Χαροκόπειο Πανεπιστήμιο
Τμήμα Πληροφορικής & Τηλεματικής

Πτυχιακή Εργασία

Εξόρυξη γνώμης με χρήση ταξινομητών και αποτύπωση σε γραφήματα

Χαράτσεβ Φίλιππος

Επιβλέπον

Ηρακλής Βαρλάμης, Λέκτορας

Μέλη Εξεταστικής Επιτροπής

Δημοσθένης Αναγνωστόπουλος, Καθηγητής

Δημήτριος Μιχαήλ, Λέκτορας

Αθήνα
Ιούλιος 2012

Περιεχόμενα

Πίνακας Εικόνων

Σύνοψη

Το πρόβλημα του χαρακτηρισμού της άποψης που μεταφέρει μια φράση ή ένα κείμενο, έχει μεγάλο ερευνητικό και πρακτικό ενδιαφέρον. Ενδιαφέρον το οποίο γίνεται ακόμη πιο έντονο με την έλευση των εφαρμογών κοινωνικής δικτύωσης. Σε αυτές συχνά οι χρήστες εκφράζουν την άποψή τους για ορισμένα προϊόντα, το περιεχόμενο άλλων χρηστών και άλλο συναφές περιεχόμενο, χρησιμοποιώντας σύντομες φράσεις με έντονο όμως σημασιολογικό περιεχόμενο. Αντίστοιχο είναι και το ενδιαφέρον για την ανάλυση του συναισθήματος που μεταφέρουν τα λεγόμενα των χρηστών. Το πρόβλημα της εξόρυξης γνώμης συχνά ανάγεται σε ένα πρόβλημα κατηγοριοποίησης κάθε φράσης ή μέρους αυτής σε προκαθορισμένες κατηγορίες. Οι αλγόριθμοι που έχουν αναπτυχθεί ως τώρα χρησιμοποιούν πρότερη γνώση (κείμενα ή φράσεις που γνωρίζουμε την κατηγορία τους) για την εκπαίδευση του ταξινομητή, και μέτρα ομοιότητας μεταξύ φράσεων ώστε να κατατάξουν τις νέες φράσεις στην καταλληλότερη κατηγορία.

Στόχος της εργασίας είναι να αυτοματοποιήσει τη διαδικασία από την πρώτη φάση της συλλογής των γνωμών μέχρι την τελική αποτύπωσή τους σε συγκεντρωτικά γραφήματα με απώτερο στόχο να διευκολύνει την παρακολούθηση της «κοινής γνώμης», όπως αυτή διατυπώνεται με βαθμολογίες, ποιοτικούς χαρακτηρισμούς κλειστού τύπου (καλό, πολύ καλό, μέτριο κλπ.) ή και ελεύθερο κείμενο. Παράλληλα, μέσα από την εργασία, αναδεικνύεται η όποια απόκλιση υπάρχει μεταξύ της άμεσα διατυπωμένης γνώμης μέσω π.χ. βαθμολογιών (implicit ratings) και της έμμεσης γνώμης που διατυπώνεται με κείμενο και ποσοτικοποιείται με τεχνικές εξόρυξης γνώμης (implicit ratings).

Στα πλαίσια της εργασίας, τα δεδομένα τα οποία επιλέχθηκαν προς εξόρυξη γνώμης, αντλήθηκαν από το διαδικτυακό τόπο IMDb (http://www.imdb.com/), όπου φιλοξενούνται πληροφορίες που αφορούν ταινίες οι οποίες είτε έχουν προβληθεί, είτε θα προβληθούν στους κινηματογράφους. Ο ίδιος δικτυακός τόπος προσφέρει επίσης κριτικές χρηστών για ορισμένες από αυτές τις ταινίες. Οι κριτικές των 250 υψηλότερα βαθμολογημένων, από τους χρήστες ταινιών, αποτελούν το βασικό σύνολο δεδομένων για την εργασία . Με τη χρήση της γλώσσας προγραμματισμού JAVA, υλοποιήθηκε ένας Web crawler ο οποίος επισκέπτεται ιστοσελίδες οι οποίες περιέχουν κριτικές χρηστών για τις παραπάνω ταινίες, αντλεί το περιεχόμενο, δημιουργεί αντικείμενα JAVA, στα οποία καταχωρεί τα δεδομένα, και στη συνέχεια τα αποθηκεύει με τη μορφή αρχείων αντικειμένων στο δίσκο, για κάθε ταινία ξεχωριστά. Στη συνέχεια γίνεται κατηγοριοποίηση των δεδομένων αυτών με τη χρήση ενός αλγορίθμου ταξινόμησης κειμένων που βασίζεται στο Μηχανισμό Διανυσμάτων Υποστήριξης (Support Vector Machines-SVM) και συγκεκριμένα μιας υλοποίησης του αλγορίθμου LibSVM που είναι διαθέσιμη στις βιβλιοθήκες του περιβάλλοντος Weka (http://www.cs.waikato.ac.nz/ml/weka/) και ενσωματώθηκε στην εφαρμογή.

Η πληροφορία που συγκεντρώνεται για κάθε ταινία καθώς και η πληροφορία κατηγοριοποίησης της αποθηκεύεται σε νέα αρχεία αντικείμενων, ενώ εξάγεται και σε αρχεία XML ώστε να είναι πιο εύκολη η επισκόπηση της πληροφορίας κάθε ταινίας.

Όλες οι προηγούμενες λειτουργίες ενσωματώνονται στη συνέχεια σε μια JAVA εφαρμογή, η οποία επιτρέπει στον χρήστη να επιλέξει τον τίτλο ταινίας της αρεσκείας του, να δει την επεξεργασμένη πληροφορία της γνώμης των χρηστών για την συγκεκριμένη ταινία, και να επιλέξει από μια σειρά οπτικοποιήσεων και εναλλακτικών γραφημάτων την πορεία της γνώμης των θεατών για την ταινία μέσα στα χρόνια. Τα γραφήματα, υλοποιούνται με τη χρήση των βιβλιοθηκών JFreeChart (http://www.jfree.org/jfreechart/) που είναι γραμμένα σε JAVA και ενσωματώνονται στην εφαρμογή.

Στις συνεισφορές της εργασίας συγκαταλέγονται: α) οι εναλλακτικές μορφές οπτικοποίησης που ενσωματώθηκαν και περιλαμβάνουν τόσο τεχνικές για την αποτύπωση βαθμολογίας και κατηγοριοποιημένης γνώμης, όσο και τεχνικές για την αποτύπωση του περιεχομένου των σχολίων ανά κατηγορία γνώμης, β) η μεθοδολογία η οποία αναπτύχθηκε τόσο για την εξόρυξη γνώμης, η οποία χρησιμοποιεί έναν αλγόριθμο SVM, αλλά μπορεί εύκολα να λειτουργήσει και με περισσότερους αλγορίθμους, γ) η εξαγωγή της πληροφορίας σε αρχεία XML, στα οποία αποθηκεύθηκαν οι γνώμες χρηστών ανά ταινία και τα οποία μπορούν να χρησιμοποιηθούν ως Dataset, για οποιαδήποτε μελλοντική ερευνητική και μη χρήση, δ) η δυνατότητα προσθήκης επιπλέον ταινιών στο σύνολο των αρχικών 250 ταινιών με αυτόματο τρόπο μέσα από την εφαρμογής.

1

Εισαγωγή

1.1 Εξόρυξη Γνώμης και Οπτικοποίηση Γνώμης

Ο όρος Εξόρυξη Γνώμης (Opinion Mining) αφορά την επεξεργασία κειμένων γραμμένων σε φυσική γλώσσα με τη χρήση υπολογιστικών και στατιστικών μεθόδων, με σκοπό την εξαγωγή υποκειμενικών πληροφοριών. Ο γενικότερος στόχος της Εξόρυξη Γνώμης είναι ο προσδιορισμός της άποψης του χρήστη για ένα συγκεκριμένο θέμα, προϊόν κλπ. ή και η γενικότερη σημασιολογική πολικότητα του περιεχομένου ενός εγγράφου (contextual polarity).

Η άποψη του εκάστοτε χρήστη μπορεί να προκύπτει είτε από την προσωπική κρίση, είτε από τη συναισθηματική του κατάσταση, είτε από τη συναισθηματική επιρροή την οποία ασκεί στο χρήστη ο συγγραφέας του κειμένου, είτε σαφώς από ένα συνδυασμό των παραπάνω.

Στις υπολογιστικές και στατιστικές μεθόδους οι οποίες χρησιμοποιούνται στην εξόρυξη γνώμης από κείμενα συμπεριλαμβάνονται: η επεξεργασία φυσικής γλώσσας (Natural Language Processing), η λεξικογραφική ανάλυση κειμένου μέσω υπολογιστικών αλγορίθμων (Computational linguistics), όπως και η εξόρυξη δεδομένων από κείμενο (Text Mining). Στις μεθόδους Text Mining δίνεται μεγάλη έμφαση στην αναγνώριση προτύπων, με τη μορφή λέξεων, συνδυασμών λέξεων, συχνότητας λέξεων ή όρων, και ακολουθείται μια διαδικασία κατά την οποία αναθέτονται χαρακτηρισμοί (labels) στο εκάστοτε κείμενο. Μια χαρακτηριστική μέθοδος Text Mining είναι η κατηγοριοποίηση (classification) κατά την οποία κομμάτια κειμένου μπορεί να χωριστούν σε ομάδες (κλάσεις), ανάλογα με μια υποκειμενική κατηγοριοποίηση του περιεχομένου αυτών.

Ως Οπτικοποίηση δεδομένων(Data Visualization) ορίζεται ουσιαστικά η οπτική αναπαράσταση δεδομένων, τα οποία έχουν εξαχθεί είτε από κείμενο είτε από άλλου τύπου πηγές, με τη μορφή σχηματικών δομών και γραφημάτων. «Κύριος σκοπός τη μεθόδου αυτής είναι, να διαδώσει πληροφορία ξεκάθαρα και αποτελεσματικά με τη χρήση γραφικών μέσων» (Friedman, 2008).

Οι οπτικές αναπαραστάσεις επωφελούνται από την ικανότητα του ανθρώπινου ματιού, στο να μεταφέρει ταχύτατα και αποτελεσματικά ένα ευρύ φάσμα πληροφοριών στον εγκέφαλο, ώστε να επιτραπεί στους χρήστες να δουν, να παρατηρήσουν, να εξερευνήσουν και να κατανοήσουν άμεσα μεγάλο όγκο δεδομένων (Thomas & Cook, 2005). Έμφαση δίνεται

στην δημιουργία μεθόδων, ώστε να μεταδίδεται πληροφορία , μέσω της διαισθητικής ικανότητας του κάθε χρήστη.

Τα τελευταία χρόνια έχουν παρατηρηθεί οπτικοποιήσεις στον παγκόσμιο ιστό, οι οποίες μπορούν να περιγράφουν από πολιτική θεματολογία μέχρι και άρθρα από εφημερίδες. Οι οπτικοποιήσεις αυτές είναι προσβάσιμες από χιλιάδες χρήστες, και είναι επόμενο το ερώτημα του τι είδους ευκαιρίες παρουσιάζονται όταν οι οπτικοποιήσεις μπορούν να αξιολογηθούν από μεγάλα πλήθη ανθρώπων. Οι διαδραστικές οπτικοποιήσεις, όχι μόνον αποτελούν ένα κομβικό μέσο επικοινωνίας σε ένα κόσμο γεμάτο πληροφορία, αλλά αφήνεται να εννοηθεί σε πολλές αναφορές ότι δυνητικά μπορούν να έχουν καταλυτική επιρροή στις αφηγηματικές τεχνικές (storytelling) και στην ανάλυση συλλογικών δεδομένων (collective data analysis) (Many Eyes, 2007).

Οι πιο τυπικές τεχνικές ανάλυσης πληροφορίας μέσω της οπτικοποίησης είναι τα ιστογράμματα, γραφικές παραστάσεις, δεντρικά διαγράμματα κ.ά., μέσω των οποίων μπορεί να γίνει εξαγωγή περεταίρω πληροφορίας και σημαντικών συμπερασμάτων, χωρίς να είναι απαραίτητη η εξειδίκευση του χρήστη σε κάποιο επιστημονικό τομέα. Η φαντασία και η δημιουργικότητα και μόνο μπορούν να παράγουν σημαντικά συμπεράσματα, τα οποία θα ήταν πολύ δύσκολο να εξαχθούν χωρίς τη χρήση τεχνικών οπτικοποίησης. Συμπεράσματα τα οποία σαφώς μπορούν να επεξεργαστούν περεταίρω μέσω στατιστικών αναλύσεων και άλλων συναφών μεθόδων.

1.2 Σκοπός της εργασίας

Σκοπός της παρούσας εργασίας είναι η υλοποίηση μιας ολοκληρωμένης εφαρμογής εξαγωγής γνώμης από κείμενα και οπτικοποίησης της. Μέσα από την εργασία μελετώνται και υλοποιούνται διάφοροι τρόποι οπτικοποίησης με παράλληλη αξιοποίηση διαθέσιμων APIs.

Δίνεται έμφαση στην παραμετροποίηση των γραφημάτων, ώστε ο χρήστης να έχει τη μέγιστη δυνατή διάδραση με αυτά και να μπορεί να τα αντιπαραβάλει διαμορφώνοντας έτσι άποψη α) για το πώς κατανέμονται οι γνώμες με το χρόνο, β) για την αντιστοιχία της εξαγόμενης ποιοτικής γνώμης από τα κείμενα με τις εμφανείς ποσοτικοποιημένες γνώμες (βαθμολογίες), γ) για την πόλωση στις γνώμες και τις λέξεις που χρησιμοποιούνται σε κάθε πόλο.

Τέλος δίνεται έμφαση στη δημιουργία αρχείου με τις γνώμες (αρχικές και επεξεργασμένα αποτελέσματα) ώστε να είναι και μελλοντικά διαθέσιμες.

Μικρότερη έμφαση δίνεται στον αλγόριθμο κατηγοριοποίησης των άρθρων, λόγω όμως της σχεδίασης της εφαρμογής είναι δυνατή η ενσωμάτωση οποιασδήποτε άλλης υλοποίησης (σε java).

1.3 Μελέτη Περίπτωσης

Η συλλογή και οπτικοποίηση απόψεων αποτελεί ένα ευρύ πλαίσιο μελέτης και εξαγωγής συμπερασμάτων. Η περίπτωση του διαδικτυακού τόπου IMDB και των γνωμών των χρηστών, η οποία μελετάται στη παρούσα εργασία, παρουσιάζει ιδιαίτερο ενδιαφέρον.

Τα σχόλια και η γνώμες των χρηστών ανά ταινία ποικίλουν, και ενδεχομένως επηρεάζονται ή διαμορφώνονται από μια πληθώρα παραγόντων, όπως για παράδειγμα το πότε ένας χρήστης παρακολούθησε μια ταινία, σε σχέση με την ημερομηνία κυκλοφορίας της ταινίας στο ευρύ κοινό, ή το πόσο δημοφιλείς είναι μια ταινία κάποια συγκεκριμένη χρονική περίοδο. Όπως επίσης η γνώμη ενός χρήστη μπορεί να επηρεαστεί διαβάζοντας τις γνώμες άλλων χρηστών για μια ταινία, κάτι που πιθανότατα θα ισχύει και τη βαθμολογία.

Ανεξάρτητα από το είδος και τα γενικότερα χαρακτηριστικά των ταινιών, οι απόψεις των χρηστών ποικίλουν και η γενική άποψη του συνόλου των χρηστών παρουσιάζει διακυμάνσεις, άλλοτε ελάχιστες και άλλοτε πολύ σημαντικές επηρεάζοντας τη γενικότερη «εικόνα» για μια ταινία πολλές φορές ριζικά, ανά διάφορες χρονικές περιόδους. Φαινόμενο εξαιρετικά ενδιαφέρον προς μελέτη, καθώς οπτικοποιόντας τα δεδομένα μπορούμε να εξάγουμε σημαντικά συμπεράσματα σε ελάχιστο χρόνο, για τη γενικότερη άποψη του κοινού για μια ταινία. Αξιοποιώντας δε τις παρατηρήσεις που κάνουμε βάση της οπτικοποίησης μπορούμε να εξάγουμε στοιχεία τα οποία είναι ικανά να αποτελέσουν τη βάση για τη μελέτη ορισμένων φαινομένων και τα αίτια αυτών. Όπως για παράδειγμα να ερευνήσουμε το λόγο για τον οποίο οι ταινίες X και Y, του Z σκηνοθέτη, παρουσίασαν σημαντική άνοδο στις προτιμήσεις των χρηστών (θετικά σχόλια) κατά τη διάρκεια ενός συγκεκριμένου έτους. Στοιχεία τα οποία θα ήταν σχεδόν αδύνατο να εξαχθούν και να αξιοποιηθούν χωρίς τη χρήση οπτικοποίησης, κυρίως λόγο του όγκου των δεδομένων από τα οποία προέκυψαν.

Η εν λόγω εφαρμογή δέχεται ως είσοδο τα σχόλια των χρηστών του διαδικτυακού τόπου IMDb (http://www.imdb.com/), σχόλια σε μορφή κειμένου, που αφορούν ορισμένες ταινίες οι οποίες έχουν προβληθεί στους κινηματογράφους. Οι γνώμες των χρηστών για κάθε ταινία ποικίλουν, όπως επίσης ο κάθε χρήστης έχει τη δυνατότητα βαθμολόγησης της εκάστοτε ταινίας βάσει των προσωπικών του υποκειμενικών κριτηρίων. Οι γνώμες, οι βαθμολογίες όπως και άλλα στοιχεία συλλέγονται μέσω κατάλληλου Crawler που έχει προσαρμοστεί στη δομή της σελίδας σχολίων του imbd και κατηγοριοποιούνται σε θετικές, αρνητικές ή ουδέτερες. Η κατηγοριοποίηση γίνεται είτε με βάση το περιεχόμενο του κάθε σχολίου χρήστη, με χρήση ενός ταξινομητή SVM, είτε με βάση τη βαθμολογία του χρήστη. Αφού γίνει η ταξινόμηση για το σύνολο των σχολίων των χρηστών για μία συγκεκριμένη ταινία, ο χρήστης της εφαρμογής έχει τη δυνατότητα να επιλέξει από μια λίστα έναν από τους τρόπους οπτικοποίησης που παρέχει η εφαρμογή, ώστε να εμφανιστεί στο χρήστη, οπτικοποιημένο, το αποτέλεσμα της κατηγοριοποίησης της γνώμης των χρηστών του IMDB για τη εκάστοτε ταινία. Επίσης έχει τη δυνατότητα να αλλάξει κλίμακα εστιάζοντας σε μια συγκεκριμένη περίοδο, η ομαδοποιώντας τα σχόλια ανά μήνα, τρίμηνο κλπ. Τέλος έχει τη δυνατότητα

να προσθέσει δυναμικά νέες ταινίες μέσα από την εφαρμογή και να εξάγει τα δεδομένα για οποιαδήποτε ταινία σε μορφή XML.

