating=3**



Εδώ παρατηρειται ότι ο αλγοριθμος βρισκει πολλους κανονες οι τοπ 10 κανονες εμπλεκουν 14 ταινιες από τις οποιες η ταινια που επιρεαζει περισσοτερο τους κανονες είναι η 6874 είναι η Kill Bill: Vol. 1 (2003)Action|Crime|Thriller

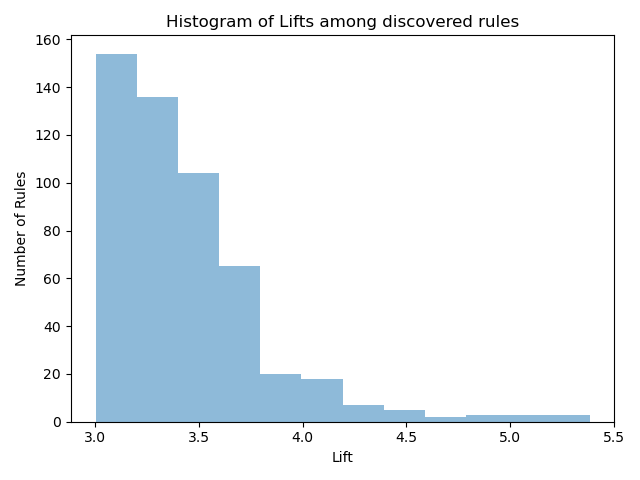
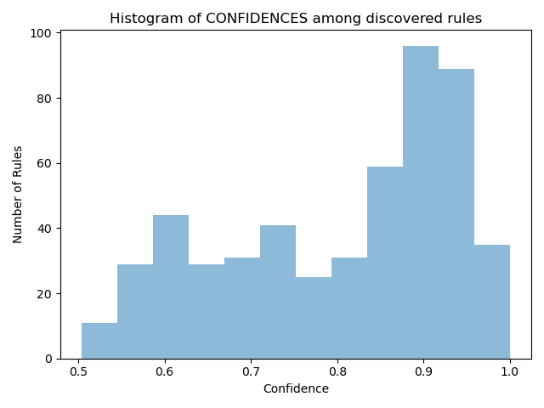
Και η πιο επιρεασμενη από κανονες είναι η 7438 η Kill Bill: Vol. 2 (2004)

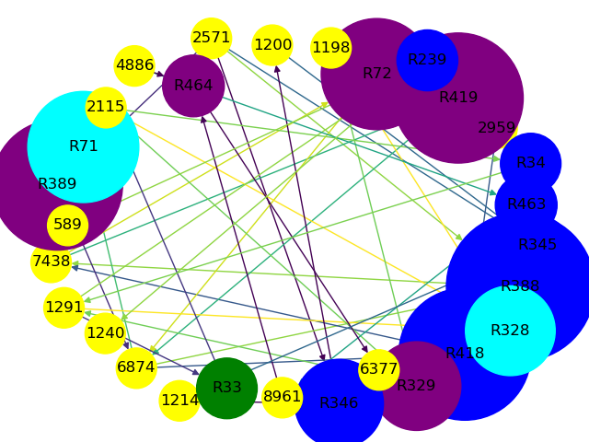
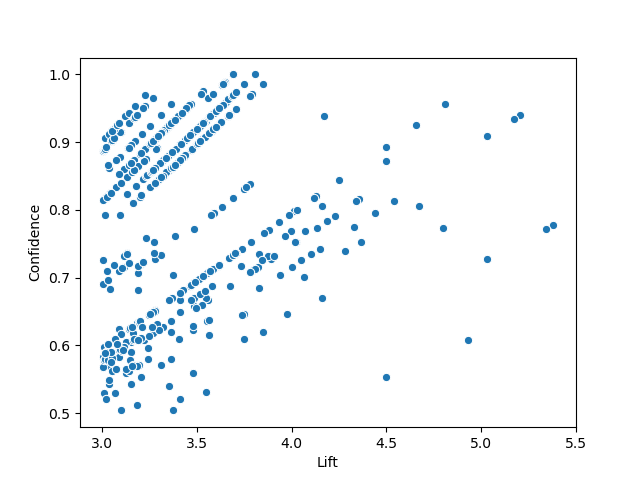
Action|Drama|Thriller το οποιο είναι απολυτα λογικο καθως υπαρχει ο κανονας R360,R2013 λογω της στοιχησης δεν φαινετι ακριβως πιος είναι από τους 2 αλλα υπαρχει ο κανονας 7438->6874.Στο γραφημα με τους κανονες επισης βλεπουμε ότι οι πιο ηψηλοι κανονες είναι R2027 ,R1509,R360 καθως οι μεγαλοι κομβοι σε μεγεθως εχουν ψηλοτερο lift ενώ οι πιο σκουροι εχουν μεγαλυτερο confidence πρωτα μωβ μετα μπλε μετα θαλασσι.Οι εμπλεκομενες ταινιες είναι 14 αυτό μας δειχνει ότι κυριαρχουν στο συνολο των κανονων μας.

Στο σκαττερ διαγραμμα συσχετησης εμπιστοσυνης – κλιμακωσης φαινεται ότι υπαρχει θετικη συσχετηση.Στα ιστογραμματα η εμπιστοσυνη εχει μεγαλυτερη συγκεντρωση κοντα στο 0.70 και μια λιγο μικροτερη κοντα στο 0.90.Στο ιστογραμμα κλιμακωσης βλεπουμε ότι οι περισσοτεροι κανονες εχουν κλιμακωση κοντα στο 3-3.5 και οσο μεγαλωνει η τιμη ολο και λιγοτεροι κανονες υπαρχουν αυτό μας δειχνει ότι οι περισσοτεροι κανονες δηλωνουν οτι ένα συνολο ταινιων ότι περιπου τριπλασιζει την πιθανοτητα ενφανισης ενός αλλου.

**Για minfrequency=0.1\_min\_rating=3.5\_max\_length=3\_min\_lift=3**

Αυξησαμε το min rating από (προηγουμενο) 3 σε 3.5





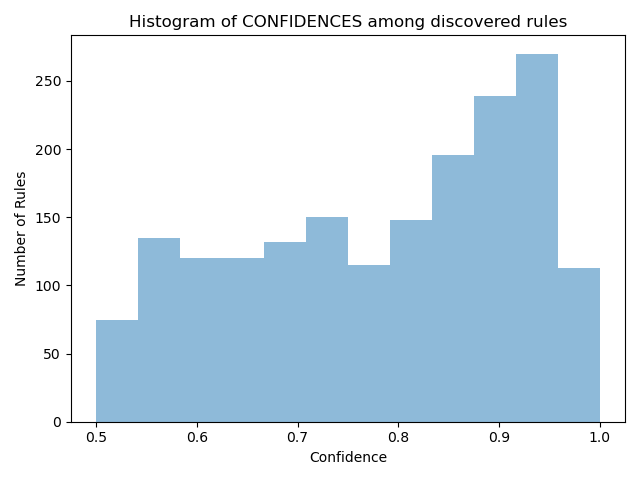
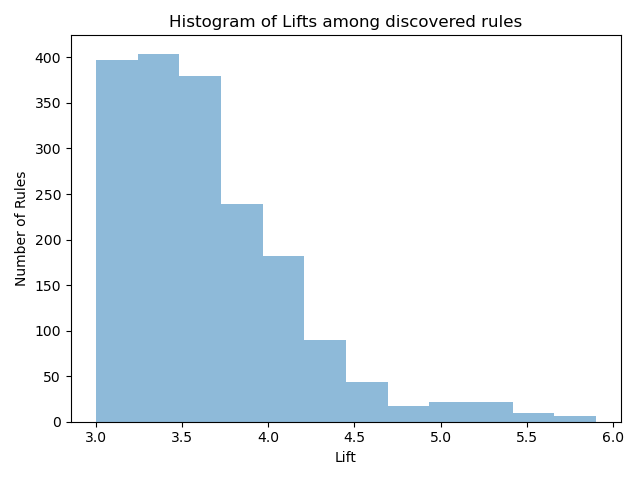
Εδώ παρατηρειται ότι καθως αυξησαμε το min rating από 3 σε 3.5 και ότι υπαρχει θετικη συσχετιση αναμεσα στην εμπιστοσυνη και τη κλιμακωση.Στα ιστογραμματα η οι περισσοτεροι κανονες εχουν εμπιστοσυνη κοντα στο 0.9 και παλι η κλιμακωση εχει φθηνουσα πορια ως προς την κλιμακωση και τον αριθμο των κανονων με σχεδον ολους τους κανονες που βρισκουμε να εχουν κλιμακωση κοντα στο 3-3,5.Εδω βλεπουμε 13 ταινιες να εμπλεκονται στους τοπ 15 κανονες ως προς κλιμακωση.Οι περισσοτερες δεν είναι ιδιες με το προηγουμενο πειραμα αλλα και εδώ βλεπουμε τους κανονες 6874,7438 οι οποιο στο προηγουμενο πειραμα ηταν οι τοπ influenced and influencer ταινιες εδώ όμως οι τοπ influenced/influencer είναι οι 1291,2571,και οι 6874,7438 .

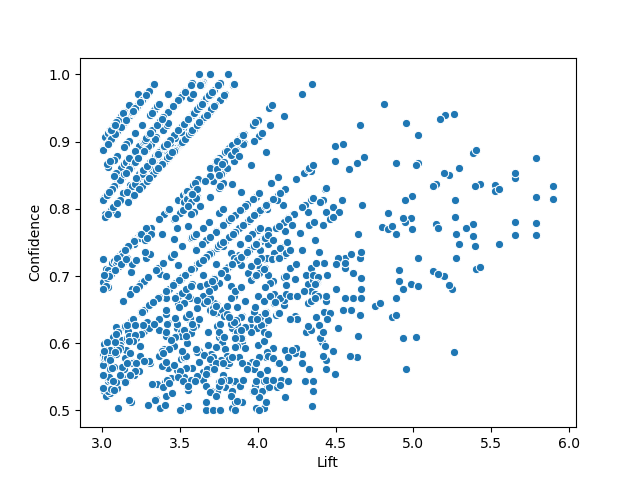
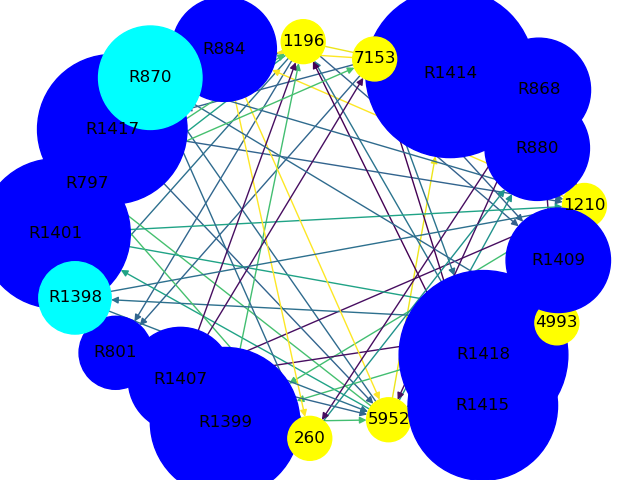
Indiana Jones and the Last Crusade (1989), Matrix, The (1999), Kill Bill: Vol. 1,Kill Bill: Vol. 2 (2004)

ολες αυτές οι ταινιες είναι λογικο να συσχετιζονται σοτυς τοπ κανονες καθως ολες είναι δρασης θριλερ και περιπετειας.