1.4 Προσωπικά οφέλη

Στα πλαίσια της παρούσας εργασίας αποκόμισα ορισμένα σημαντικά οφέλη.

Εξοικειώθηκα σε μεγάλο βαθμό με τη γλώσσα προγραμματισμού JAVA, κατανοώντας σε μεγάλο βαθμό ορισμένες πτυχές και δυνατότητες τις συγκεκριμένης γλώσσας, τόσο κατά τη κατασκευή του Web Crawler για τη συλλογή δεδομένων, όσο και κατά την υλοποίηση της κύριας εφαρμογής. Απέκτησα επίσης εμπειρία χειρισμού προγραμματιστικών λαθών, κάτι το οποίο σαφέστατα, θα με ωφελήσει και σε μελλοντικές εργασίες και έρευνες.

Πολύ σημαντική επίσης θεωρώ την εξοικείωση που απέκτησα με έννοιες και όρους στον τομέα της Οπτικοποίησης Γνώμης και δεδομένων, τομέα της πληροφορικής για τον οποίο είχα πλήρη άγνοια πριν ασχοληθώ ενεργά με τη παρούσα εργασία. Ένα πολύ ενδιαφέρον τομέα, η χρησιμότητα και επιρροή στη επιστήμη του οποίου, δεν είναι εμφανείς στο μέσο άνθρωπο, αλλά ταυτόχρονα παράγει γνώση και αξιοποιήσιμη πληροφορία με γρήγορο και αποτελεσματικό τρόπο.

Με δεδομένο το μεγάλο όγκο πληροφορίας που δημοσιεύεται και συγκεντρώνεται καθημερινά στο διαδίκτυο, οι τεχνικές οπτικοποίησης κρίνονται πλέον απαραίτητες για τη γρήγορη και αποτελεσματική συνόψιση της πληροφορίας και αυτή η εργασία μου έδωσε σημαντικές γνώσεις προς την κατεύθυνση αυτή.

2

Υπόβαθρο

2.1 Text & Opinion Mining . IMDb.

Στο πεδίο της εξόρυξης πληροφοριών από κείμενο(Text mining), επιχειρείται η εξαγωγή πληροφοριών με ουσιαστικό νόημα, από κείμενο γραμμένο σε φυσική γλώσσα. Με την ευρεία έννοια μπορεί να χαρακτηριστεί ως μια διαδικασία εξαγωγής πληροφορίας η οποία είναι χρήσιμη και αξιοποιήσιμη για κάποιο συγκεκριμένο σκοπό. Σε σύγκριση με τον τύπο των δεδομένων που αποθηκεύεται στις βάσεις δεδομένων, το κείμενο σε φυσική γλώσσα δεν έχει σαφή δομή και μορφή, και ο χειρισμός του αλγοριθμικά είναι πολύ δύσκολος. Παρόλα αυτά η ανταλλαγή και διάδοση πληροφοριών μέσω κειμένου, είναι πλέον ο πιο διαδεδομένος τρόπος επικοινωνίας.

Η εξόρυξη γίνεται συνήθως από κείμενο το οποίο περιέχει έγκυρες πληροφορίες ή γνώμες, ενώ υπάρχει πάντα υψηλό κίνητρο για την αυτοματοποίηση της διαδικασίας έστω και αν το αποτέλεσμα δεν είναι τέλειο.

Το πεδίο της ανάλυσης συναισθήματος και εξόρυξης γνώμης, μελετά τα συναισθήματα, τις γνώμες, τις αξιολογήσεις και τη συμπεριφορά των ανθρώπων, το σύνολο των οποίων αποτυπώνετε στο γραπτό λόγο. Είναι ένα από τα περισσότερο ενεργά πεδία μελέτης το οποίο συνδέεται με την επεξεργασία φυσική γλώσσας όπως επίσης μελετάται ευρέως στα πεδία της επιστήμης των υπολογιστών, της εξόρυξης πληροφορίας από κείμενο και το διαδίκτυο. Η ανάπτυξη μέσων ενημέρωσης, όπως τα διαδικτυακά forum, τα προσωπικά ιστολόγια(blogs) και οι υπηρεσίες κοινωνικής δικτύωσης προσδίδει ιδιαίτερη αξία και σημαντικότητα στη διαδικασία εξόρυξης γνώμης. Πλέον κυρίως λόγο των παραπάνω υπηρεσιών, υπάρχει διαθέσιμος προς ανάλυση ένας τεράστιος όγκος πληροφοριών, αποθηκευμένων σε ψηφιακή μορφή, ο οποίος είναι κυρίως παράγωγο γνωμών και απόψεων(Liu, 2012).

Η αντίληψη του κόσμου, οι προσωπικές πεποιθήσεις του καθενός καθώς και οι επιλογές οι οποίες κάνουμε επηρεάζονται έντονα από τον τρόπο που αντιλαμβάνεται τον κόσμο ο κοινωνικός μας περίγυρος και όχι μόνο. Για το λόγο αυτό συνήθως ζητούμε τις γνώμες άλλων ανθρώπων προτού πάρουμε μια απόφαση. Γεγονός το οποίο ισχύει εξίσου για μεμονωμένα άτομα αλλά και επιχειρήσεις.

Με τη λογική αυτή επιλέγουμε να αγοράσουμε ένα προϊόν αφού πρώτα διαβάσουμε κριτικές για αυτό και ανταγωνιστικά του προϊόντα, επιλέγουμε να δούμε μια ταινία αφού πρώτα δούμε τις βαθμολογίες που έχει λάβει από κριτές και θεατές ή αφού πρώτα διαβάσουμε τις γνώμες άλλων θεατών. Ο ιστότοπος IMDb έχει σχεδιαστεί με αυτό το

σκεπτικό, με τρόπο ώστε να συλλέγει πέρα από γενικές πληροφορίες για τις ταινίες και τις γνώμες των θεατών γι' αυτές. Μπορεί κάλλιστα να χαρακτηρισθεί ως μια τεράστια πηγή κειμένου το οποίο περιέχει γνώμες και απόψεις, όπως επίσης είναι ένα χαρακτηριστικό παράδειγμα υπηρεσίας την οποία μπορεί να συμβουλευτεί ο οποιοσδήποτε ενδιαφερόμενος ώστε να διαβάσει/συμβουλευτεί γνώμες άλλων χρηστών. Ο συγκεκριμένος ιστότοπος φιλοξενεί πολλές φορές, χιλιάδες μεμονωμένες απόψεις χρηστών για μία συγκεκριμένη ταινία, με τη μορφή κειμένου. Απόψεις προσβάσιμες σε όλους όσους επισκέπτονται τον ιστότοπο. Αποτελεί ουσιαστικά μια τεράστια βάση δεδομένων μη επεξεργασμένης πληροφορίας, ιδανική για τη διενέργεια μελέτης στα πεδία της ανάλυσης συναισθήματος και εξόρυξης γνώμης, όπως εξαιρετικό ενδιαφέρον παρουσιάζει η οπτικοποίηση της πληροφορίας η οποία θα εξαχθεί από τις γνώμες των χρηστών, προς τη διερεύνηση κοινωνικών φαινομένων και τάσεων που αφορούν την παρακολούθηση και υποκειμενική αξιολόγηση των διαθέσιμων κινηματογραφικών ταινιών.

2.2 Τεχνικές Visualization

2.2.1 Μέθοδοι απεικόνισης κειμένου

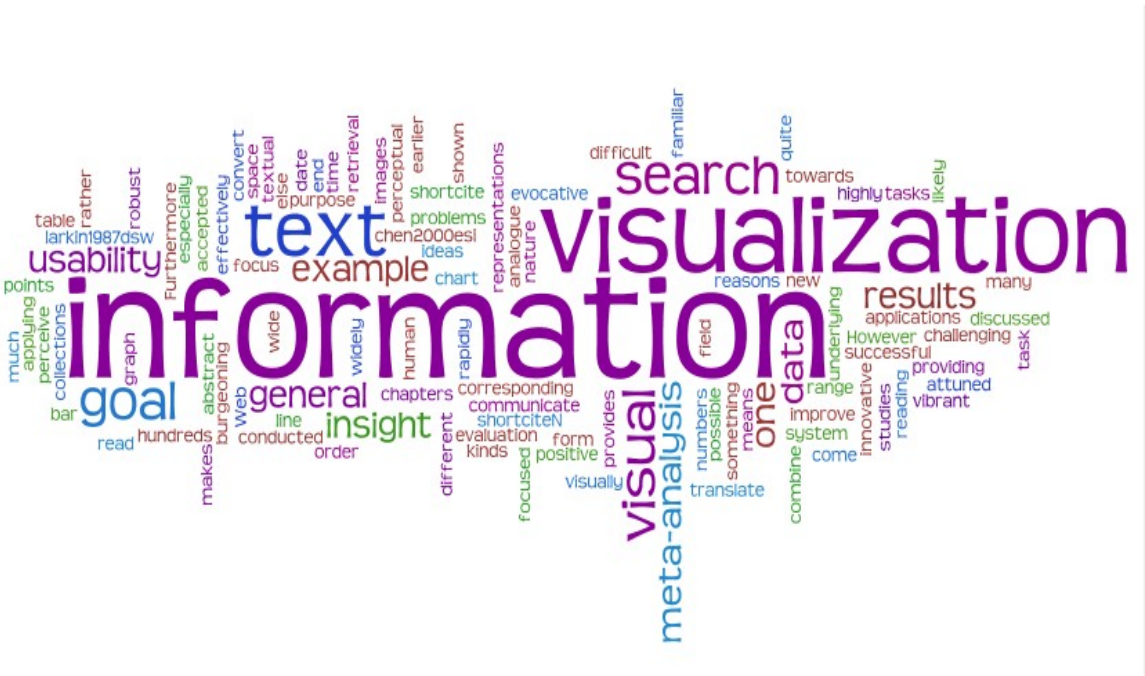
Υπάρχουν πολλοί και διαφορετικοί μέθοδοι αναπαράστασης του περιεχομένου ενός κειμένου. Οι μέθοδοι αυτοί εστιάζουν στη παρουσίαση των συχνοτήτων λέξεων ή φράσεων, στη συσχέτισή τους ή στη σύγκριση δυο διαφορετικών κειμένων. Παρακάτω παρατίθενται ορισμένες από αυτές τις μεθόδους.

Tag Clouds και Wordles

Το Tag Cloud είναι ίσως η πιο συχνή μέθοδος οπτικοποίησης κειμένου. Αποτελείτε ουσιαστικά από μια λίστα λέξεων, φράσεων ή περιγραφών, οι οποίες έχουν μέγεθος ανάλογο με τη συχνότητα εμφάνισης κάθε λέξης, δηλαδή το πόσες φορές εμφανίζεται μια συγκεκριμένη λέξη σε ένα κείμενο. Χρησιμοποιούνται συχνά για να τονίσουν τις κύριες πτυχές ενός κειμένου.

Εικόνα 1. Tag Cloud, Πηγή: http://searchuserinterfaces.com

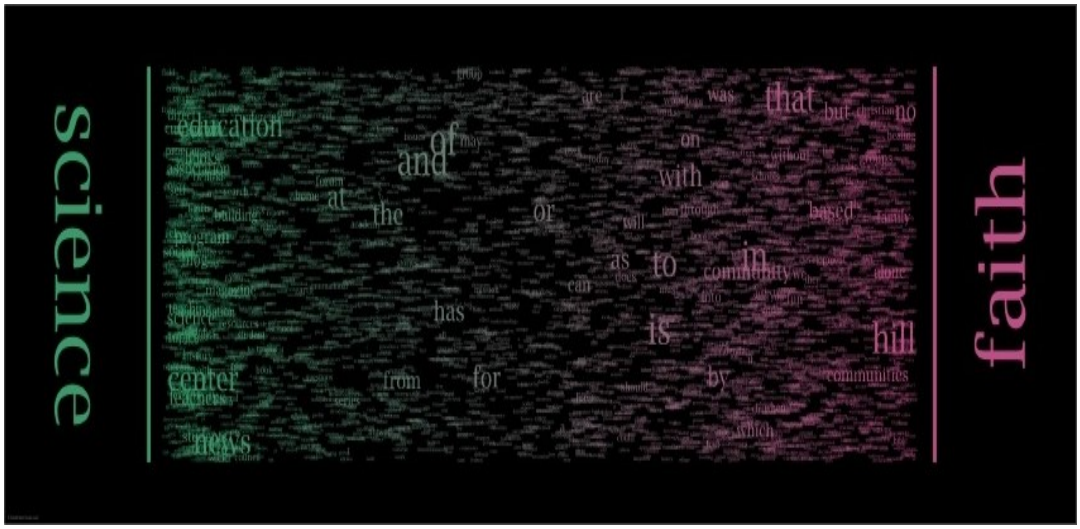
Τα Wordles είναι παρόμοια με τα Tag Clouds. Αποτελούν μια περισσότερο «καλλιτεχνική» εκδοχή, όπου οι λέξεις απεικονίζονται με διαφορετικά χρώματα και διάταξη. Τείνουν να είναι λιγότερο ενημερωτικά, αλλά προσδίδουν μια περισσότερο προσωπική αποτύπωση στην οπτικοποίηση του εγγράφου.



Εικόνα 2. Wordle, Πηγή: http://searchuserinterfaces.com

Διαγράμματα Φάσματος Λέξεων

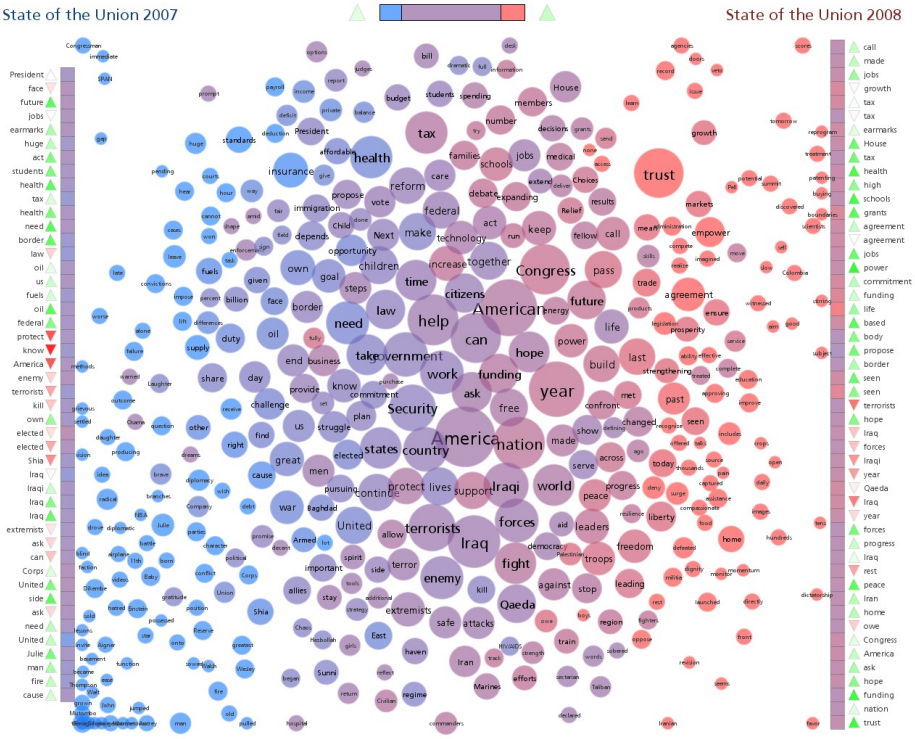
Τα διαγράμματα παρουσιάζουν τις σχέσεις μεταξύ δύο λέξεων ή φράσεων. Απεικονίζουν τις σχέσεις, ουσιαστικά των σύνολο των κοινών λέξεων, που «συνοδεύουν» συνήθων ένα ευρέως διαδεδομένο ζεύγος λέξεων. Για παράδειγμα το μέγεθος αναπαράστασης της λέξης education ορίζεται από το άθροισμα των συχνοτήτων εμφάνισης του ζεύγους "science education" και του ζεύγους "faith education". Η θέση της λέξης education ορίζεται από την αναλογία των συχνοτήτων εμφάνισης των δυο ζευγών.



Εικόνα 3. Διάγραμμα φάσματος λέξεων, Πηγή:
http://www.chrisharrison.net/index.php/Visualizations/WordSpectrum

Διαγράμματα αντίθεσης λέξεων

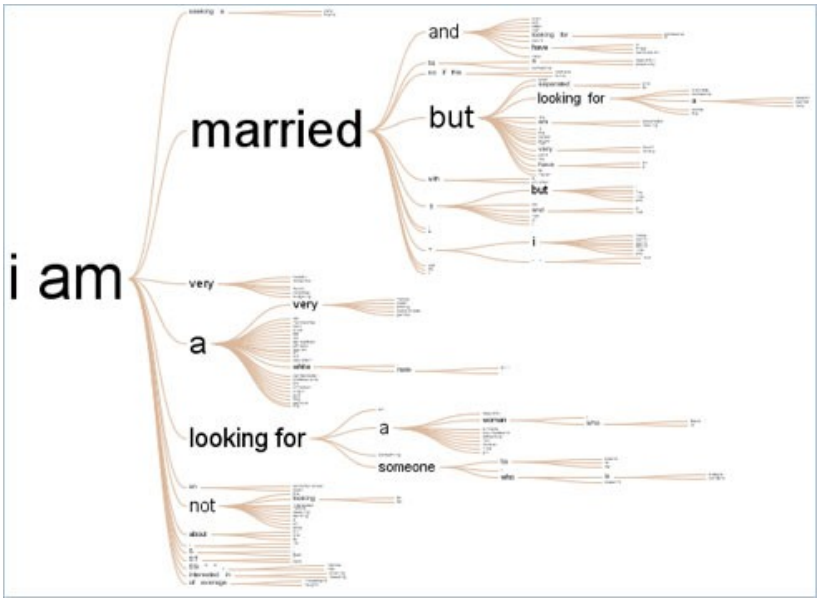
Έχουν τη δυνατότητα να τονίσουν τις κοινές λέξεις μεταξύ δύο κειμένων. Τονίζοντας για παράδειγμα τη συχνότητά τους με το μέγεθος και το βαθμό συναισθηματικής φόρτισης με το χρώμα τις αναπαράστασης κάθε λέξης. Στο εικονιζόμενο παράδειγμα επιχειρείται σύγκριση δύο αναφορών γραμμένων με διαφορά ενός έτους, οι οποίες αφορούν το ίδιο θέμα.