Για minfrequency=0.1\_min\_rating=3.5\_max\_length=4\_min\_lift=3

Αυξησαμε το max\_length σε 4 από 3 το προηγουμενο

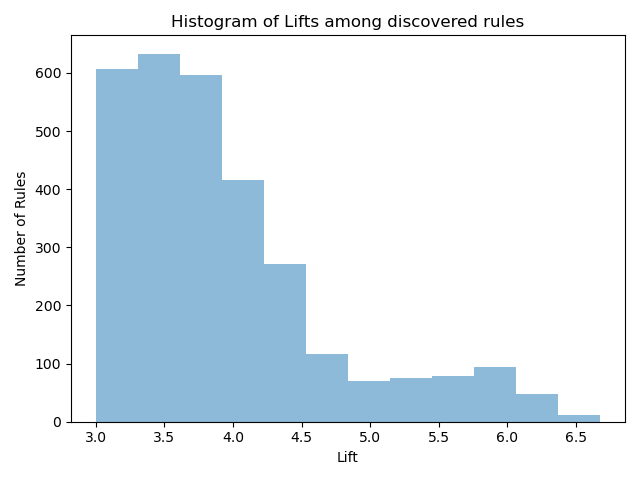
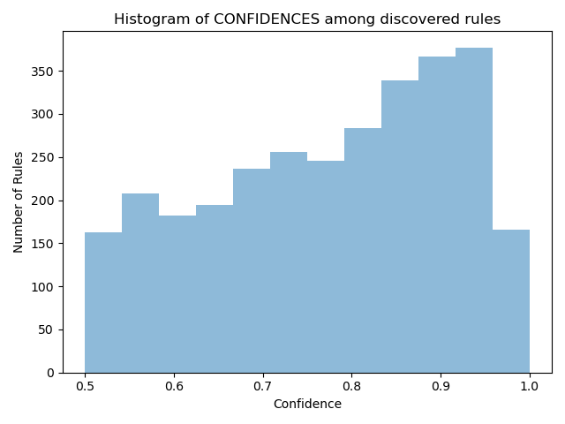
Εδώ παρατηρειται καθως αυξησαμε το max\_length σε 4 από 3 (το προηγουμενο) ότι υπαρχει θετικη συσχετιση αναμεσα στην εμπιστοσυνη και τη κλιμακωση.Στα ιστογραμματα η οι περισσοτεροι κανονες εχουν παλι εμπιστοσυνη κοντα στο 0.9 και παλι η κλιμακωση εχει φθηνουσα πορια με σχεδον ολους τους κανονες που βρισκουμε να εχουν κλιμακωση κοντα στο 3-3,5.Εδω βλεπουμε 6 ταινιες να εμπλεκονται στους τοπ 15 κανονες ως προς κλιμακωση.Δεν είναι ιδιες με τα προηγουμενα πειραματα αλλα και εδώ βλεπουμε τους περισσοτερους κανονες να εχουν και ηψηλη εμπιστοσυνη και ηψηλη κλιμακωση(σκουροχρωμοι μεγαλες μπαλες).Εδω οι κανόνες εγκλωβίζονται σε συγκεκριμένες ταινίες τις

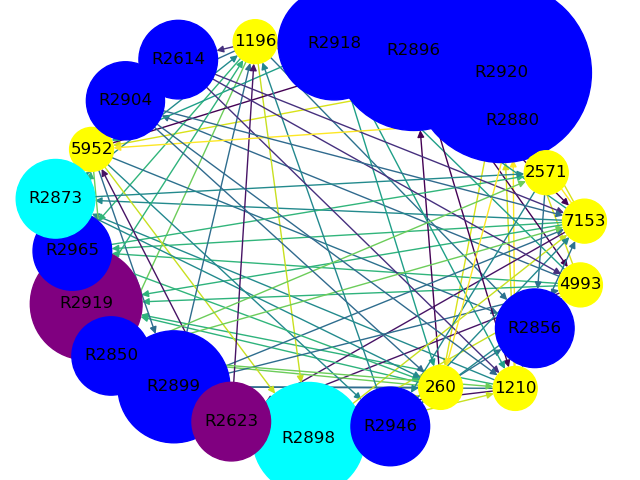
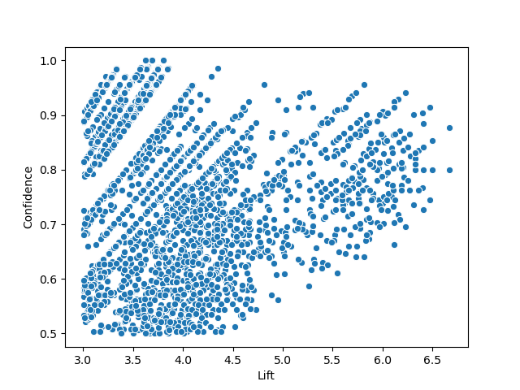
5952,1196,1210 τοπ επιρεασμενες καθως εχουν πολλα εισερχομενα βελακια.Εχουμε τις

Lord of the Rings: The Two Towers, The (2002), Star Wars: Episode V - The Empire Strikes Back (1980), Lord of the Rings: The Return of the King, The (2003), Star Wars: Episode VI - Return of the Jedi (1983) .Αρα και εδώ βλεπουμε ότι εμπλεκονται κανονες κυριως που μας λενε ότι αν δεις το ένα επισοδειο μιας ταινιας θα δεις και το άλλο αλλα επισης εχουμε και αρκετους κανονες που μας συσχετιζουν ωραια τις ταινιες με βαση τα ειδη τους.