Εικόνα 4. Διάγραμμα αντίθεσης κειμένων, Πηγή: http://www.neoformix.com

Δέντρα λέξεων

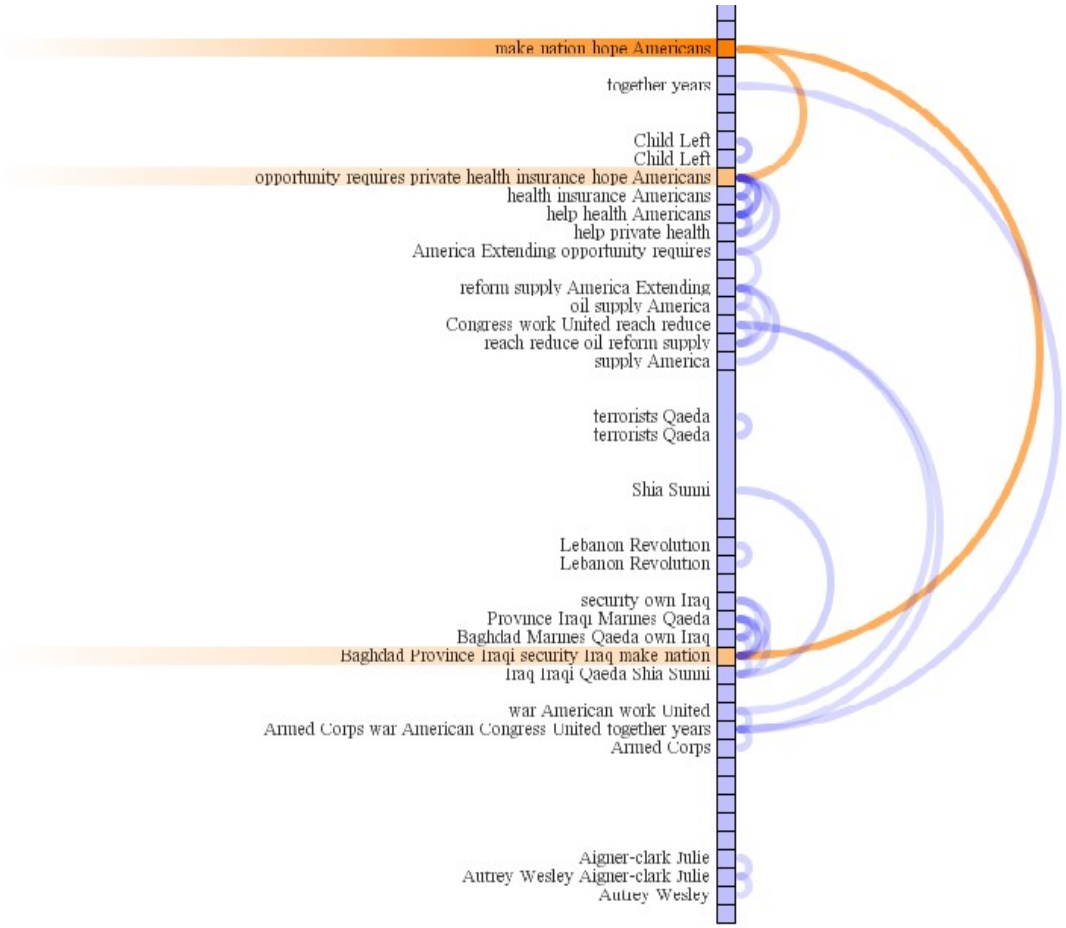
Τα Δέντρα λέξεων παρέχουν ένα πλαίσιο σε ένα κείμενο χωρίς δομή, αναδεικνύοντας τις σχέσεις μεταξύ των σημαντικότερων λέξεων ή φράσεων και αυτών των οποίων τις ακολουθούν στο κείμενο. Το μέγεθος τις αναπαράστασης αντιπροσωπεύει τη σημαντικότητα ή τη συχνότητα εμφάνισης στο κείμενο, ενώ η δομή βοηθά τον εκάστοτε χρήστη να ακολουθήσει «σημασιολογικές» διαδρομές από μια έννοια που απαντάται στο κείμενο, προς μία διαφορετική.



Εικόνα 5. Δέντρο λέξεων, Πηγή: http://hint.fm

Διαγράμματα Τόξων

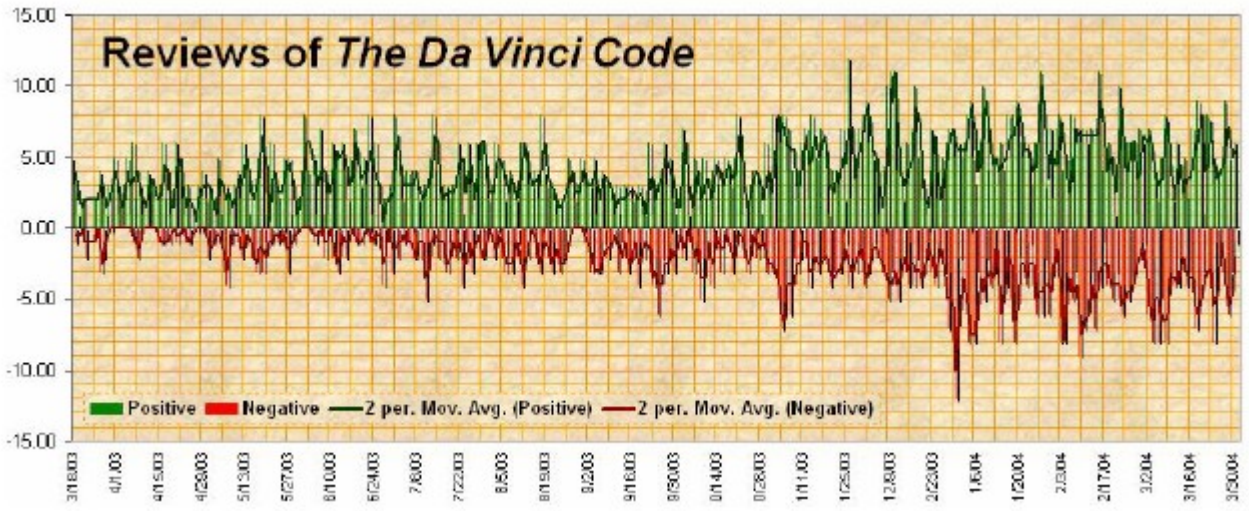
Απεικονίζουν την ομοιότητα στη δομή ενός κειμένου. Με τη χρήση τόξων, συνδέονται οι φράσεις ενός κειμένου οι οποίες περιέχουν παρόμοιο λεξιλόγιο. Το παρακάτω παράδειγμα παρουσιάζει τη χρήση διαγράμματος τόξων σε κείμενο από προεκλογική εκστρατεία στις ΗΠΑ.



Εικόνα 6. Διάγραμμα τόξων, Πηγή: http://www.neoformix.com

Χρονική Ανάλυση Αντικρουόμενων Σχολίων

Με αυτό τον τρόπο οπτικοποίησης δίνεται η δυνατότητα στο χρήστη να διαπιστώσει την χρονική εξέλιξη θετικών και αρνητικών σχολίων. Η πιο διαδεδομένη δομή αυτής της μεθόδου ορίζει ένα σύστημα αξόνων, όπου στον οριζόντιο άξονα βρίσκεται ο χρόνος και στον κάθετο το σύνολο των θετικών και αρνητικών σχολίων με διαφορετικό χρώμα.



Εικόνα 7. Χρονική Ανάλυση Αντικρουόμενων Σχολίων, Πηγή: Chen (2006)

2.2.2 Διαθέσιμα Εργαλεία

Many Eyes (http://www-958.ibm.com/software/data/cognos/manyeyes/)

Δωρεάν προς χρήση διαδικτυακό εργαλείο, το οποίο προσφέρει πολλών ειδών οπτικοποιήσεις κειμένου. Όχι απαραίτητα ιδανικές και τα δεδομένα του ιστότοπου IMDb, παρόλα αυτά υποστηρίζει είσοδο δεδομένων σε μορφή απλού κειμένου χωρίς δομή ώστε να γενικεύσει το εύρος των δεδομένων που μπορεί να οπτικοποιήσει. Το κυριότερο σενάριο χρήσης είναι η αποστολή μέρους δεδομένων προς το σύστημα και η παραγωγή ενός ή περισσότερων οπτικοποιήσεων, προσβάσιμων προς το ευρύ κοινό. Κατά το χρόνο συγγραφής της παρούσας εργασίας, δεν υπάρχει τρόπος αξιοποίησης του εργαλείου αυτού από εξωτερικές εφαρμογές.

2.3 Σχετικές ερευνητικές εργασίες

Οι Viigas et al (2007) περιγράφουν την ανάγκη μέσων οπτικοποίησης και υλοποιούν μια εφαρμογή η οποία παρέχει πολλών ειδών οπτικοποιήσεις δωρεάν, δεχόμενη ως είσοδο τα δεδομένα των χρηστών που την χρησιμοποιούν. Η μορφή εισόδου μπορεί να είναι και απλό κείμενο. Η συγκεκριμένη εφαρμογή αποτελεί την πιο ολοκληρωμένη δωρεάν εφαρμογή οπτικοποίησης στο διαδίκτυο.

Στο (Annet & Kondrak, 2008), προτείνεται μια νέα προσέγγιση στη ανάλυση συναισθήματος μέσω της χρήσης Support Vector Machines, και συγκρίνεται με παλαιότερες προσεγγίσεις., καθώς επίσης παρουσιάζεται η διενέργεια πειραμάτων και των αποτελεσμάτων αυτών. Αναλύεται επίσης η θεωρητική εφαρμογή των αποτελεσμάτων στο eNulog, ένα εργαλείο οπτικοποίησης απόψεων που περιέχονται σε διαδικτυακά blogs.

Ο Chen (2006), διενεργεί έρευνα με βάση τις κριτικές, χρηστών του διαδικτυακού τόπου Amazon.com, οι οποίες αφορούν το βιβλίο «The Da Vinci Code». Μέσα από την ανάλυση των θετικών και αρνητικών κριτικών γίνεται προσπάθεια απάντησης σε κρίσιμα ερωτήματα όπως το ποιές είναι οι διαφορές μεταξύ θετικών και αρνητικών κριτικών, από το τι αυτές οι διαφορές πηγάζουν, το πώς οι γνώμες αυτό των κριτικών διαμορφώνονται με το πέρασμα του χρόνου κα.

Οι Miloš Radovanović και Mirjana Ivanović (2008), αναλύουν ορισμένες προσεγγίσεις (automated classification και clustering) στο τομέα του προσδιορισμού των παγκόσμιων δομικών χαρακτηριστικών κειμένου(text patterns), μέσω της αναπαράστασης «Bag-of-Words». Εφαρμογές των προσεγγίσεων αυτών αναπαρίστανται μέσω εργαλείων οπτικοποίησης αποτελεσμάτων αναζήτησης στο παγκόσμιο ιστό, και οπτικοποίηση σχέσεων μεταξύ κοινών συγγραφέων(coauthorship) σε ένα σύνολο ερευνητικών εγγράφων.

Οι Diana Maynar et al (2012), αναλύουν οι δυσκολίες ανάπτυξης εργαλείων εξόρυξης γνώμης, καθώς και η ανάγκη ανάπτυξης εργαλείων αυτού του τύπου λόγω του αυξανόμενου όγκου πληροφοριών που παράγουν οι υπηρεσίες κοινωνικής δικτύωσης.

Οι Oelke et al (2009) παρουσιάζουν μια νέα προσέγγιση ανάλυσης μεγάλου όγκου κριτικών πελατών, ανάλογα με το ποία στοιχεία κρίνονται ως θετικά ή αρνητικά. Επιπλέον παρουσιάζονται τεχνικές οπτικοποίησης οι οποίες υποστηρίζεται ότι βοηθούν τον αναλυτή να αξιοποιήσει αποτελεσματικά μεγάλο αλλά και δομημένο όγκο πληροφοριών.

Οι Martin Potthas και Stephen Becker (2010) παρουσιάζουν τη τεχνολογία OpinionCloud, κύρια λειτουργία της οποίας αποτελεί η σύνοψη των γνωμών των χρηστών του διαδικτύου για οποιοδήποτε ζήτημα, ώστε να αποφεύγεται η προσπάθεια εξαγωγής συμπερασμάτων μέσω της ανάλυσης μεγάλου όγκου δεδομένων.

2.4 Εργαλεία

Για την ολοκλήρωση της παρούσας εργασίας χρησιμοποιήθηκαν τα εξής εργαλεία:

JFreeChart (http://www.jfree.org/jfreechart/)

Δωρεάν προς χρήση βιβλιοθήκες JAVA, οι οποίες επιτρέπουν τη δημιουργία πολλών ειδών διαγραμμάτων, όπως ιστογράμματα, γραφήματα τύπου πίτας(Pie Chart), γραφήματα με μπάρες(Bar Charts) κα. Αποτέλεσε το κορμό του μέρους οπτικοποίησης της παρούσας εργασίας.

OpenCloud (http://opencloud.mcavallo.org/)

Δωρεάν προς χρήση βιβλιοθήκες JAVA, οι οποίες υποστηρίζουν τη δημιουργία Tag Cloud αναπαραστάσεων σε μορφή HTML. Χρησιμοποιήθηκε για την υλοποίηση των δυο τύπων οπτικοποίησης που υποστηρίζει η εφαρμογή οι οποίες αφορούν την συχνότητα εμφάνισης όρων σε κείμενα.

Weka (http://www.cs.waikato.ac.nz/ml/weka/)

Δωρεάν προς χρήση εφαρμογή και βιβλιοθήκες JAVA. Αποτελεί εργαλείο Εξόρυξης Δεδομένων και περιέχει πληθώρα υλοποιημένων αλγορίθμων για το σκοπό αυτό. Αξιοποιήθηκε ο αλγόριθμος LibSVM τον οποίο περιέχει υλοποιημένο στο κομμάτι της κατηγοριοποίησης απόψεων.

IcePdf (http://icesoft.org)

Δωρεάν προς χρήση βιβλιοθήκες JAVA, οι οποίες υποστηρίζουν την ανάγνωση αρχείων σε μορφή PDF . Χρησιμοποιήθηκε ώστε να ενσωματωθεί στη εφαρμογή η δυνατότητα ανάγνωσης και προβολής αρχείων τύπου PDF, με κύριο σκοπό τη προβολή του εγχειριδίου χρήσης μέσα από το περιβάλλον της εφαρμογής.

3

Σχεδίαση

3.1 Ροή Δεδομένων

Η εφαρμογή η οποία υλοποιήθηκε συλλέγει, επεξεργάζεται και παρουσιάζει δεδομένα από τις κριτικές ταινιών από τον ιστότοπο IMDb(http://www.imdb.com). Τα στάδια χειρισμού των δεδομένων αυτών είναι:

- Συλλογή των δεδομένων από το IMDb με την χρήση Web Crawler οποίος υλοποιείται στα πλαίσια της υλοποίησης του συνόλου της εφαρμογής.
- Επεξεργασία και εγγραφή των δεδομένων σε δυο τύπους αρχείων στο δίσκο, XML και JAVA Serializable.
- Χειρισμός των JAVA Serializable αρχείων σύμφωνα με επιλογή του χρήστη με δύο τρόπους:

    1. - Κατηγοριοποίηση των δεδομένων.

       - Εγγραφή κατηγοριοποιημένων δεδομένων σε νέα JAVA Serializable αρχεία στο δίσκο.

       - Χρήση των νέων αρχείων για τη δημιουργία οπτικοποίησης και προβολή στο χρήστη.

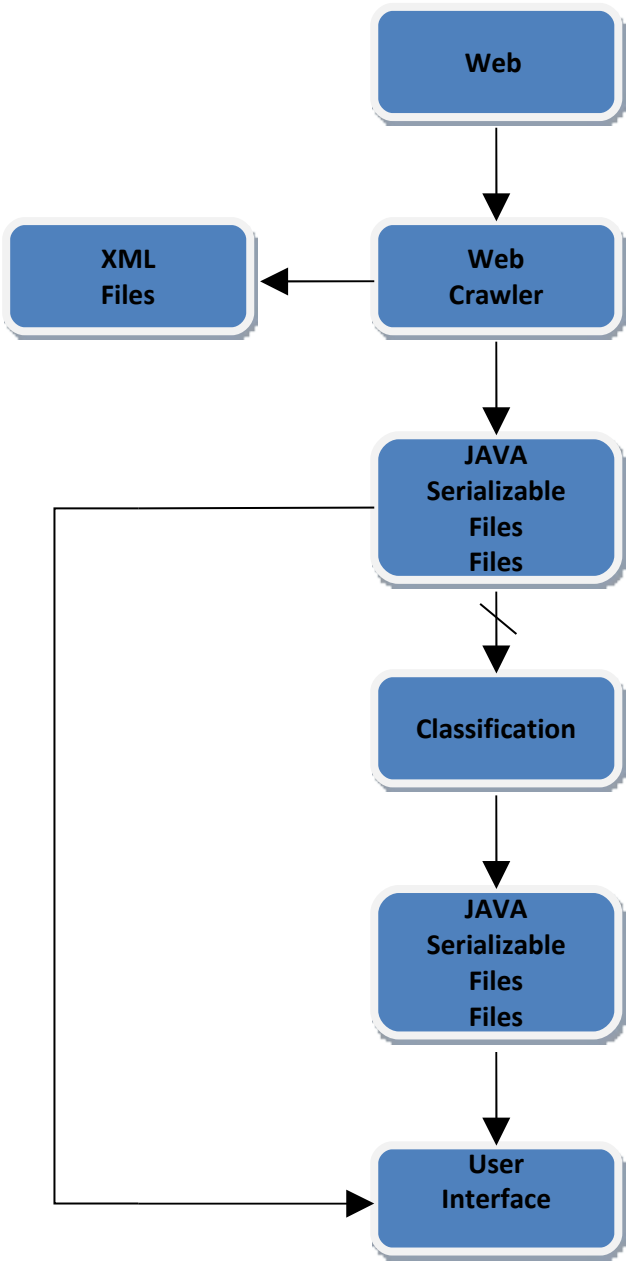
    2. Χρήση των αρχικών JAVA Serializable αρχείων για τη δημιουργία οπτικοποίησης και προβολή στο χρήστη.

Ο πρώτος τρόπος χειρισμού δεδομένων είναι και ο προτεινόμενος(Default).

Επισημάνεται ότι η συλλογή των δεδομένων από το διαδίκτυο υλοποιείται αρχικά επαναληπτικά για ένα σύνολο ταινιών, ώστε η εφαρμογή να διαθέτει έναν αρχικό όγκο δεδομένων προς χειρισμό.

Πέραν του γεγονότος αυτού, η εφαρμογή παρέχει στο χρήστη τη δυνατότητα άντλησης δεδομένων για μία μεμονωμένη ταινία ώστε να εμπλουτίσει τα διαθέσιμα προς οπτικοποίηση δεδομένα.

Παρατίθεται σχηματική αναπαράσταση τις ροής δεδομένων στην εφαρμογή:



Εικόνα 8. Ροή Δεδομένων

3.2 Web

3.2.1 Δομή δεδομένων IMDb

Όπως έχει είδη αναφερθεί ο ιστότοπος IMDb(http://www.imdb.com) φιλοξενεί κριτικές χρηστών για χιλιάδες κινηματογραφικές, και όχι μόνο, ταινίες.

Οι κριτικές αυτές τοποθετούνται σε σελίδες του ιστοτόπου ανά ομάδες των δέκα (10) σχολίων, και κάθε κριτική ακολουθεί πιστά μια συγκεκριμένη μορφοποίηση αφού υποβληθεί από τον εκάστοτε χρήστη. Η μορφοποίηση αυτή αποτελεί ουσιαστικά μια διάταξη των επιμέρους χαρακτηριστικών μιας κριτικής με τη χρήση της γλώσσας μορφοποίησης περιεχομένου HTML.

Κάθε ιστοσελίδα η οποία αφορά μια συγκεκριμένη ταινία, αποτελείτε από τα εξής επιμέρους χαρακτηριστικά:

- Τίτλος ταινίας
- Φίλτρα διάταξης κριτικών χρηστών
- Υπερσύνδεσμοι πλοήγησης για τις κριτικές, ανά ιστοσελίδα
- Συνολικός αριθμός κριτικών χρηστών οι οποίες έχουν υποβληθεί

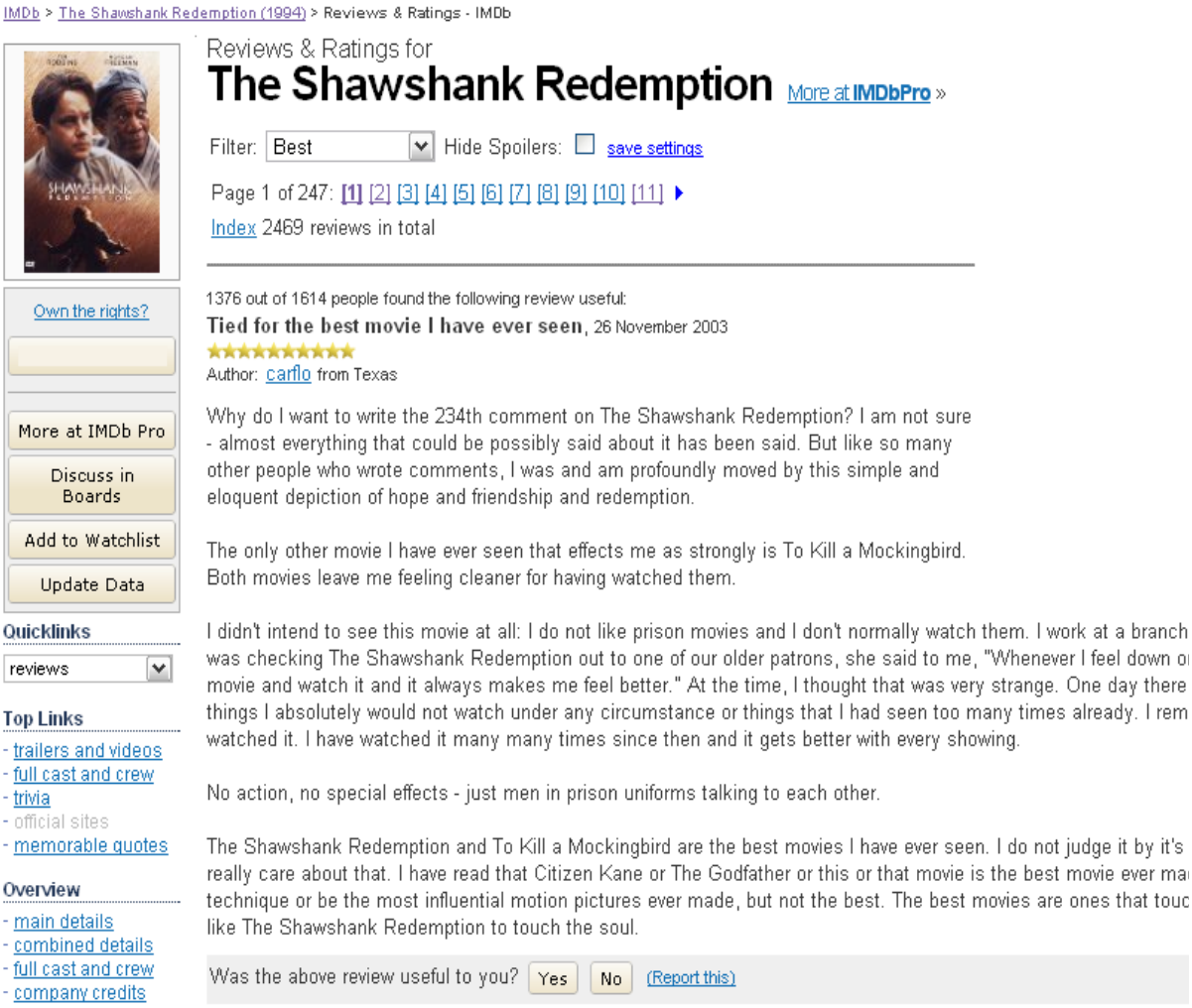
Κάθε κριτική αποτελείτε από τα εξής κύρια χαρακτηριστικά:

- Τίτλος σχολίου
- Ημερομηνία υποβολής από το χρήστη
- Αξιολόγηση ταινίας από το χρήστη σε κλίμακα 0 έως 10(άριστο)
- Το όνομα χρήστη στο IMDb, που υπέβαλε τη κριτική
- Τοποθεσία χρήστη
- Το κύριο σώμα του σχολίου, όπου περιέχεται και η άποψη του χρήστη
- Την αναλογία ποσοτήτων χρηστών οι οποίοι θεώρησαν ότι το συγκεκριμένο σχόλιο ήταν κατατοπιστικό

Πρέπει σε αυτό το σημείο να επισημανθεί ότι εκ των παραπάνω χαρακτηριστικών, η Αξιολόγηση ταινίας και η Τοποθεσία χρήστη, δεν είναι υποχρεωτικό να υποβληθούν από τον εκάστοτε χρήστη. Γεγονός το οποίο σημαίνει ότι πολλές κριτικές μπορεί να μη περιέχουν τα χαρακτηριστικά αυτά.

Το σύνολο των κριτικών αυτής της μορφής για κάθε ταινία, αποτελεί τα δεδομένα μελέτης και αξιοποίησης της εφαρμογής που αναπτύχθηκε στα πλαίσια της παρούσας εργασίας(Dataset).

Παρατίθεται εικόνα(screenshot) μιας σελίδας κριτικών από την ταινία «The Shawshank Redemption», όπου είναι εμφανή τα παραπάνω χαρακτηριστικά.



Εικόνα 9. Κριτική IMDb, Πηγή: http://www.imdb.com/title/tt0111161/reviews

3.3 Web Crawler

Ο Web Crawler αποτελεί το κομμάτι της εφαρμογής το οποίο είναι υπεύθυνο για τη συλλογή των κατάλληλων δεδομένων προς επεξεργασία για τις ταινίες του IMDb.

Η λειτουργία του μπορεί να περιγραφεί ως εξής:

- Επισκέπτεται μια συγκεκριμένη ιστοσελίδα μέσω ενός μοναδικού URL
- Εντοπίζει τα σημεία τις σελίδας τα οποία περιέχουν τα χαρακτηριστικά της ταινίας που αντιπροσωπεύει το URL, και αποθηκεύει στη μνήμη τη πληροφορία
- Εντοπίζει τα σημεία της σελίδας τα οποία περιέχουν τα χαρακτηριστικά των κριτικών για την ταινία και σαρώνει στη μνήμη όλες τις κριτικές των χρηστών. Οι κριτικές χρηστών σαφώς και βρίσκονται σε πολλές διαφορετικές ιστοσελίδες οπότε ο Web Crawler επισκέπτεται όλες τις σχετικές σελίδες.
- Αποθηκεύει τα δεδομένα σε δυο ειδών αρχεία. XML και JAVA Serializable.

Αφού επιτελεστεί επιτυχώς η διαδικασία που περιγράφεται με τα παραπάνω βήματα, το αποτέλεσμα είναι δυο αρχεία τα οποία βρίσκονται στο δίσκο και αποτελούν την επεξεργάσιμη μορφή δεδομένων για την εφαρμογή.

Η εφαρμογή χρησιμοποιεί τα αρχεία με μορφή JAVA Serializable, ενώ τα αρχεία σε μορφή XML αποτελούν το Dataset που παράγεται και μπορεί να αξιοποιηθεί μελλοντικά με πληθώρα τρόπων εντός και εκτός του πλαισίου της παρούσας εργασίας.

Παρατίθεται η δομή των XML αρχείων με τη χρήση XML Schema:

```
<?xml version="1.0" encoding="UTF-8" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">

<xs:element name="movie">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="review">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="rating" type="xs:string"/>
            <xs:element name="author" type="xs:string"/>
            <xs:element name="area" type="xs:string"/>
            <xs:element name="title" type="xs:string"/>
            <xs:element name="date"/>
                    <xs:complexType>
                      <xs:sequence>
                        <xs:element name="year" type="xs:string"/>
                        <xs:element name="month" type="xs:string"/>
                        <xs:element name="day" type="xs:string"/>
                      </xs:sequence>
                    </xs:complexType>
            </xs:element>
            <xs:element name="people" type="xs:string"/>
            <xs:element name="people1" type="xs:string"/>
            <xs:element name="comment" type="xs:string"/>
          </xs:sequence>
        </xs:complexType>
     </xs:element>
    </xs:sequence>
  </xs:complexType>
</xs:element>

</xs:schema>
```

3.4 Σχεδιασμός και χρήση Αρθρωμάτων κώδικα

Η εφαρμογή η οποία υλοποιήθηκε είναι γραμμένη εξολοκλήρου σε JAVA. Αποτελείτε από κώδικα ο οποίος γράφτηκε αποκλειστικά για την παρούσα εφαρμογή. Αλλά και από κώδικα ο οποίος έχει συγγραφεί εκ των προτέρων προς χρήση σε άλλου είδους εφαρμογές πέρα από τα πλαίσια της παρούσας. Κώδικας δωρεάν προς χρήση (OpenSource) με τη μορφή βιβλιοθηκών JAVA.

3.4.1 Σχεδιασμός Αρθρωμάτων

Άρθρωμα Web Crawler

Έχει υλοποιηθεί εξολοκλήρου για χρήση στη παρούσα εφαρμογή. Η μεθοδολογία άντλησης και συλλογής των δεδομένων από τις ιστοσελίδες είναι επίσης αποτέλεσμα της παρούσας εργασίας και μόνο.

Άρθρωμα δημιουργίας αρχείων

Έχει υλοποιηθεί εξολοκλήρου για χρήση στη παρούσα εφαρμογή. Αποτελείται κυρίως από τις κλάσεις JAVA οι οποίες συνθέτουν το περιεχόμενο εγγραφής σε αρχεία.

Άρθρωμα επεξεργασίας δεδομένων

Έχει υλοποιηθεί εξολοκλήρου για χρήση στη παρούσα εφαρμογή. Οι μέθοδοι οπτικοποίησης οι οποίες χρησιμοποιούνται απαιτούν συγκεκριμένη μορφή εισόδου των δεδομένων. Τα μέρη κώδικα αυτής της κατηγορίας είναι υπεύθυνα για τον μετασχηματισμό αυτό.

Άρθρωμα Γραφικής Διεπαφής χρήστη

Το περιβάλλον διεπαφής με το χρήστη, καθώς και ο σχεδιασμός αυτού, έχει υλοποιηθεί εξολοκλήρου για χρήση στη παρούσα εφαρμογή. Το σύνολο των συνθετικών στοιχείων του περιβάλλοντας (Components) υλοποιείται με τη χρήση των βιβλιοθηκών JAVA Swing & AWT.

Άρθρωμα Κατηγοριοποίησης δεδομένων

Για τον σκοπό της κατηγοριοποίησης των δεδομένων τα οποία αντλήθηκαν από το IMDb, γίνεται χρήση των βιβλιοθηκών JAVA του εργαλείου εξόρυξης δεδομένων Weka (http://www.cs.waikato.ac.nz/ml/weka/). Πιο συγκεκριμένα γίνεται η χρήση στη οποία βρίσκεται υλοποιημένος ο αλγόριθμος LibSVM, οποίος χρησιμοποιείται για τν κατηγοριοποίηση των κριτικών των χρηστών.

Άρθρωμα Δημιουργίας Οπτικοποιήσεων

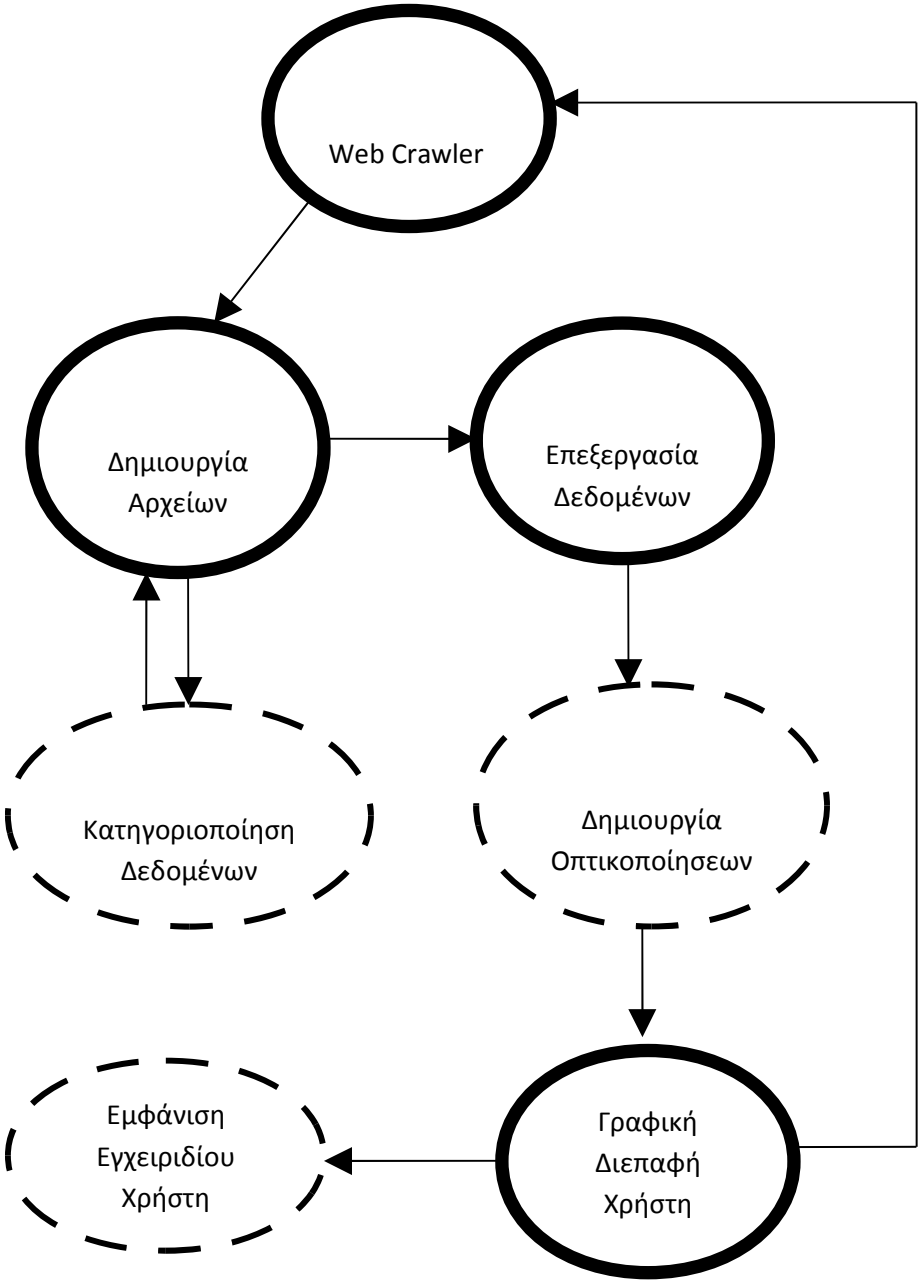
Για την υλοποίηση των οπτικοποιήσεων που είναι διαθέσιμες στην εφαρμογή χρησιμοποιούνται κλάσεις από τις OpenSource βιβλιοθήκες JFreeChart (http://www.jfree.org/jfreechart/) όπως επίσης κλάσεις από τη OpenSource βιβλιοθήκη

OpenCloud (http://opencloud.mcavallo.org/). Οι κλάσεις αυτές επεκτείνονται κατάλληλα σε κάθε περίπτωση, ώστε να παραχθεί το τελικό αποτέλεσμα.

Άρθρωμα εμφάνισης εγχειριδίου χρήστη

Η εφαρμογή παρέχει στο χρήστη τη δυνατότητα να συμβουλευτεί σχετικό εγχειρίδιο χρήσης ώστε να ικανοποιήσει κατάλληλα τις ανάγκες του. Το εν λόγω εγχειρίδιο είναι μορφής PDF και για την ανάγνωσή του μέσα από το γραφικό περιβάλλον της εφαρμογής γίνεται χρήση των OpenSource βιβλιοθηκών IcePdf (http://icesoft.org).