**Για min\_frequency=0.1\_min\_rating=3.5\_max\_length=6\_min\_lift=3**

**Αυξηση στο max\_length από 4 σε 6**





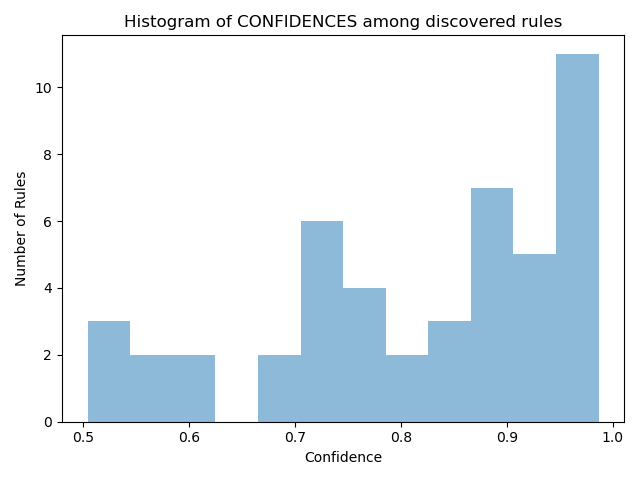
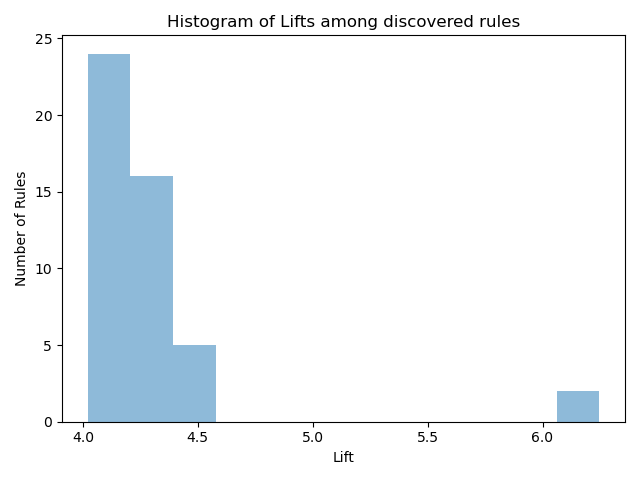
Εδώ παρατηρειται καθως αυξησαμε το max\_length σε 6 από 4 (το προηγουμενο) ότι υπαρχει ακομα πιο θετικη συσχετιση από το προηγουμενο πειραμα και αναμεσα στην εμπιστοσυνη και τη κλιμακωση.Στα ιστογραμματα η μεγλυτερη συγκεντρωση από κανονες ειναι παλι εμπιστοσυνη κοντα στο 0.9 αλλα λιγο πιο ομοιομορφα και παλι η κλιμακωση εχει φθηνουσα πορια με σχεδον ολους τους κανονες που βρισκουμε να εχουν κλιμακωση κοντα στο 3-3,5.Εδω βλεπουμε 7 ταινιες να εμπλεκονται στους τοπ 15 κανονες. Εδώ βλεπουμε τους τοπ 15 κανονες να εγκλωβίζονται σε συγκεκριμένες ταινίες όπως οι 1196,1210,260 ειναι influenced,

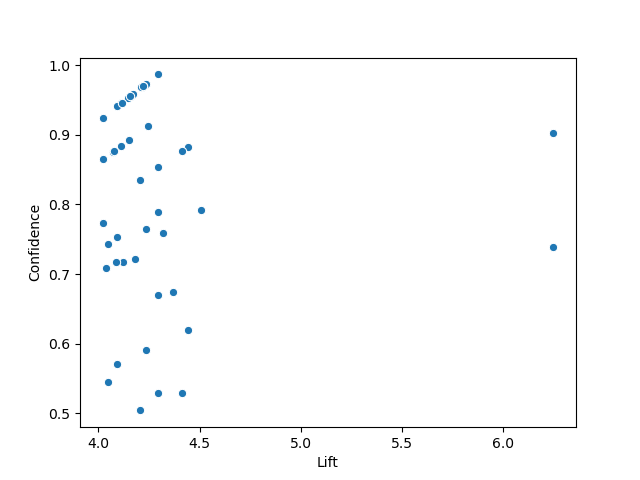
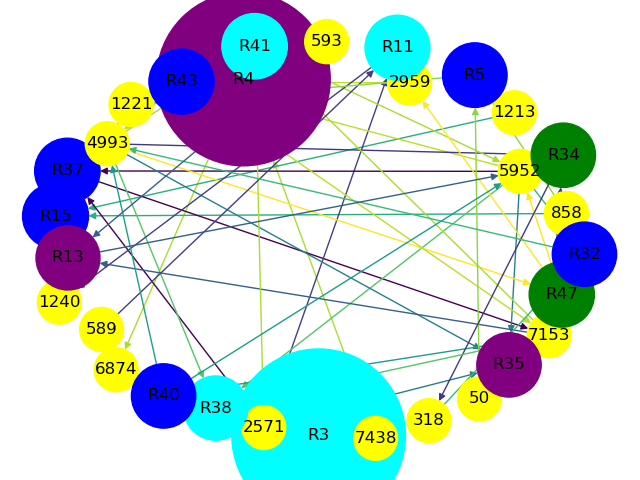
Ενώ η 7153,4993 είναι influencer.Οι κανονες με μεγαλυτερη κλιμακωση είναι οι R 2918, R 2896, R 2920, R 2880 ενώ σε εμπιστοσυνη R 2919 και ο R2623.Εδω οι ταινιες ειναι Lord of the Rings: The Return of the King, The (2003), Lord of the Rings: The Fellowship of the Ring, The (2001), Star Wars: Episode V - The Empire Strikes Back (1980), Star Wars: Episode VI - Return of the Jedi (1983), Star Wars: Episode IV - A New Hope (1977) και παλι βλεπω ότι εγκλωβιζονται παλι καποιες ταινιες

αλλα φαινεται η συσχετιση Star Wars και Lord of the Rings καθως είναι παρομοιων κατηγοριων.