Στη συνέχεια παρατίθεται διάγραμμα αλληλεπίδρασης Αρθρωμάτων:



Εικόνα 10. Αλληλεπίδραση Αρθρωμάτων

3.4.2 Δομή Γραφικής Διεπαφής Χρήστη

Η εφαρμογή η οποία αναπτύχθηκε στα πλαίσια της παρούσας εργασίας προσφέρει ένα γραφικό περιβάλλον ώστε ο εκάστοτε χρήστης να έχει τη δυνατότητα προβολής και χειρισμού των οπτικοποιήσεων για τα σχόλη των χρηστών του ιστοτόπου IMDb. Το γραφικό περιβάλλον, τα ονόματα και οι περιγραφές των στοιχείων τα οποία το αποτελούν είναι στην Αγγλική γλώσσα. Ακολουθεί αναλυτική περιγραφή της δομής της κεντρικής οθόνης του περιβάλλοντος.

Η κεντρική οθόνη της εφαρμογής περιέχει τα εξής στοιχεία:

- Το κεντρικό μενού, όπου υπάρχει οι επιλογές File,Chart Options και Help.
- Τη λίστα επιλογής ταινίας. Η λίστα περιέχει αρχικά τους τίτλους των 250 δημοφιλέστερων ταινιών του ιστότοπου IMDb.
- Τη λίστα επιλογής τύπου οπτικοποίησης, όπου ο χρήστης επιλέγει το είδος οπτικοποίησης που επιθυμεί να δει, και το οποίο θα αφορά τις κριτικές χρηστών για την ταινία η οποία είναι επιλεγμένη στη λίστα επιλογής ταινίας. Τα διαθέσιμα προς επιλογή είδη οπτικοποίησης είναι:

    1. Deviation Chart
       Καμπύλες συνόλων απόψεων των χρηστών του IMDb, η οποίες παρουσιάζουν τη μεταβολή της άποψης στο χρόνο.

    2. Deviation Rating Mean Chart
       Καμπύλη η οποία παρουσιάζει την εξέλιξη στο χρόνο, του μέσου όρου των τιμών αξιολόγησης ταινιών από τους χρήστες του IMDb.

    3. Spider Web Chart
       Συγκεντρωτικό διάγραμμα απόψεων χρηστών ανά περιόδους, με πολλαπλούς άξονες οι οποίοι αντιπροσωπεύουν τις περιόδους αυτές.

    4. Stacked Bar Chart
       Διάγραμμα με μπάρες, το οποίο παρουσιάζει το ποσοστό επί τις εκατό, των διαφορετικών γνωμών χρηστών ανά χρονικές περιόδους.

    5. Tag Cloud Chart
       Οπτικοποίηση του συνόλου λέξεων με τη μεγαλύτερη συχνότητα εμφάνισης στις κριτικές χρηστών.

    6. Frequent Words Chart
       Οπτικοποίηση των συνόλων λέξεων με τη μεγαλύτερη συχνότητα εμφάνισης στις «Θετικές» και «Αρνητικές» κριτικές χρηστών.

- Το κουμπί «Show Plot» το οποίο ενεργοποιεί τη διαδικασία δημιουργίας οπτικοποίησης.
- Η οπτικοποίηση προβάλλεται σύμφωνα με τις αρχικές ρυθμίσεις στο αρχικά κενό πλαίσιο το οποίο υπάρχει στη κεντρική οθόνη.

4

Υλοποίηση

Η εφαρμογή η οποία αναπτύχθηκε επιτελεί τρείς επιμέρους, βασικές λειτουργίες ώστε να υπάρξει το τελικό αποτέλεσμα οπτικοποίησης:

- Συλλέγει, επεξεργάζεται και αποθηκεύει στο δίσκο, τα δεδομένα από τις κριτικές χρηστών του διαδικτυακού τόπου IMDb.
- Κατηγοριοποιεί τις κριτικές των χρηστών σε Θετικές, Αρνητικές και Ουδέτερες, και αποθηκεύει τη κατηγοριοποιημένη πληροφορία σε νέα αρχεία στο δίσκο.
- Οπτικοποιεί τις κριτικές των χρηστών χρησιμοποιώντας είτε τα κατηγοριοποιημένα είτε τα αρχικά δεδομένα, σε μία από τις έξι μορφές οπτικοποίησης οι οποίες έχουν υλοποιηθεί, και εμφανίζει το αποτέλεσμα στο χρήστη.

Ακολουθεί αναλυτική περιγραφή της μεθόδου υλοποίησης της εφαρμογής, περιγράφοντας σε βάθος τις επιμέρους λειτουργίες που επιτελεί. Θα παρουσιαστούν επίσης σενάρια εκτέλεσης της εφαρμογής, καθώς και αξιολόγηση των αποτελεσμάτων.

4.1 Δομικές Κλάσεις

4.1.1 Review

Τα δεδομένα από τις κριτικές των χρηστών πρέπει να συλλέγονται και να αποθηκεύονται με δομημένο τρόπο ώστε να επιτρέπεται η μελλοντική ανάγνωση, επεξεργασίας και γενικότερα αξιοποίηση των δεδομένων.

Για το σκοπό αυτό δημιουργείται μια νέα κλάση JAVA, με το όνομα **Review**. Στη κλάση Review ενσωματώνεται η λειτουργικότητα της κλάσης-διεπαφής(Interface) , java.io.Serializable[1]. Με τον τρόπο αυτό εξασφαλίζεται στη κλάση Review η ιδιότητα σειριοποίησης (Serialization), γεγονός το οποίο σημαίνει ότι γίνεται δυνατή η αποθήκευση σε αρχείο αντικειμένων τύπου Review, με συγκεκριμένη κωδικοποίηση. Η αποθήκευση με αυτό τον τρόπο επιτρέπει τη μετέπειτα, ταχύτατη ανάγνωση και επαναφορά στη μνήμη των δεδομένων σε μορφή αντικειμένων Review.

Η κλάση Review σχεδιάστηκε ώστε να αντιπροσωπεύει ως αντικείμενο μία κριτική χρήστη για μία ταινία. Οπότε λόγω της απαραίτητης αντιστοιχίας η οποία θα πρέπει να υπάρχει, η κλάση φέρει τα εξής χαρακτηριστικά:

- *rating*
  Αντιστοιχεί στη βαθμολογία την οποία έχει κάθε χρήστης τη δυνατότητα να προσάψει στη κριτική του, ως βαθμό αξιολόγησης της ταινίας. Είναι τύπου Int, ακέραιος.

- people & people1
  Τα χαρακτηριστικά αυτά αντιστοιχούν στα σύνολα άλλων χρηστών του IMDb, οι οποίοι αξιολόγησαν μια κριτική του χρήστη ως χρήσιμη.
  Σε αυτό το παράδειγμα:
  "20 out of 41 people found the following review useful" στο χαρακτηριστικό people εκχωρείται ο αριθμός 20, ενώ στο people1 ο αριθμός 41.
  Ουσιαστικά η αναλογία των δύο αυτών χαρακτηριστικών αποτελεί το ποσοστό των χρηστών οι οποίοι θεώρησαν ότι η κριτική είναι χρήσιμη προς τον αριθμό των χρηστών οι οποίοι διάβασαν τη κριτική γενικότερα. Και τα δυο χαρακτηριστικά είναι τύπου Int.

- title
  Αποτελεί το τίτλο τον οποίο έδωσε στη κριτική του ένας χρήστης. Είναι τύπου String, αλφαριθμητικό .

- area
  Περιέχει τη ονομασία της περιοχής από όπου ο χρήστης υπέβαλε τη κριτική του. Είναι τύπου String.

- author
  Περιέχει το όνομα χρήστη ο οποίος υπέβαλε τη κριτική, όπως αυτό εμφανίζεται στο IMDb. Είναι τύπου String.

- comment
  Περιέχει το περιεχόμενο/κύριο σώμα της κριτικής του χρήστη. Είναι τύπου String.

- date
  Περιέχει την ημερομηνία την οποία υπέβαλε ο χρήστης τη κριτική του. Είναι τύπου java.util.Date[2].

- URI
  Περιέχει τη διεύθυνση URI της κριτικής χρήστη. Είναι τύπου String.

- OpinionLevel
  Χαρακτηριστικό το οποίο αξιοποιείται κατά τη φάση της κατηγοριοποίησης γνώμης του χρήστη. Πιθανές τιμές του χαρακτηριστικού αυτού είναι οι: Agree, Strongly Agree, Neutral, Disagree, Strongly Disagree. Το χαρακτηριστικό αυτό καθώς και ο τρόπος ανάθεσης τιμών σε αυτό θα περιγραφούν αναλυτικά στο κεφάλαιο της κατηγοριοποίησης. Το χαρακτηριστικό είναι τύπου String.

- Training
  Χαρακτηριστικό το οποίο αξιοποιείται κατά τη φάση της κατηγοριοποίησης γνώμης του χρήστη. Χρησιμοποιείται ώστε να καθοριστεί εάν το αντικείμενο Review, θα χρησιμοποιηθεί για την εκπαίδευση του αλγορίθμου κατηγοριοποίησης ή όχι. Το χαρακτηριστικό αυτό καθώς και ο τρόπος ανάθεσης τιμών σε αυτό θα περιγραφούν αναλυτικά στο κεφάλαιο της κατηγοριοποίησης. Το χαρακτηριστικό είναι τύπου boolean, δυαδική.

Επίσης υλοποιούνται στη κλάση μέθοδοι, ώστε να είναι εύκολη η προσπέλαση των τιμών των χαρακτηριστικών κάθε αντικειμένου.

Για να μη διακόπτεται η εκτέλεση της εφαρμογής του crawler όταν υπάρχουν προβλήματα στη σελίδα έχει δημιουργηθεί μια επιπλέον κλάση, η:

4.1.2 FileFormatException

Η κλάση αυτή δημιουργήθηκε για το χειρισμό σφαλμάτων τα οποία είναι πιθανόν να συναντήσει ο Web Crawler αναζητώντας τη κατάλληλη πληροφορία. Αποτελεί επέκταση της java.lang.Exception[3].

Παρατίθεται ο κώδικας της κλάσης **FileFormatException**

```
public class FileFormatException extends Exception{
    String movie_id;
    public FileFormatException(String movie_id){
        this.movie_id=movie_id;
    }
    public String toString(){
        return this.movie_id;
    }

}
```

4.2 Συλλογή και σειριοποίηση κριτικών

4.2.1 Λειτουργία SerializeAndXML

Η μέθοδος αυτή είναι υπεύθυνη για την επίσκεψη των ιστοσελίδων του IMDb οι οποίες περιέχουν κριτικές χρηστών, τον εντοπισμό στον κώδικα HTML των σελίδων αυτών του κατάλληλου περιεχομένου, την δημιουργίας αντικειμένων τύπου Review, και την αποθήκευση αυτών σε αρχεία στο δίσκο. Ουσιαστικά αποτελεί τον Web Crawler.

Κάθε σχετική ιστοσελίδα περιγράφεται από μία διεύθυνση URL τύπου: http://www.imdb.com/title/tt000000/reviews?start=0

Όπου tt000000, ορίζεται ο μοναδικός αναγνωριστικός αριθμός τις κάθε ταινίας που βρίσκεται στο IMDb. Κάθε αναφορά σε αυτό το κωδικό θα γίνεται πλέον με τον όρο ID. Όπου start=0, ορίζεται ο αύξον αριθμός της σελίδας κριτικών για τη ταινία η οποία ορίζεται από το ID στη διεύθυνση URL. Το βήμα για την μετάβαση στην επόμενη σελίδα είναι 10, δηλαδή η πρώτη σελίδα σχολίων θα αντιπροσωπεύεται από το URL:

http://www.imdb.com/title/tt000000/reviews?start=0
η δεύτερη από το:

http://www.imdb.com/title/tt000000/reviews?start=10 και ούτω καθεξής.

Χάριν παραδείγματος το ID για την συγκεκριμένη υποθετική ταινία είναι 000000.

Ο Web Crawler επισκέπτεται τη κάθε σελίδα της ταινίας στην οποία αντιστοιχεί το ID, το οποίο δέχεται η μέθοδος ως παράμετρο, με τη χρήση της κλάσης java.net.URLConnection[4], και αποθηκεύει το σύνολο του κώδικα HTML τις σελίδας στη μνήμη. Ύστερα αναζητά λέξεις κλειδιά μέσα στο σύνολο του κώδικα HTML, οι οποίες σηματοδοτούν την ύπαρξη χρήσιμης πληροφορίας σε αυτό το σημείο στο κώδικα. Για παράδειγμα ο HTML όρος `"<small>Author:",` ακολουθείτε πάντοτε από το όνομα του χρήστη ο οποίος υπέβαλε μία κριτική, ένα από τα βασικά στοιχεία που χρειάζονται, άρα αποτελεί όρο κλειδί. Εφόσον ο Web Crawler «συναντήσει» έναν όρο κλειδί, εντοπίζει και αποθηκεύει τη σχετική πληροφορία, διαδικασία η οποία επαναλαμβάνεται για το σύνολο των κριτικών, για το σύνολο των ιστοσελίδων οι οποίες περιέχουν κριτικές. Αφού εντοπιστεί το κύριο σώμα της κριτικής κάθε χρήστη, προτού αποθηκευθεί καθαρίζεται από κατάλοιπα κώδικα HTML τα οποία είναι πιθανό να περιέχει.

Επισημάνεται επίσης, ότι ο χρήστης του IMDb δεν υποχρεούται να υποβάλει βαθμό αξιολόγησης, αρά είναι πιθανό το χαρακτηριστικό rating να μην έχει η τιμή. Σε αυτή τη περίπτωση εκχωρείται η τιμή πέντε(5) για λόγους ομοιομορφίας, καθώς αποτελεί τη διάμεσο του εύρους τιμών.

Το τελικό αποτέλεσμα είναι μια λίστα στη μνήμη η οποία περιέχει αντικείμενα τύπου Review, για τη ταινία με το προσδιορισμένο ID. Πρέπει να σημειωθεί σε αυτό το σημείο ότι τα γνωρίσματα OpinionLevel και Training κάθε αντικειμένου δεν παίρνουν περιεχόμενο

ανάλογα με κάποια τιμή του κώδικα HTML κατά τη διάρκεια της διαδικασίας που περιγράφεται. Στη παρούσα φάση ορίζονται απλώς ορίζονται ως κενά

Τα αντικείμενα αυτά εγγράφονται κωδικοποιημένα σε αρχείο τύπου PHC, με όνομα αρχείου το ID της ταινίας. Επίσης κατά τη διάρκεια της αναζήτησης της πληροφορίας στο κώδικα HTML, ο Web Crawler εγγράφει με τη σειρά τη πληροφορία που αντιστοιχεί σε μία κριτική, σε ένα αρχείο τύπου XML, με όνομα επίσης το ID της ταινίας.

Παρατίθεται δείγμα ενός αρχείου XML

```
<?xml version="1.0" encoding="UTF-8"?>
<movie title="Sherlock Jr.">
  <review>
    <rating>9</rating>
    <author>imogensara_smith</author>
    <area>New York City</area>
    <title>Through the Movie Screen</title>
    <date>
        <year>2006</year>
        <month>9</month>
        <day>13</day>
    </date>
    <people>24</people>
    <people1>26</people1>
    <comment>  Buster Keaton's most surreal movie sprang from his
insistence on logic and realism. His tribute …
</comment>
  </review>
  <review>
    <rating>5</rating>
    <author>Snow Leopard</author>
    <area>Ohio</area>
    <title>Astounding Creativity</title>
    <date>
    ……

</movie>
```

4.3 Κατηγοριοποίηση

4.3.1 Λειτουργία ClassifySingleMovie

Η μέθοδος είναι υπεύθυνη για τη κατηγοριοποίηση των κριτικών χρηστών για μία συγκεκριμένη ταινία, της οποίας το ID, δέχεται ως παράμετρο.

Αποκωδικοποιεί το αντίστοιχο αρχείο PHC που παρήγαγε ο Web Clawler, το οποίο βρίσκεται στο δίσκο, και σηκώνει όλα τα δεδομένα στη μνήμη με τη μορφή λίστας αντικειμένων Review.

Για το σκοπό της κατηγοριοποίησης πρέπει τα αντικείμενα στη λίστα να χωριστούν σε δύο ομάδες. Η μία ομάδα θα αποτελέσει τα δεδομένα εκπαίδευσης του αλγορίθμου, και η άλλη τα δεδομένα ελέγχου. Το ποσοστό των κριτικών οι οποίες θα χρησιμοποιηθούν ως δεδομένα εκπαίδευσης ορίζεται στο 5%.

Κατά τη διαδικασία επιλογής των δεδομένων εκπαίδευσης, τα αντικείμενα τα οποία τελικώς επιλέγονται τροποποιούνται. Το χαρακτηριστικό Training λαμβάνει τιμή “true”, ενώ το χαρακτηριστικό OpinionLevel λαμβάνει μια εκ των τιμών Agree, Strongly Agree, Neutral, Disagree και Strongly Disagree ανάλογα με τη τιμή γνωρίσματος αξιολόγησης ταινίας του χρήστη του IMDb, δηλαδή το χαρακτηριστικό rating του αντικειμένου. Ο η παραπάνω τιμές αποτελούν ένα τρόπο υποκειμενικής απόδοσης του χαρακτηριστικού rating, στην γενικότερη προσπάθεια ανάλυσης συναισθήματος.

Τα αντικείμενα τα οποία δεν επιλέγονται ως δεδομένα εκπαίδευσης, οπότε η τιμή του χαρακτηριστικού Training δεν αλλάζει(false), και το χαρακτηριστικού OpinionLevel παραμένει δίχως τιμή, καθώς πρέπει να λάβει τιμή βάση των αποτελεσμάτων του αλγόριθμου κατηγοριοποίησης ο οποίος θα χρησιμοποιηθεί.

Αφού ολοκληρωθεί η διαδικασία διαχωρισμού σε δεδομένα εκπαίδευσης και ελέγχου, δημιουργείται στο δίσκο φάκελος με όνομα “training”,όπου δημιουργούνται πέντε υποφάκελοι με ονόματα Agree, Strongly Agree, Neutral, Disagree και Strongly Disagree. Κατ’ αντιστοιχία με τη τιμή του χαρακτηριστικού OpinionLevel, το κύριο σώμα κάθε κριτικής αποθηκεύεται σε έναν από αυτούς τους φακέλους με τη μορφή αρχείου TXT, και όνομα αρχείου τη τιμή του χαρακτηριστικού URI. Δημιουργείται επίσης φάκελος με όνομα “test” όπου με το ίδιο τρόπο αποθηκεύονται οι κριτικές οι οποίες επιλέχθηκαν ως δεδομένα ελέγχου.