**Για minfrequency=0.1\_min\_rating=4\_max\_length=4\_min\_lift=4**

**Αυξηση στο min\_rating από 3.5 σε 4**

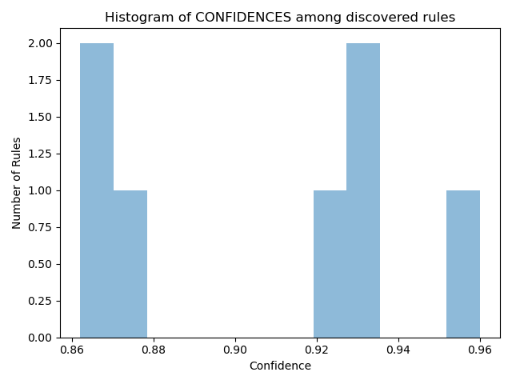
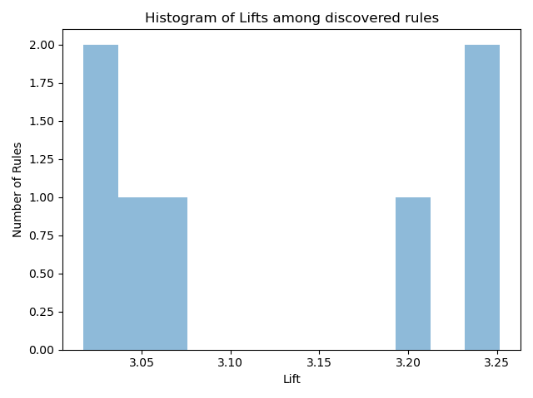
 

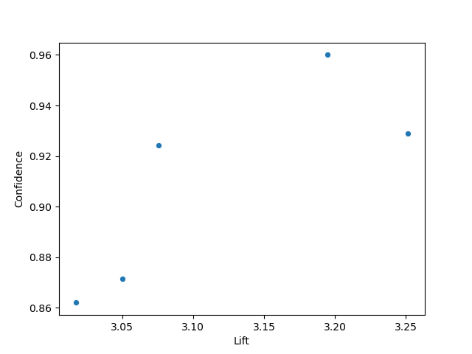
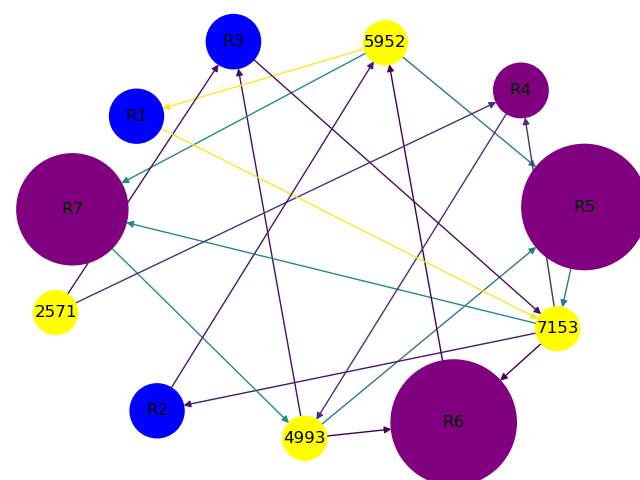
 

Από πανω βλεπουμε καθως αυξησαμε min\_rating από 3.5 σε 4 (από το προηγουμενο) ότι υπαρχει μια σταθερη συσχετιση αναμεσα στην εμπιστοσυνη και τη κλιμακωση.Επισης βλεπουμε ότι εχουν λιγοτερες βουλες αρα λιγοτεροι κανονες βρεθηκαν. Στα ιστογραμματα η μεγαλυτερη συγκεντρωση από κανονες ειναι εμπιστοσυνη κοντα στο 0.7-0.95 και καποιοι λιγοι στο 0.5-0.6 και η κλιμακωση εχει σχεδον ολους τους κανονες που βρισκουμε να εχουν κλιμακωση κοντα στο 4-4,5 λογικο αφου την αυξησαμε και καποιους λιγους μετα το 6.Αρα παρατηρουμε ότι κατά την αυξηση του σκορ βρεθηκαν λιγοτεροι κανονες αλλα η μεση κλιμακωση των κανονων που βρεθηκαν είναι μεγαλυτερη.Εδω βλεπουμε 15 ταινιες να εμπλεκονται στους τοπ 15 κανονες κατι το οποιο μας δινει πιο καλη πληροφορια.Ο κανονας 4 εχει την μεγαλυτερη εμπιστοσυνη και κλιμακωση.**Εδω βλεπουμε και πιο μη συχνες ταινιες όπως η 1213 Goodfellas (1990) συσχετιζεται με Lord of the Rings: The Fellowship of the Ring, με τον R32 οι 2 αυτές δεν είναι ιδιου ιδους η πρωτη είναι εγληματος,δραματος και η δευτερη περιπετειας φαντασιας.Αυτο μας δειχνει ότι για ηψηλοτερα min\_rating δίνουν κάποιες απρόβλεπτες συστάσεις και μπορουμε να βρουμε πιο ποιοτικες συσχετισης με πιο μη συχνες ταινιες για να προτεινουμε**.Αλλα και εδώ δεν λειπουν οι κανονες που συσχετιζουν επεισοδια από ιδιες ταινιες όπως η 589 Terminator 2: Judgment Day (1991) με τον κανονα 11 στην 1240 Terminator 2: Judgment Day (1991).