Η παραπάνω διαδικασία επεξεργασίας δεδομένων, εκτελείται ώστε να τροποποιηθούν τα δεδομένα κατάλληλα ώστε να μπορεί να τα χειριστεί ο αλγόριθμος κατηγοριοποίησης LibSVM των βιβλιοθηκών του περιβάλλοντος εξόρυξης δεδομένων Weka.

Γίνεται κλίση του αλγορίθμου LibSVM με είσοδο τα δεδομένα ελέγχου, και αφού ολοκληρωθεί η διαδικασία κατηγοριοποίησης, τα κατηγοριοποιημένα πλέον αντικείμενα εγγράφονται σε νεο αρχείο τύπου PHC, σε διαφορετικό φάκελο όπου αποθηκεύονται τα αποτελέσματα κατηγοριοποίησης , με όνομα αρχείου το ID της εκάστοτε ταινίας.

Η κλήση του κώδικα κατηγοριοποίησης γίνεται όπως παρουσιάζεται παρακάτω. Η κλήση των βιβλιοθηκών Weka γίνεται μέσω έτοιμου κώδικα ο οποίος προσαρμόζεται στην εφαρμογή.

```
public static void processFilipp(ArrayList<Review>
testReviewsForFilm) {
        TextDirectoryLoader m = new TextDirectoryLoader();
        m.lang = "EN";
        m.convertTextToArff("ISO-8859-7", "survey\\training",
"survey\\training.arff", "training");
        m.convertToVectorArff("survey\\training.arff",
"survey\\trainingVector.arff", "training", 10);
        m.convertTextToArff("ISO-8859-7", "survey\\test",
"survey\\Test.arff", "test");
        m.convertTestToVectorArff("survey\\Test.arff",
"survey\\TestVector.arff");
        m.evaluateExternal("survey\\trainingVector.arff",
"survey\\TestVector.arff", testReviewsForFilm);

    }
```

4.3.2 Λειτουργία Deserialize

Η μέθοδος αυτή διαβάζει το αρχείο “movies_html.txt”, το οποίο περιέχει τα ID τα οποία αντιστοιχούν στις διακόσιες πενήντα (250) δημοφιλέστερες ταινίες του ιστοτόπου IMDb. Με τη χρήση των ID σηκώνει στη μνήμη, σε μορφή λίστας αντικειμένων Review, όλες τις κριτικές από τις 250 αυτές ταινίες. Οι κριτικές υπάρχουν στο δίσκο με τη μορφή αρχείων PHC.

4.3.3 Λειτουργία DeserializeAndClassify

Αποτελεί την μέθοδο μαζικής κατηγοριοποίησης. Ουσιαστικά καλεί τι μέθοδο ClassifySingleMovie για κάθε ID το οποίο περιέχεται στο αρχείο “movies_html.txt”, με τελικό αποτέλεσμα την κατηγοριοποίηση του συνόλου των 250 δημοφιλέστερων ταινιών.

4.4 Προετοιμασία για Οπτικοποίηση

4.4.1 Λειτουργία MakeSpiderWeb

Μέθοδος υπεύθυνη για την διαμόρφωση των δεδομένων κριτικών μίας ταινίας, σε μορφή κατάλληλη ώστε να δημιουργηθεί οπτικοποίηση τύπου Spider Web των βιβλιοθηκών JFreeChart.

Γίνεται ανάγνωση αρχείου τύπου PHC το οποίο αντιστοιχεί στο ID ταινίας το οποίο δέχεται η μέθοδος ως παράμετρο. Παράγεται λίστα αντικειμένων τύπου Review.

Η αρχική λίστα διασπάται σε νέες λίστες. Οι λίστες περιέχουν τις κριτικές οι οποίες υποβλήθηκαν σε συγκεκριμένες χρονικές περιόδους . Οι χρονικές περίοδοι είναι διάρκειας δυο ετών, και κάθε κριτική της οποίας η ημερομηνία υποβολής περιέχεται στην εκάστοτε περίοδο, προστίθεται στη αντίστοιχη λίστα. Στη συνέχεια μετράται η συχνότητα εμφάνισης, ανά χρονική περίοδο, της κάθε τύπου κριτικής ανάλογα με το περιεχόμενο του χαρακτηριστικού OpinionLevel κάθε αντικειμένου Review. Για παράδειγμα στα πλαίσια της διαδικασίας θα μετρηθεί η συχνότητα εμφάνισης των κριτικών με OpinionLevel “Strongly Agree”, την περίοδο 2006 έως 2008.

Το τελικό αποτέλεσμα της παραπάνω διαδικασίας είναι ένα αντικείμενο τύπου java.util.HashMap[5], το οποίο φέρει πέντε λίστες αντικειμένων τύπου Review, με κάθε λίστα να αντιστοιχεί και σε μια από τις πιθανές τιμές του OpinionLevel. Οι λίστες αυτές θα αποτελέσουν ουσιαστικά τους άξονες του τελικού διαγράμματος.

Καλείται η μέθοδος **opinionClassifierDemo.showResults** με παράμετρο το αντικείμενο HashMap ώστε να αρχίσει η αλληλουχία κλήσεων των κατάλληλων μεθόδων για την δημιουργία της τελικής οπτικοποίησης.

4.4.2 Λειτουργία MakeStackedBarChart

Μέθοδος υπεύθυνη για την διαμόρφωση των δεδομένων κριτικών μίας ταινίας, σε μορφή κατάλληλη ώστε να δημιουργηθεί οπτικοποίηση τύπου Stacked Bar των βιβλιοθηκών JFreeChart.

Γίνεται ανάγνωση αρχείου τύπου PHC το οποίο αντιστοιχεί στο ID ταινίας το οποίο δέχεται η μέθοδος ως παράμετρο. Παράγεται λίστα αντικειμένων τύπου Review.

Όπως και στην οπτικοποίηση τύπου Spider Web, η αρχική λίστα διασπάται σε νέες λίστες. Οι λίστες περιέχουν τις κριτικές οι οποίες υποβλήθηκαν σε συγκεκριμένες χρονικές περιόδους.

Υπολογίζονται τα ποσοστά Θετικών (Agree + Strongly Agree), Ουδέτερων (Neutral) και Αρνητικών (Disagree + Strongly Disagree) κριτικών ανά περίοδο (τιμές κατηγοριοποίησης OpinionLevel) και ανάγονται σε ποσοστά επί τις εκατό (% normalization). Τα αποτελέσματα

εισάγονται σε δομή δυσδιάστατου πίνακα φυσικών αριθμών, ώστε να είναι κατάλληλα προς επεξεργασία από τη μέθοδο **showresults** της κλάσης **StackedBarChart,** ώστε να δημιουργηθεί η τελική οπτικοποίηση.

4.4.3 Λειτουργία MakeStackedbarChartUsr

Μέθοδος όμοια με την MakeStackedBarChart. Η διαφοροποίηση έγκειται στο γεγονός ότι η παρούσα μέθοδος υπολογίζει τα ποσοστά Θετικών, Ουδέτερων και Αρνητικών κριτικών βάσει της τιμής του χαρακτηριστικού αξιολόγησης χρήστη, και όχι βάσει της τιμής του χαρακτηριστικού OpinionLevel, προϊόντος κατηγοριοποίησης. Μια κριτική χαρακτηρίζεται ως Θετική εφόσον η τιμή του χαρακτηριστικού rating είναι μεγαλύτερη του έξι(6), Αρνητική εφόσον δε ξεπερνά τη τιμή τέσσερα (4) και Ουδέτερη σε οποιαδήποτε άλλη περίπτωση. Υπενθυμίζεται ότι το εύρος τιμών που λαμβάνει το χαρακτηριστικό rating είναι το [1-10] .

4.4.4 Λειτουργία MakeDeviationRenderer

Μέθοδος υπεύθυνη για την διαμόρφωση των δεδομένων κριτικών μίας ταινίας, σε μορφή κατάλληλη ώστε να δημιουργηθεί μία εκ των δύο διαθέσιμων από την εφαρμογή οπτικοποιήσεων τύπου Deviation Renderer.

Γίνεται ανάγνωση αρχείου τύπου PHC το οποίο αντιστοιχεί στο ID ταινίας το οποίο δέχεται η μέθοδος ως παράμετρο. Παράγεται λίστα αντικειμένων τύπου Review.

Γίνεται ταξινόμηση με αύξουσα σειρά της λίστας, με βάση την ημερομηνία υποβολής κριτικής (χαρακτηριστικό date), με τη χρήση της μεθόδου **SortReviewsByDate.**

Στη συνέχεια γίνεται κλήση της μεθόδου showresults είτε της κλάσης DeviationChart είτε της κλάσης DeviationChartMean ώστε να ξεκινήσει η δημιουργία της αντίστοιχης οπτικοποίησης. Η λειτουργία των κλάσεων θα περιγραφεί αναλυτικά στη συνέχεια του παρόντος εγγράφου.

4.4.5 Λειτουργία MakeTagCloud

Μέθοδος υπεύθυνη για την διαμόρφωση των δεδομένων κριτικών μίας ταινίας, σε μορφή κατάλληλη ώστε να δημιουργηθεί οπτικοποίηση τύπου Tag Cloud, με χρήση της κλάσης Cloud των βιβλιοθηκών OpenCloud.

Γίνεται ανάγνωση αρχείου τύπου PHC το οποίο αντιστοιχεί στο ID ταινίας το οποίο δέχεται η μέθοδος ως παράμετρο. Παράγεται λίστα αντικειμένων τύπου Review.

Γίνεται καθαρισμός του κύριου σώματος της κάθε κριτικής ταινίας (χαρακτηριστικό comment) από κοινές και συχνές στη λόγο λέξεις (stopwords) ώστε να βελτιωθεί το τελικό αποτέλεσμα, με τη χρήση της μεθόδου **RemoveStopWords** η οποία δέχεται τη λίστα αντικειμένων Review ως παράμετρο.

Καλείται η μέθοδος **showresults** τις κλάσης Tag Cloud ώστε να ξεκινήσει η δημιουργία της αντίστοιχης οπτικοποίησης.

4.4.6 Λειτουργία MakeFrequentWords

Μέθοδος υπεύθυνη για την διαμόρφωση των δεδομένων κριτικών μίας ταινίας, σε μορφή κατάλληλη ώστε να δημιουργηθεί οπτικοποίηση τύπου FrequentWords. Η οπτικοποίηση αυτή υλοποιείται ουσιαστικά με τη χρήση δυο αντικειμένων Cloud, των βιβλιοθηκών OpenCloud. Το ένα αντικείμενο περιέχει τις λέξεις με την υψηλότερη συχνότητα εμφάνισης στις κριτικές χρηστών οι οποίες αξιολογούνται ως Θετικές, ενώ το άλλο περιέχει τις Αρνητικές.

Γίνεται ανάγνωση αρχείου τύπου PHC το οποίο αντιστοιχεί στο ID ταινίας το οποίο δέχεται η μέθοδος ως παράμετρο. Παράγεται λίστα αντικειμένων τύπου Review.

Γίνεται καθαρισμός του κύριο σώματος της κάθε κριτικής ταινίας (χαρακτηριστικό comment) από κοινές και συχνές στη λόγο λέξεις (stopwords) ώστε να βελτιωθεί το τελικό αποτέλεσμα, με τη χρήση της μεθόδου **RemoveStopWords** η οποία δέχεται τη λίστα αντικειμένων Review ως παράμετρο.

Γίνεται διάσπαση τις αρχικής λίστας αντικειμένων Review σε δύο επιμέρους λίστες, ώστε να διαχωριστούν οι Θετικές από τις Αρνητικές κριτικές. Κριτήριο διάσπασης αποτελεί η τιμή του χαρακτηριστικού rating. Εφόσον αυτή ξεπερνά το πέντε(5), η κριτική χαρακτηρίζεται ως Θετική. Σε διαφορετική περίπτωση χαρακτηρίζεται ως αρνητική.

Δημιουργούνται δυο αντικείμενα τύπου org.mcavallo.opencloud.Cloud[6], με τη χρήση των λιστών Θετικών και Αρνητικώς κριτικών, αφού καθαριστούν πρώτα από stopwords, με τη χρήση της μεθόδου RemoveStopWords.

Επιτελείται «ζύγισμα» με TF (πόσες φορές συνολικά εμφανίζεται η λέξη) ή IDF (σε πόσες διαφορετικές κριτικές εμφανίζεται η λέξη), στο σύνολο των Θετικών ή Αρνητικών κριτικών αντίστοιχα. Κάθε συχνότητα συνδέεται με την αντίστοιχη λέξη.

Καλείται η μέθοδος **showresults** της κλάσης FrequentWords, με παραμέτρους τα δυο αντικείμενα τύπου Cloud, ώστε να ξεκινήσει η δημιουργία της αντίστοιχης οπτικοποίησης.

4.5 Βοηθητικές λειτουργίες

4.5.1 Λειτουργία StoreToLogStoreToLog

Μέθοδος η οποία επιτελεί λειτουργία δημιουργίας ιστορικού δημιουργίας οπτικοποιήσεων. Καλείτε μέσα από κάθε μία από τις μεθόδους επεξεργασίας δεδομένων. Εγγράφει σε αρχείο με όνομα “log.txt”, τον χρόνο συστήματος (ημερομηνία), το ID ταινίας για την οποία δημιουργήθηκε οπτικοποίηση, καθώς και τον τύπο οπτικοποίησης.

4.5.2 Λειτουργία RemoveStopWords

Μέθοδος υπεύθυνη για τον καθαρισμό των κριτικών χρηστών από stopwords, τα οποία σε περίπτωση που δεν αφαιρεθούν επηρεάζουν δραματικά το τελικό αποτέλεσμα οπτικοποίησης. Οπότε η ανάγκη αφαίρεσης των stopwords είναι επιτακτική.

Συγχωνεύει τα χαρακτηριστικά comment (κύριο σώμα κριτικής), των αντικειμένων Review της λίστας την οποία δέχεται η μέθοδος ως παράμετρο, σε μία κοινή μεταβλητή τύπου γραμματοσειράς. Στην συνέχεια αφαιρεί τους ειδικούς με τη χρήση της μεθόδου String.Split[7].

Γίνεται ανάγνωση αρχείου με όνομα “stopwords.txt” το οποίο περιέχει τα stopwords τα οποία πρέπει να αφαιρεθούν, και με τη χρήση επανάληψης ελέγχεται εάν ο κάθε όρος στη μεταβλητή η οποία περιέχει τις κριτικές χρηστών, ταυτίζεται η όχι με κάποιο από τα stopwords. Εάν υπάρχει ταύτιση, ο όρος αφαιρείται.

Επιστρέφεται μεταβλητή τύπου γραμματοσειράς η οποία περιέχει τις κριτικές χρηστών, «καθαρές» από stopwords.

4.5.3 Λειτουργία sortReviewsByDate

Μέθοδος υπεύθυνη για τη ταξινόμηση λίστας αντικειμένων Review σε αύξουσα σειρά με βάση το χαρακτηριστικό date(ημερομηνία υποβολής κριτικής στο IMDb).

4.5.4 Λειτουργία deleteDir

Μέθοδος υπεύθυνη για τη διαγραφή των φακέλων οι οποίοι δημιουργούνται κατά τη διαδικασία κατηγοριοποίησης των κριτικών για μία ταινία. Η διαγραφή των φακέλων αυτών είναι απαραίτητη ώστε να μη υπάρχει σύγχυση από πλευράς αλγορίθμου κατηγοριοποίησης κατά την κατηγοριοποίηση διαφορετικών ταινιών.

Η διαγραφή των φακέλων γίνεται με τη χρήση αναδρομής.

4.6 Οπτικοποίηση

Οι κλάσεις οπτικοποίσης επεκτείνουν τη λειτουργικότητα της κλάσης java.swing.JFrame. Σε όλες υλοποιούνται δυο κατασκευαστές(constructors) καλώντας αρχικά τον κατασκευαστή της κλάσης JFrame.

Μέσα από το σώμα κώδικα των constructors καλούνται οι εκάστοτε μέθοδοι διαμόρφωσης του οπτικού αποτελέσματος και κυρίως οι μέθοδοι οι οποίες προσαρμόζουν τα επεξεργασμένα δεδομένα κριτικών ταινιών, τα οποία δέχονται ως παράμερο, στις τελικές οπτικοποιήσεις.

Οι constructors διαφοροποιούνται με το γεγονός ότι ο ένας εκ των δύο, δημιουργεί οπτικοποίηση σε νέο παράθυρο διεπαφής χρήστη.

Κάθε τύπος οπτικοποίησης, υλοποιείται σε ξεχωριστή κλάση.

4.6.1 Κλάση DeviationChart

4.6.1.1 Λειτουργία Mean

Επιστρέφει το μέσω όρο λίστας πραγματικών αριθμών.

4.6.1.2 Λειτουργία StandardDeviation

Επιστρέφει το μέτρο της τυπική απόκλισης από το μέσο, των τιμών λίστας πραγματικών αριθμών. Ο υπολογισμός του μέσου, γίνεται με τη χρήση της μεθόδου Mean.

4.6.1.3 Λειτουργία datediff

Επιστρέφει ακέραιο ο οποίος ισούται με τη διαφορά σε ημέρες μεταξύ δύο ημερομηνιών τύπου java.util.Date. Το αποτέλεσμα επιτυγχάνεται με τον υπολογισμό της σε χιλιοστά του δευτερολέπτου διαφοράς μεταξύ των δυο σημείων στο χρόνο, και η μετατροπή αυτής σε ημέρες.

4.6.1.4 Λειτουργία createDataset

Υπεύθυνη για τη προσαρμογή των επεξεργασμένων δεδομένων κριτικών των χρηστών του IMDb, στην τελική οπτικοποίηση.

Δημιουργούνται αντικείμενα τύπου org.jfree.data.xy.YIntervalSeries[8], τα οποία αντιπροσωπεύουν τις καμπύλες μέσου όρου, στο διάγραμμα διασπορών το οποίο δημιουργείται. Οι καμπύλες αντιπροσωπεύουν αντίστοιχα τα σύνολα των Αρνητικών, Ουδέτερων και Θετικών κρητικών.

Ως αρχικό σημείο στο χρόνο ορίζεται η ημερομηνία της πρώτης χρονικά κτητικής η οποία έχει υποβληθεί, και η οποία αντιστοιχεί στο αντικείμενο που βρίσκεται στη πρώτη θέση της ταξινομημένης λίστας αντικειμένων Review την οποία δέχεται η κλάση ως παράμετρο εισόδου.

Ο άξονας του χρόνου στο διάγραμμα χωρίζεται σε περιόδους. Η διάρκεια περιόδου ορίζεται αρχικά στον ένα μήνα ή τριάντα (30) ημέρες. Ο χρήστης της εφαρμογής έχει τη δυνατότητα μεταβολής της διάρκειας του διαστήματος αυτού μέσω του γραφικού περιβάλλοντος.

Υπολογίζεται η ημερολογιακή διαφορά μεταξύ του πρώτου και τελευταίου αντικειμένου τις λίστας αντικειμένων Review, σε ημέρες και σε συνδυασμό με τη διάρκεια περιόδου η οποία έχει οριστεί, υπολογίζεται το πλήθος των σημείων τα οποία θα αποτελέσουν τις καμπύλες απόψεων.

Δημιουργούνται τρείς (3) λίστες πραγματικών αριθμών, με σκοπό τον υπολογισμό και αποθήκευση των συνόλων Θετικών (Agree +Strongly Argee), Αρνητικών (Disagree + Strongly Disagree) και Ουδέτερων σχολίων, ανά περίοδο, βάσει κατηγοριοποίησης. Τα σύνολα αυτά υπολογίζονται και αποθηκεύονται στις λίστες.

Επαναληπτικά για το σύνολο των σημείων των καμπυλών:

- Δημιουργείται αντικείμενο τύπου org.jfree.data.time.RegularTimePeriod[9], το οποίο αντιπροσωπεύει ένα σημείο στο χρόνο.
- Για κάθε μία από τις τρεις καμπύλες (αντικείμενα τύπου YIntervalSeries), με τη χρήση της μεθόδου StandardDeviation, υπολογίζεται η τυπική απόκλιση των τιμών των παραπάνω λιστών, και σε συνδυασμό με το σύνολο των απόψεων ανά περίοδο, και του αντικειμένου RegularTimePeriod, προστίθεται στις καμπύλες από ένα σημείο.

Δημιουργείται αντικείμενο org.jfree.data.xy.YIntervalSeriesCollection, προσαρμόζονται σε αυτό τα τρία αντικείμενα YIntervalSeries, και το αντικείμενο επιστρέφεται.

4.6.1.5 Λειτουργία createChart

Υπεύθυνη για τη δημιουργία του γραφικού αντικειμένου της οπτικοποίησης. Αποτελείται κυρίως από κλήσεις μεθόδων διαμόρφωσης χρωμάτων και γραφικών. Λαμβάνει τα δεδομένα προσαρμοσμένα στην οπτικοποίηση, ως παράμετρο εισόδου. Επιστρέφει αντικείμενο τύπου org.jfree.chart.JFreeChart[10].

4.6.1.6 Λειτουργία createDemoPanel

Περιέχει κλήση προς τις μεθόδους createDataSet και createChart, επιστρέφοντας αντικείμενο τύπου javax.swing.JPanel, το οποίο είναι και το τελικό αντικείμενο το οποίο θα προσαρμοστεί στο γραφικό περιβάλλον διεπαφής χρήστη, και περιέχει την οπτικοποίηση.

4.6.1.7 Λειτουργία showresults

Μέθοδος η οποία καλεί τον ανάλογο κατασκευαστή της κλάσης, βάσει της επιλογής του χρήστη περί εμφάνισης της οπτικοποίησης στο κεντρικό παράθυρο διεπαφής ή σε ένα νέο παράθυρο.

4.6.2 Κλάση DeviationRatingMean

4.6.2.1 Λειτουργία Mean

Επιστρέφει το μέσω όρο λίστας πραγματικών αριθμών.

4.6.2.2 Λειτουργία StandarDeviation

Επιστρέφει το μέτρο της τυπική απόκλισης από το μέσο, των τιμών λίστας πραγματικών αριθμών. Ο υπολογισμός του μέσου, γίνεται με τη χρήση της μεθόδου Mean.

4.6.2.3 Λειτουργία datediff

Επιστρέφει ακέραιο ο οποίος ισούται με τη διαφορά σε ημέρες μεταξύ δύο ημερομηνιών τύπου java.util.Date. Το αποτέλεσμα επιτυγχάνεται με τον υπολογισμό της σε χιλιοστά του δευτερολέπτου διαφοράς μεταξύ των δυο σημείων στο χρόνο, και η μετατροπή αυτής σε ημέρες.

4.6.2.4 Λειτουργία createDataset

Διαθέτει κοινή μεθοδολογία με τη μέθοδο DeviationChart.createDataset, καθώς είναι υπεύθυνη για τη προσαρμογή των επεξεργασμένων δεδομένων κριτικών των χρηστών του IMDb, στην τελική οπτικοποίηση.

Η διαφορά της παρούσας μεθόδου, είναι ουσιαστικά και η διαφορά των δύο οπτικοποιήσεων. Η παρούσα οπτικοποίηση παρουσιάζει σε καμπύλη το *μέσο όρο αξιολογήσεων* της εκάστοτε ταινίας από τους χρήστες του IMDb (χαρακτηριστικό rating), μέχρι και το τέλος κάθε περιόδου. Σε αντίθεση με την οπτικοποίηση DeviationChart, η οποία παρουσιάζει τρείς (3) καμπύλες με τα σύνολα των γνωμών ανά περίοδο, οι οποίες γνώμες ήταν αποτέλεσμα κατηγοριοποίησης.

Δημιουργείται αντικείμενο org.jfree.data.xy.YIntervalSeriesCollection, προσαρμόζεται σε αυτό το αντικείμενο YIntervalSeries το οποίο αντιπροσωπεύει τη καμπύλη μέσου όρου αξιολόγησης χρηστών, και το αντικείμενο επιστρέφεται.

4.6.2.5 Λειτουργία createChart

Υπεύθυνη για τη δημιουργία του γραφικού αντικειμένου της οπτικοποίησης. Αποτελείται κυρίως από κλήσεις μεθόδων διαμόρφωσης χρωμάτων και γραφικών. Λαμβάνει τα δεδομένα προσαρμοσμένα στην οπτικοποίηση, ως παράμετρο εισόδου. Επιστρέφει αντικείμενο τύπου org.jfree.chart.JFreeChart[10].

4.6.2.6 Λειτουργία createDemoPanel

Περιέχει κλήση προς τις μεθόδους createDataSet και createChart, επιστρέφοντας αντικείμενο τύπου javax.swing.JPanel, το οποίο είναι και το τελικό αντικείμενο το οποίο θα προσαρμοστεί στο γραφικό περιβάλλον διεπαφής χρήστη, και περιέχει την οπτικοποίηση.

4.6.2.7 Λειτουργία showresult

Μέθοδος η οποία καλεί τον ανάλογο κατασκευαστή της κλάσης, βάσει της επιλογής του χρήστη περί εμφάνισης της οπτικοποίησης στο κεντρικό παράθυρο διεπαφής ή σε ένα νέο παράθυρο.

4.6.3 Κλάση StackedBarChart

4.6.3.1 Λειτουργία createDataset

Η μέθοδος δέχεται ως παράμετρο εισόδου τον δυσδιάστατο πίνακα πραγματικών αριθμών ο οποίος δημιουργήθηκε από τη μέθοδο Main.MakeStackedBarChart.
Ανατίθενται οι κατάλληλες τιμές σε ορισμένα περιγραφικά στοιχεία της οπτικοποίησης και επιστρέφεται αντικείμενο τύπου org.jfree.data.general.DatasetUtilities, το οποίο δημιουργείται συναρτήσει των στοιχείων αυτών και του δυσδιάστατου πίνακα.

4.6.3.2 Λειτουργία createChart

Ανατίθενται οι κατάλληλες τιμές σε ορισμένα περιγραφικά στοιχεία της οπτικοποίησης, και τροποποιούνται γραφικά στοιχεία.

Επιστρέφεται αντικείμενο τύπου org.jfree.chart.JFreeChart[10], στο οποίο έχουν προσαρμοστεί τα δεδομένα οπτικοποίησης τα οποία η μέθοδος λαμβάνει ως παράμετρο εισόδου.

4.6.3.3 Λειτουργία showresutls

Μέθοδος η οποία καλεί τον ανάλογο κατασκευαστή της κλάσης, βάσει της επιλογής του χρήστη περί εμφάνισης της οπτικοποίησης στο κεντρικό παράθυρο διεπαφής ή σε ένα νέο παράθυρο.

4.6.4 Κλάση PlotPoint

Κλάση η οποία αναπαριστά ένα σημείο το οποίο προστίθεται σε άξονες ενός διαγράμματος το οποίο το οποίο δημιουργείται μέσω της κλάσης SpiderWebChart. Αποτελείται από χαρακτηριστικά. Την τιμή που φέρει το σημείο, και τη χρονική περίοδο στην οποία ανήκει το σημείο.

Παρατίθεται ο κώδικας της κλάσης:

```
public class PlotPoint {
    public int value;
    public int time;

    public PlotPoint(int value, int time) {
        this.value = value;
        this.time = time;
    }

}
```

4.6.5 Κλάση SpiderWebChart

4.6.5.1 Λειτουργία createChart

Υπεύθυνη για τη δημιουργία του γραφικού αντικειμένου της οπτικοποίησης. Αποτελείτε κυρίως από κλήσεις μεθόδων διαμόρφωσης χρωμάτων και γραφικών. Λαμβάνει τα δεδομένα προσαρμοσμένα στην οπτικοποίηση, ως παράμετρο εισόδου. Επιστρέφει αντικείμενο τύπου org.jfree.chart.JFreeChart[10].

4.6.5.2 Λειτουργία createDemoPanel

Περιέχει κλήση προς τη μέθοδο createChart, επιστρέφοντας αντικείμενο τύπου javax.swing.JPanel, το οποίο είναι και το τελικό αντικείμενο το οποίο θα προσαρμοστεί στο γραφικό περιβάλλον διεπαφής χρήστη, και περιέχει την οπτικοποίηση.

4.6.5.3 Λειτουργία showResults

Μέθοδος η οποία καλεί τον ανάλογο κατασκευαστή της κλάσης, βάσει της επιλογής του χρήστη περί εμφάνισης της οπτικοποίησης στο κεντρικό παράθυρο διεπαφής ή σε ένα νέο παράθυρο.

Στο σώμα της μεθόδου προσαρμόζεται, το σύνολο των δεδομένων οπτικοποίησης στο αντικείμενο SpiderWebChart, το οποίο θα οπτικοποιηθεί. Η προσαρμογή γίνεται με τη χρήση δομής java.util.Set[11] στην οποία ενσωματώνονται τα δεδομένα, τα οποία προσαρμόζονται εκ νέα σε λίστα αντικειμένων τύπου PlotPoint. Τα αντικείμενα PlotPoint αντιπροσωπεύουν τα σημεία τα οποία προστίθενται στους άξονες του διαγράμματος, με τη χρήση επανάληψης.

4.6.6 Κλάση Tag Cloud

4.6.6.1 Λειτουργία createDataset

Υπεύθυνη για τη προσαρμογή των επεξεργασμένων δεδομένων κριτικών των χρηστών του IMDb, στην τελική οπτικοποίηση.

Δέχεται ως παράμετρο εισόδου μεταβλητή τύπου γραμματοσειράς, η οποία περιέχει όλους τους όρους οι οποίοι εμφανίζονται στις κριτικές χρηστών του IMDb για μια συγκεκριμένη ταινία.

Δημιουργείται αντικείμενο τύπου org.mcavallo.opencloud.Cloud[6], στο οποίο προσαρμόζονται οι όροι.

Επαναληπτικά, για κάθε όρο που περιέχεται στο αντικείμενο τύπου Cloud, παράγεται κατάλληλα κώδικας HTML, ώστε κάθε όρος να εμφανίζεται με μέγεθος αναπαράστασης ανάλογο με τη συχνότητα εμφάνισής του.

Ο κώδικας HTML προσαρμόζεται σε αντικείμενο τύπου γραμματοσειράς, και επιστρέφεται από τη μέθοδο.

4.6.6.2 Λειτουργία showresults

Μέθοδος η οποία καλεί τον ανάλογο κατασκευαστή της κλάσης, βάσει της επιλογής του χρήστη περί εμφάνισης της οπτικοποίησης στο κεντρικό παράθυρο διεπαφής ή σε ένα νέο παράθυρο.

4.6.7 Κλάση FrequentWords

4.6.7.1 Λειτουργία createDataset

Υπεύθυνη για τη προσαρμογή των επεξεργασμένων δεδομένων κριτικών των χρηστών του IMDb, στην τελική οπτικοποίηση.

Δέχεται ως παράμετρο εισόδου αντικείμενα τύπου org.mcavallo.opencloud.Cloud[6], τα οποία περιέχουν όλους τους όρους οι οποίοι εμφανίζονται στις κριτικές χρηστών του IMDb για μια συγκεκριμένη ταινία. Το ένα περιέχει τους όρους από τις Θετικές κριτικές, ενώ το άλλο τους όρους που βρίσκονται στο σύνολο των Αρνητικών κριτικών.

Επαναληπτικά, για κάθε όρο που περιέχεται στα αντικείμενα τύπου Cloud, παράγεται κατάλληλα κώδικας HTML, ώστε κάθε όρος να εμφανίζεται με μέγεθος αναπαράστασης ανάλογο με τη συχνότητα εμφάνισής του. Επίσης, το σύνολο των «θετικών» όρων παίρνει διαφορετικό χρώμα, από τους «αρνητικούς» όρους.

Ο κώδικας HTML προσαρμόζεται σε αντικείμενο τύπου γραμματοσειράς, και επιστρέφεται από τη μέθοδο.

4.6.7.2 Λειτουργία showresults

Μέθοδος η οποία καλεί τον ανάλογο κατασκευαστή της κλάσης, βάσει της επιλογής του χρήστη περί εμφάνισης της οπτικοποίησης στο κεντρικό παράθυρο διεπαφής ή σε ένα νέο παράθυρο.

4.7 Γραφικό Περιβάλλον

4.7.1 Κλάση DesktopApplication1

Κλάση υπεύθυνη για την αρχικοποίηση παραμέτρων της κεντρικής οθόνης. Επεκτείνει τη λειτουργικότητα της κλάσης org.jdesktop.application.SingleFrameApplication[12].

4.7.2 Κλάση DesktopApplication1View

Κλάση η οποία αντιπροσωπεύει το αντικείμενο του κεντρική παραθύρου διεπαφής με το χρήστη της εφαρμογής. Επεκτείνει τη λειτουργικότητα της κλάσης org.jdesktop.application.FrameView[13].

Στον κατασκευαστή της κλάσης αρχικοποιούνται ορισμένα γραφικά χαρακτηριστικά του παραθύρου και κατασκευάζεται το γραφικό χαρακτηριστικό μπάρας-ένδειξης αναμονής.

4.7.2.1 Λειτουργία populateList

Αναλαμβάνει της διαδικασία αρχικοποίησης της λίστα τίτλων ταινιών του γραφικού περιβάλλοντος.

Γίνεται ανάγνωση από αρχείο με όνομα “movies.txt”, των τίτλων των 250 δημοφιλέστερων ταινιών του IMDb. Οι τίτλοι καταχωρούνται ως αντικείμενα τύπου γραμματοσειράς στο αντικείμενο της λίστας, τύπου javax.swing.JComboBox[14].

Μέριμνα λαμβάνεται ώστε να υπάρχει αντιστοιχία μεταξύ τίτλων ταινιών και των ID αυτών, καθώς στο χρήστη εμφανίζονται οι τίτλοι ενώ η εφαρμογή χειρίζεται τα δεδομένα ταινιών χρησιμοποιώντας τα ID. Αυτό επιτυγχάνεται μέσω της καταχώρησης των ID, στο αρχείο “movies.txt” με την εξής μορφή:

<ID>@<Movie Title>

4.7.2.2 Λειτουργία showplot

Μέθοδος η οποία καλείται μέσω της ενέργειας χρήστη, πατήματος του κουμπιού “Show Plot” (javax.swing.JButton[15]).

«Δίνει τη τελική εντολή» δημιουργίας της οπτικοποίησης την οποία έχει επιλέξει ο χρήστης της εφαρμογής.

Στο σώμα της μεθόδου εκτελούνται έλεγχοι των παραμέτρων με τις οποίες θα παρουσιαστεί η τελική οπτικοποίηση. Κυρίως τον παραμέτρων τις οποίες μπορεί να τροποποιήσει ο χρήστης μέσω των επιλογών του μενού “Chart Options”,από τη κεντρική οθόνη. Για παράδειγμα εάν ο χρήστης έχει επιλέξει τη δημιουργία οπτικοποίησης τύπου

“DeviationChart”, ελέγχεται εάν πρέπει αυτή να προβληθεί σε νέο παράθυρο ή όχι όπως επίσης και ποιο εύρος χρονικής περιόδου, έχει επιλέξει ο χρήστης για αυτό το διάγραμμα.

Εφόσον γίνουν οι απαραίτητοι έλεγχοι, καλείται η μέθοδος της κλάσης Crawler.Main η οποία αντιστοιχεί στον επιλεγμένο τύπο οπτικοποίησης, ώστε να ξεκινήσει η διαδικασίας δημιουργίας αυτής.

Το σύνολο των λειτουργιών που επιτελούνται στο σώμα της παρούσας μεθόδου, μετατρέπεται σε αντικείμενο τύπου org.jdesktop.application.Task[16], ώστε να είναι δυνατή η παρακολούθηση της προόδου της διεργασίας άρα και η προβολή ένδειξης αναμονής στο χρήστη. Η μετατροπή επιτυγχάνεται με την τροποποίηση της μεθόδου “doInBackGround” ενός αντικειμένου Task.

Παρατίθεται υπόδειγμα κώδικα:

```
Task task = new
Task(org.jdesktop.application.Application.getInstance()) {

            @Override
            protected Void doInBackground() {
                try {

                    // κύριο σώμα κώδικα προς «παρακολούθηση»


                } catch (java.lang.Exception e) {
                    //κώδικας σε περίπτωση λάθους
                }

                return null;
            }
        };
```

4.7.2.3 Λειτουργία abort

Μέθοδος η οποία καλείται μέσω της ενέργειας χρήστη, πατήματος του κουμπιού “Cancel” (javax.swing.JButton[15]).

Διακόπτει/Ακυρώνει τη διαδικασία δημιουργίας οπτικοποίησης μέσω της αφαίρεσης της διαδικασίας εκτέλεσης της (org.jdesktop.application.Task[16]), από το αντικείμενο παρακολουθητή διαδικασιών της εφαρμογής (org.jdesktop.application.TaskMonitor[17]).