**Για minfrequency=0.2\_min\_rating=3\_max\_length=3\_minlift=3**

**Αυξηση της συχνοτητας σε 0.2**

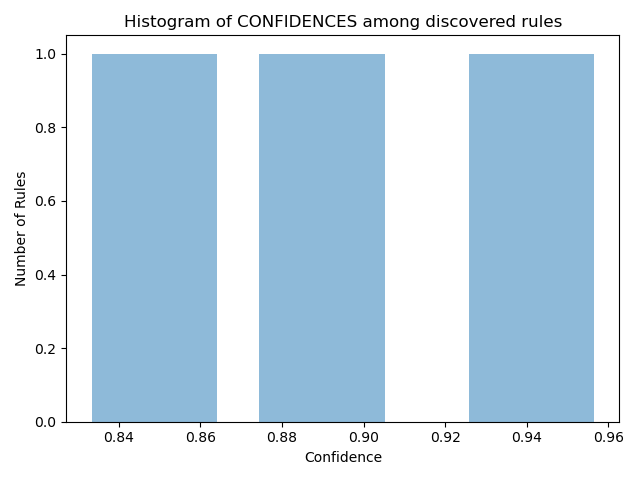
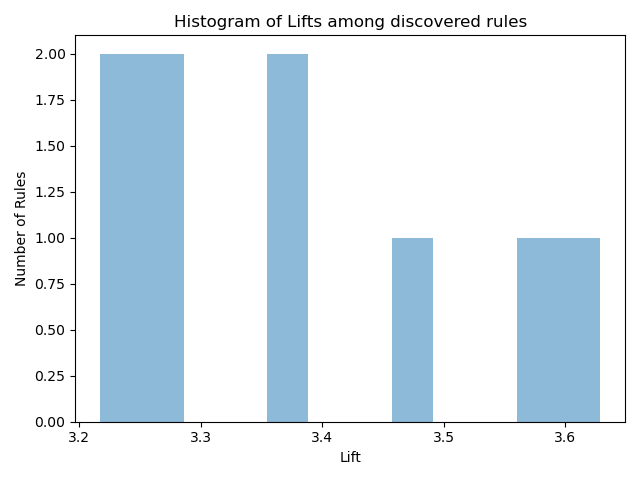
 

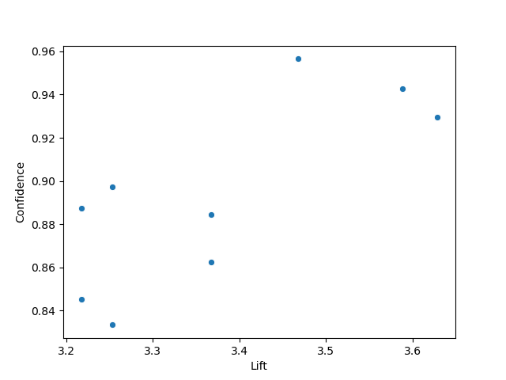
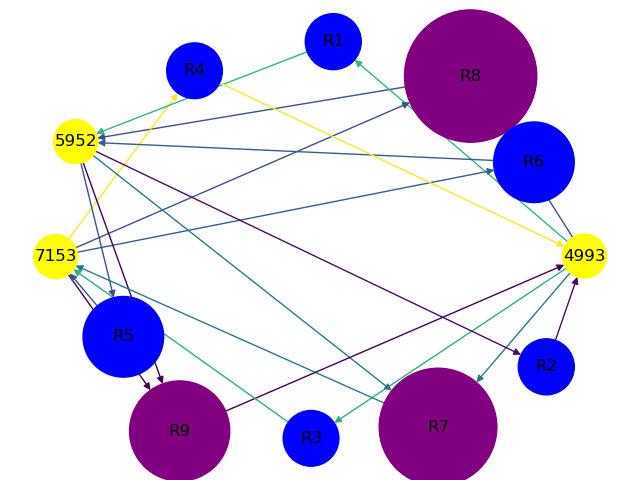
Εδώ παρατηρειται καθως αυξησαμε την συχνοτητας σε 0.2

(από το προηγουμενο) ότι υπαρχει μια θετικη συσχετιση αναμεσα στην εμπιστοσυνη και τη κλιμακωση αλλα βρισκουμε μονο 7 κανονες. Στα ιστογραμματα οι τιμες εμπιστοσυνης πανω από 0.86 αλλα δεν είναι ομοιομορφα κατανεμημενοι και η κλιμακωση εχει σχεδον ολους τους κανονες που βρισκουμε να εχουν κλιμακωση κοντα στο 3-3.5 λογικο αφου την αυξησαμε και καποιους λιγους μετα το 6.Αρα παρατηρουμε ότι κατά την αυξηση του σκορ βρεθηκαν λιγοτεροι κανονες αλλα η μεση κλιμακωση των κανονων που βρεθηκαν είναι μεγαλυτερη.Εδω βλεπουμε 4 ταινιες να εμπλεκονται στους τοπ 7 κανονες κατι το οποιο δεν μας δινει καλη πληροφορια.Οι κανονες 5,6,7 εχει την μεγαλυτερη εμπιστοσυνη και κλιμακωση.Οι ταινιες είναι Lord of the Rings: The Return of the King, Lord of the Rings: The Two Towers, The (2002, Matrix, The (1999), Lord of the Rings: The Two Towers, The (2002).Και εδώ οι περισσοτερες συσχετισεις είναι αναμεσα στα επεισοδια των Lord of the Rings αλλα εχουμε και συσχετιση με το Matrix επειδη είναι ιδιο ειδος.

**Για hist\_conf\_minfre=0.2\_min\_rating=3.5\_max\_length=6\_minlift=3**

**Αυξηση max\_length από 3 σε 6 και min\_rating από 3 σε 3.5**

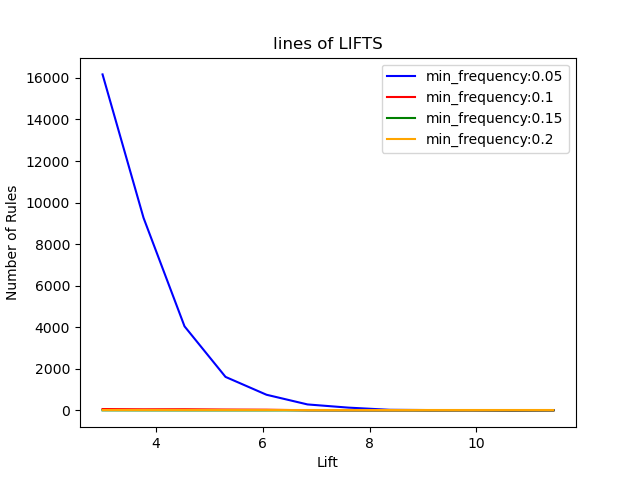
 

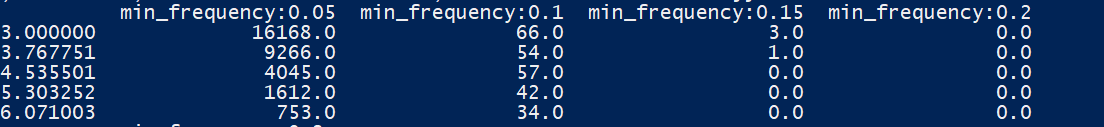
 

Εδώ παρατηρειται καθως αυξησαμε max\_length από 3 σε 6 και min\_rating από 3 σε 3.5 (από το προηγουμενο) ότι υπαρχει μια θετικη συσχετιση αναμεσα στην εμπιστοσυνη και τη κλιμακωση αλλα βρισκουμε 9 κανονες 2 παραπανω από το προηγουμενο πειραμα. Στα ιστογραμματα οι τιμες εμπιστοσυνης είναι ομοιομορφα κατανεμημενοι σε 3 καδους πανω από 0.84 και η κλιμακωση εχει περισσοτερους τκανονες που βρισκουμε να εχουν κλιμακωση κοντα στο 3-3.4 και καπιους λιγοτερους από 3.5-3.7 Φαινεται ότι κατά την αυξηση του σκορ και του max\_length βρεθηκαν λιγοι κανονες αλλα η μεση κλιμακωση και η μεση εμπιστοσυνη των κανονων που βρεθηκαν είναι μεγαλυτερη.Εδω βλεπουμε 3 ταινιες να εμπλεκονται στους τοπ 9 κανονες κατι το οποιο δεν μας δινει καλη πληροφορια.Οι κανονες 5,6,7 εχει την μεγαλυτερη εμπιστοσυνη και κλιμακωση.Και εδώ οι εγκλωβιζονται οι ταινιες Lord of the Rings στους κανονες.

**Πειραμα για 4 διαφορετικες τιμες των min\_frequency**

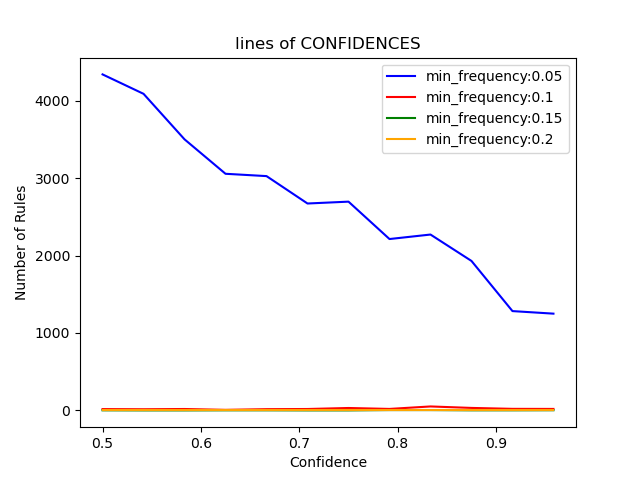
**στο συνολο=[0.05,0.1,0.15,0.2]**

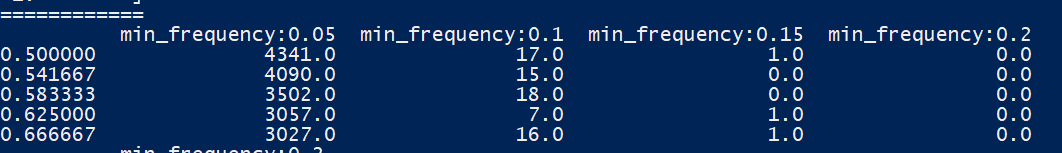




Εδώ βλεπουμε ότι οι περισσοτερες ταινιες βρισκονται για 0.05 min frequency και καθως αυξανει η κλιμακωση μειωνεται ο αριθμος των ταινιων που βρισκονται κατακορυφα .Στο DataFrame μπορουμε να δουμε τις ακριβης τιμες καθως στο διαγραμμα για όλα τα υπολοιπα min frequency οι τιμες φαινονται μηδενικες αλλα δεν είναι.Φαινονται ετσι γιατι υπαρχει τεραστια διαφορα στα νουμερα του 0.05 με τα υπολοιπα.Παρολα αυτά και στο DataFrame φαινεται ξεκαθαρα ότι και για τις άλλες τιμες ότι μειωνονται καθως αυξανεται το lift.

Καθως τεινει στο 0.2 για min\_rating 4 και min\_lift 4 βρισκουμε μηδενικες ταινιες.





Και εδώ επισης βλεπουμε ότι οι περισσοτερες ταινιες βρισκονται για 0.05 min frequency και καθως αυξανει η εμπιστοσυνη μειωνεται ο αριθμος των ταινιων που βρισκονται.Στο DataFrame μπορουμε να δουμε τις ακριβης τιμες καθως στο διαγραμμα για όλα τα υπολοιπα min frequency οι τιμες φαινονται μηδενικες αλλα δεν είναι ,φαινονται ετσι γιατι υπαρχει τεραστια διαφορα στα νουμερα του 0.05 με τα υπολοιπα.Παρολα αυτά και στο DataFrame φαινεται ξεκαθαρα ότι και για τις άλλες τιμες ότι μειωνονται καθως αυξανεται η εμπιστοσυνη.

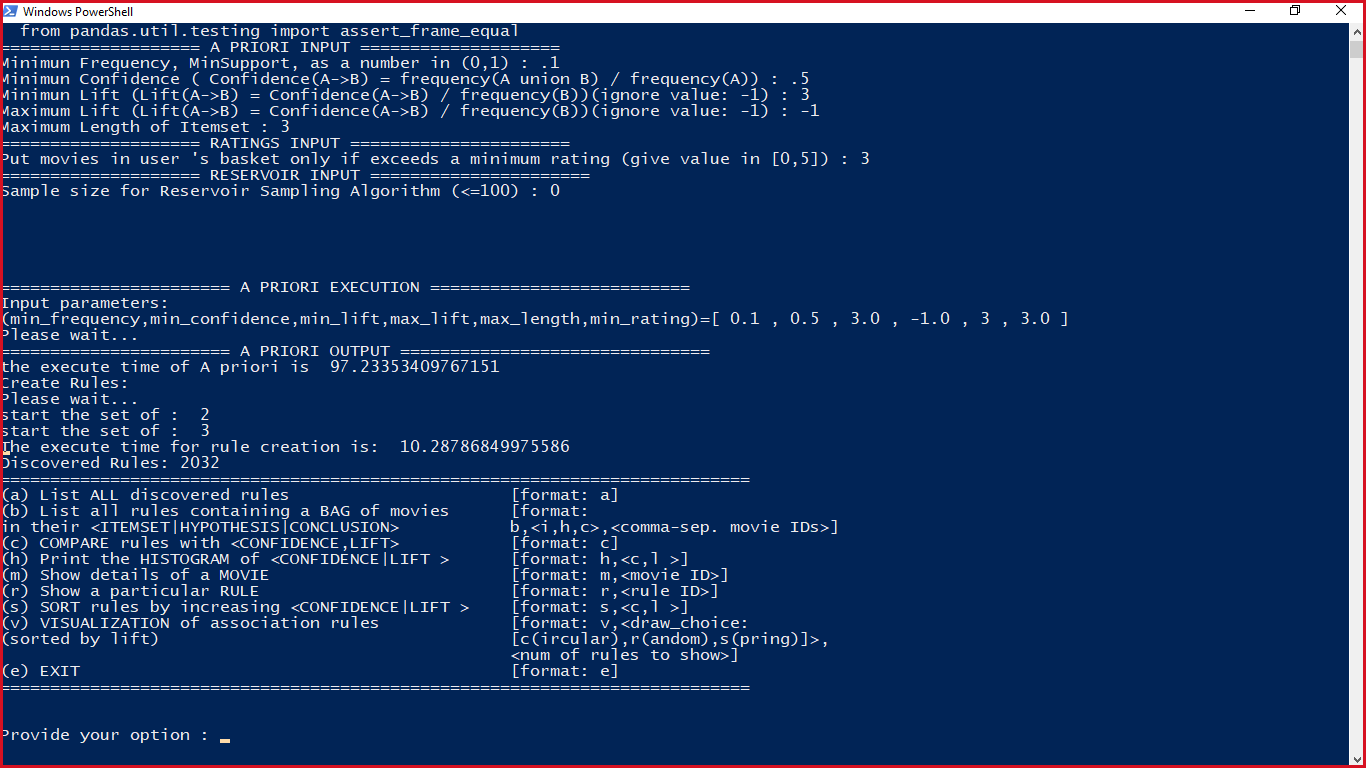
Καθως τεινει στο 0.2 για για min\_rating 4 και min\_lift 4 βρισκουμε μηδενικες ταινιες.

H ιδέα της χρήσης των Association Rules ως βασικού συστατικού για ένα Recommendation System βοηθαει γιατι βρισκουμε κανονες αναμεσα σε ταινιες που εχουν παρομοιες αλλα πολλες φορες όχι ιδιες κατηγοριες καθως και απρόβλεπτες συσχετισεις. Υπάρχουν πραγματικά ενδιαφέροντες κανόνες, που δίνουν κάποιες απρόβλεπτες συστάσειςαλλα στις περισσοτερες φορες μου λένε μόνο ότι αφού είδα το πρώτο επεισόδιο μιας σειράς ταινιών μάλλον θέλω να δω και τα επόμενα η ότι ειναι σχετικες επειδη είναι κοντινες οι περισσοτερες κατηγουριες στις οποιες εμπιπτουν.

Καθώς αποτυπώνουμε όλο και περισσότερους κανόνες, εμπλέκονται καποιες λιγες αναλογα τις παραμετρους που δινουμε ταινίες οι οποιες είναι απρόβλεπτες ή όχι τοσο συχνες αλλα συνηθως οι περισσοτεροι κανόνες εγκλωβίζονται σε συγκεκριμένες ταινίες.

Οι κανόνες που παράγονται σε λιγότερο συχνές ταινίες, δηλαδή, στη "μακριά ουρά" προϊόντων είναι συνηθως λιγοι.

Το είδος επιπρόσθετη πληροφορία που θα μπορούσαμε να αξιοποιήσουμε, προκειμένου να ανακαλύψουμε (μεταξύ των κανόνων που βρήκαμε) εκείνους τους κανόνες που μας οδηγούν να προτείνουμε στοχευμένα ταινίες που ανήκουν στη "μακριά ουρά" είναι τα ειδη -κατηγοριες στις οποιες εμπιπτουν οι ταινιες.



Εδώ φαινεται το μενου του προγραμματος υπαρχει και η τιμη -1 για να μην κανει το βημα 5 με το Sampled Αpriori.Στο e βγαινει από το μενου αλλα εκτελει την διαδικασια υπολογισμου του γραφηματος με τις γραμμες για τις 4 τιμες min frequency.