Παρατίθεται ο κώδικας της μεθόδου:

```
@Action
    public void abort() {
        taskMonitor.getTasks().remove(0);
        taskMonitor.getForegroundTask().cancel(true);

        this.abort.setVisible(false);
        showplot.setEnabled(true);
    }
```

4.7.2.4 Λειτουργία showUserManual

Καλείται μέσω της ενέργεια χρήστη, επιλογής από το κεντρικό μενού της εφαρμογής

“Help -> UserManual”.

Δημιουργεί ένα νέο αντικείμενο τύπου UserManual.

4.7.2.5 Λειτουργία Addmovie

Καλείται μέσω της ενέργεια χρήστη, επιλογής από το κεντρικό μενού της εφαρμογής

“File -> Add a Movie”.

Δημιουργεί ένα νέο αντικείμενο τύπου InputID.

4.7.3 Κλάση InputID

Κλάση η οποία αντιπροσωπεύει το παράθυρο εισαγωγής δεδομένων νέας ταινίας στην εφαρμογή. Στον κατασκευαστή της κλάσης, όπως και γενικότερα στις κλάσεις οι οποίες επεκτείνουν τη κλάση javax.swing.JFrame[18], αρχικοποιούνται γραφικά χαρακτηριστικά.

4.7.3.1 Λειτουργία isNumeric

Ελέγχει εάν η είσοδος του χρήστη, έχει τη μορφή ακεραίου αριθμού. Εφόσον το ID κάθε ταινίας είναι ακέραιος αριθμός, στις περισσότερες περιπτώσεις επταψήφιος, ο έλεγχος είναι απαραίτητος για την επιτυχή λειτουργία της διαδικασίας εισαγωγής δεδομένων νέας ταινίας.

4.7.3.2 Λειτουργία add

Μέθοδος η οποία καλείται μέσω της ενέργειας χρήστη, πατήματος του κουμπιού “Add” (javax.swing.JButton[15]).

Διενεργούνται οι εξής λειτουργίες ελέγχων:

- Έλεγχος μορφής εισόδου χρήστη με τη χρήση της μεθόδου isNumeric
- Έλεγχος έγκυρου ID ταινίας, σε περίπτωση την οποία ο αριθμός που εισάγει ο χρήστης δεν αντιστοιχεί σε ταινία του IMDb. Το κύριο σώμα της διαδικασίας ελέγχου βρίσκεται στη μέθοδο Main.SerializeAndXML, το αποτέλεσμα της οποίας σε αυτή τη περίπτωση είναι το “404 Error”, σηματοδοτώντας το λάθος.
- Έλεγχος ύπαρξης κριτικών χρηστών για τη ταινία που αντιστοιχεί στο ID το οποίο εισήγαγε ο χρήστης. Το κύριο σώμα της διαδικασίας ελέγχου βρίσκεται στη μέθοδο Main.SerializeAndXML, το αποτέλεσμα της οποίας σε αυτή τη περίπτωση είναι το “no_rev”, σηματοδοτώντας το λάθος.

Εφόσον η διαδικασία εξελιχθεί ιδανικά, καλείται η μέθοδος Main.ClassifySingleMovie, ώστε να επιτελεστεί η διαδικασία κατηγοριοποίησης και να δημιουργηθούν τα αντίστοιχα αρχεία.

Το όνομα της ταινίας προστίθεται στη αντίστοιχη λίστα του γραφικού περιβάλλοντος ,όπως επίσης η εφαρμογή ενημερώνεται γενικότερα(ενημέρωση αρχείων) με τις πληροφορίες της νέας ταινίας.

Εάν η ταινία η οποία επιλέχθηκε, υπάρχει ήδη στη λίστα ταινιών της εφαρμογής, τα υπάρχοντα αρχεία της ταινίας αντικαθίστανται από τα νέα αρχεία. Ουσιαστικά η εφαρμογή ενημερώνεται με τις καινούργιες κριτικές χρηστών, οι οποίες υπεβλήθησαν στο IMDb μέχρι και τη στιγμή εκτέλεσης της λειτουργίας προσθήκης/ενημέρωσης νέας ταινίας.

4.7.4 Κλάση UserManual

Κλάση η οποία αντιπροσωπεύει το παράθυρο προβολής του εγχειριδίου χρήσης της εφαρμογής. Στον κατασκευαστή της κλάσης, όπως και γενικότερα στις κλάσεις οι οποίες επεκτείνουν τη κλάση javax.swing.JFrame[18], αρχικοποιούνται γραφικά χαρακτηριστικά.

Επίσης καλούνται οι κατάλληλες μέθοδοι της βιβλιοθήκης IcePdf, ώστε να δημιουργηθεί επέκταση αντικειμένου javax.swing.JPanel[19] , ικανού να προβάλει έγγραφο μορφής PDF.

4.8 Δομή Γραφικής Διεπαφής Χρήστη

Η εφαρμογή η οποία αναπτύχθηκε στα πλαίσια της παρούσας εργασίας προσφέρει ένα γραφικό περιβάλλον ώστε ο εκάστοτε χρήστης να έχει τη δυνατότητα προβολής και χειρισμού των οπτικοποιήσεων για τα σχόλη των χρηστών του ιστοτόπου IMDb. Το γραφικό περιβάλλον, τα ονόματα και οι περιγραφές των στοιχείων τα οποία το αποτελούν είναι στην Αγγλική γλώσσα. Ακολουθεί αναλυτική περιγραφή της δομής των οθονών τις οποίες περιέχει το περιβάλλον.

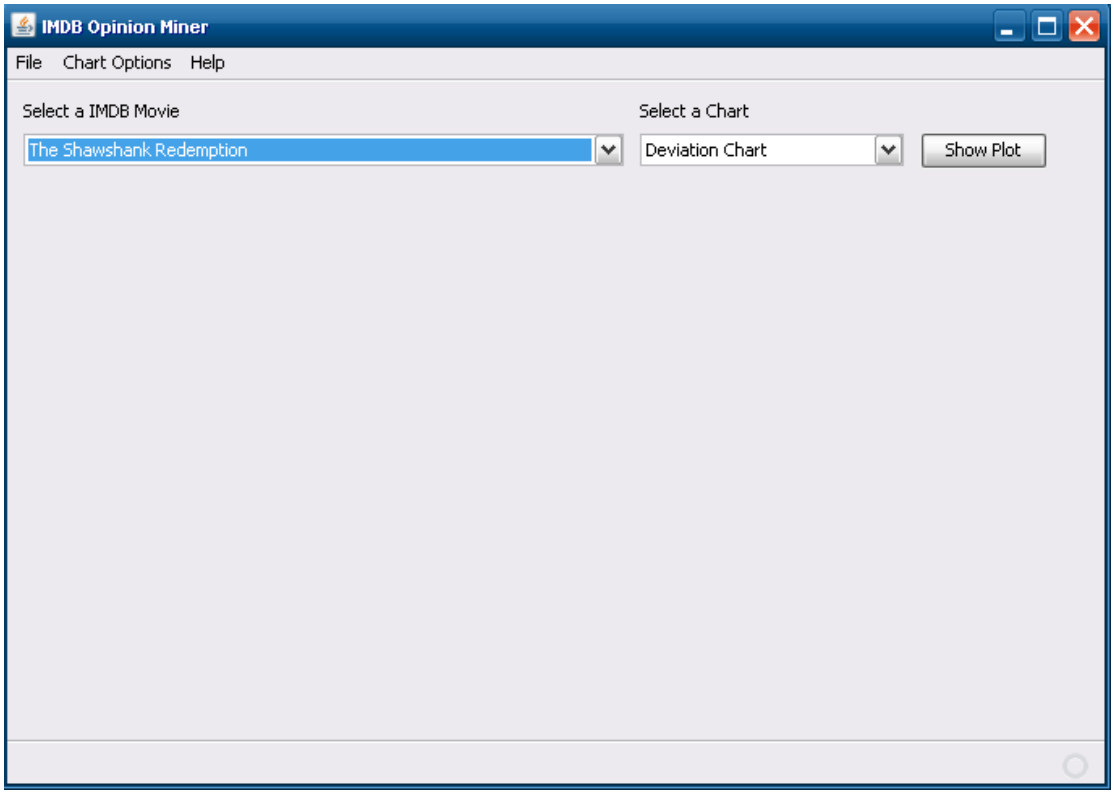
4.8.1 Κεντρική οθόνη

Η κεντρική οθόνη της εφαρμογής περιέχει τα εξής στοιχεία:

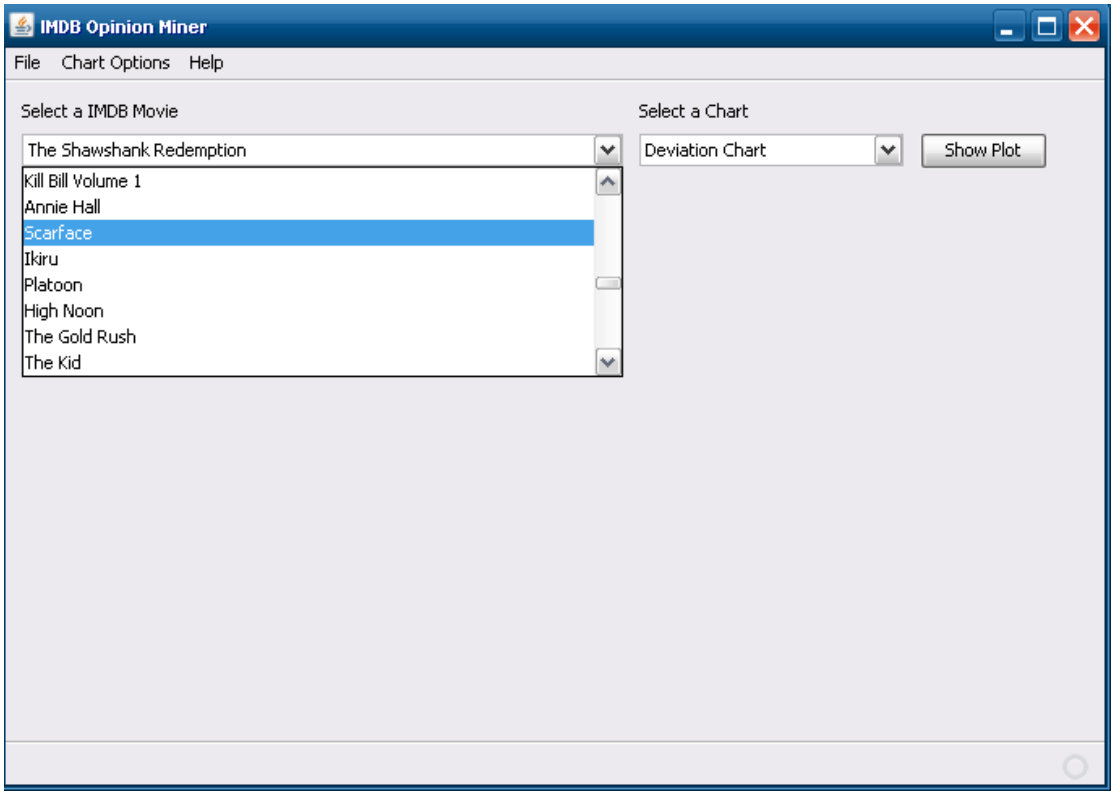
- Το κεντρικό μενού, όπου υπάρχει οι επιλογές File,Chart Options και Help
- Τη λίστα επιλογής ταινίας. Η λίστα περιέχει αρχικά τους τίτλους των 250 δημοφιλέστερων ταινιών του ιστότοπου IMDb.
- Τη λίστα επιλογής τύπου οπτικοποίησης, όπου ο χρήστης επιλέγει το είδος οπτικοποίησης που επιθυμεί να δει, και το οποίο θα αφορά τις κριτικές χρηστών για την ταινία η οποία είναι επιλεγμένη στη λίστα επιλογής ταινίας. Τα διαθέσιμα προς επιλογή είδη οπτικοποίησης είναι:
    1. Deviation Chart
    2. Deviation Rating Mean Chart
    3. Spider Web Chart
    4. Stacked Bar Chart
    5. Tag Cloud Chart
    6. Frequent Words Chart

- Το κουμπί «Show Plot» το οποίο ενεργοποιεί τη διαδικασία δημιουργίας οπτικοποίησης.
- Η οπτικοποίηση προβάλλεται σύμφωνα με τις αρχικές ρυθμίσεις στο αρχικά κενό πλαίσιο το οποίο υπάρχει στη κεντρική οθόνη.

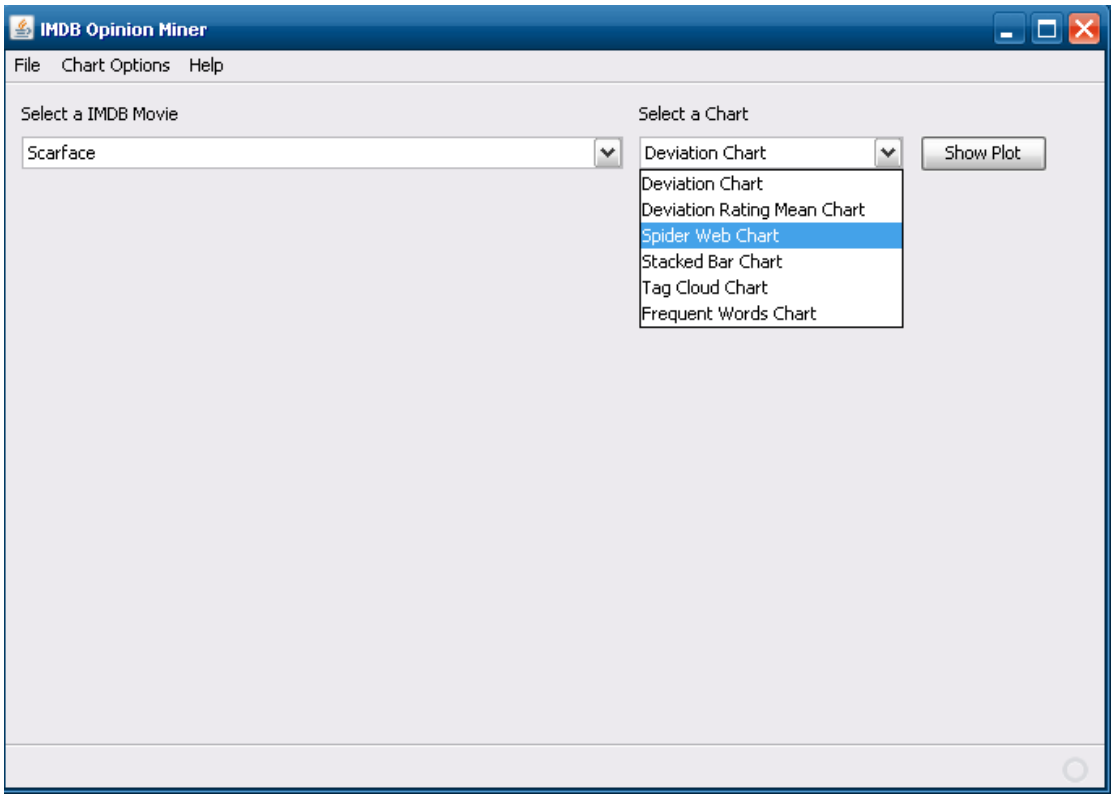
Παρατίθενται εικόνες (screenshots) της κεντρικής οθόνης:



Εικόνα 11. Αρχική Κατάσταση Οθόνης Εφαρμογής



Εικόνα 12. Λίστα Επιλογής Ταινίας



Εικόνα 13. Λίστα επιλογής τύπου οπτικοποίησης

4.8.2 Κεντρικό μενού επιλογών

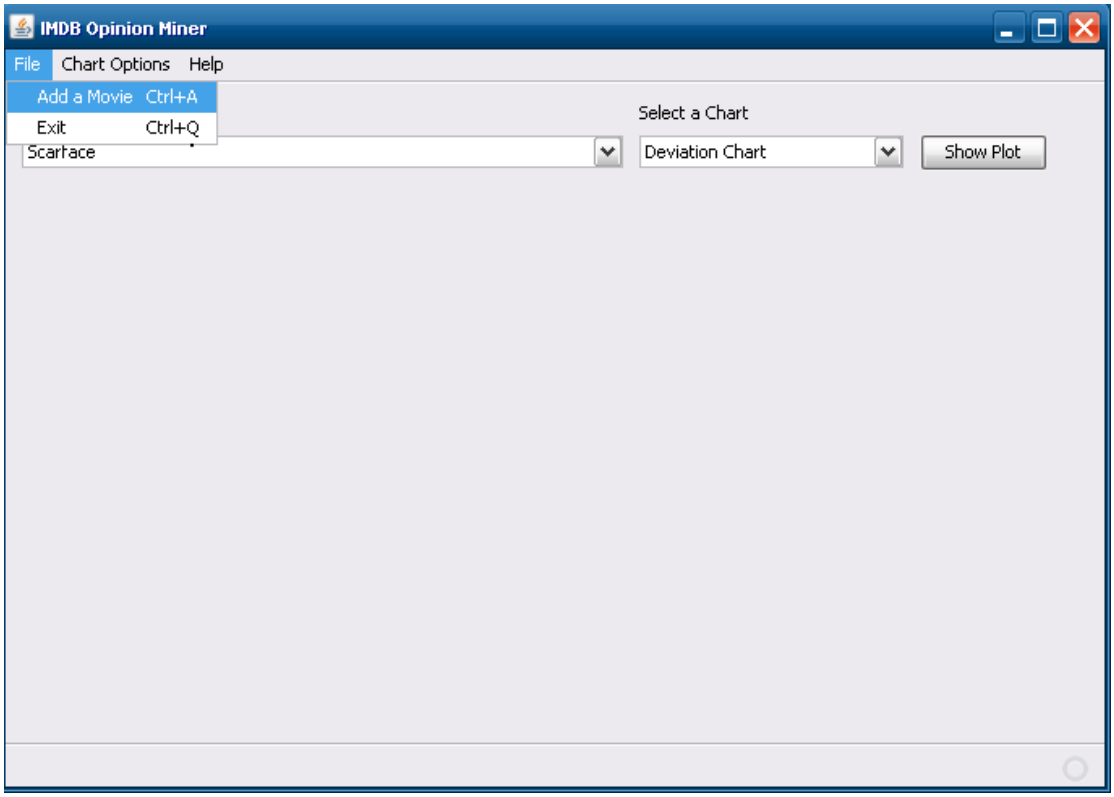
4.8.2.1 Λίστα Επιλογών “File”

Η λίστα επιλογών “File” διαθέτει δυο επιλογές:

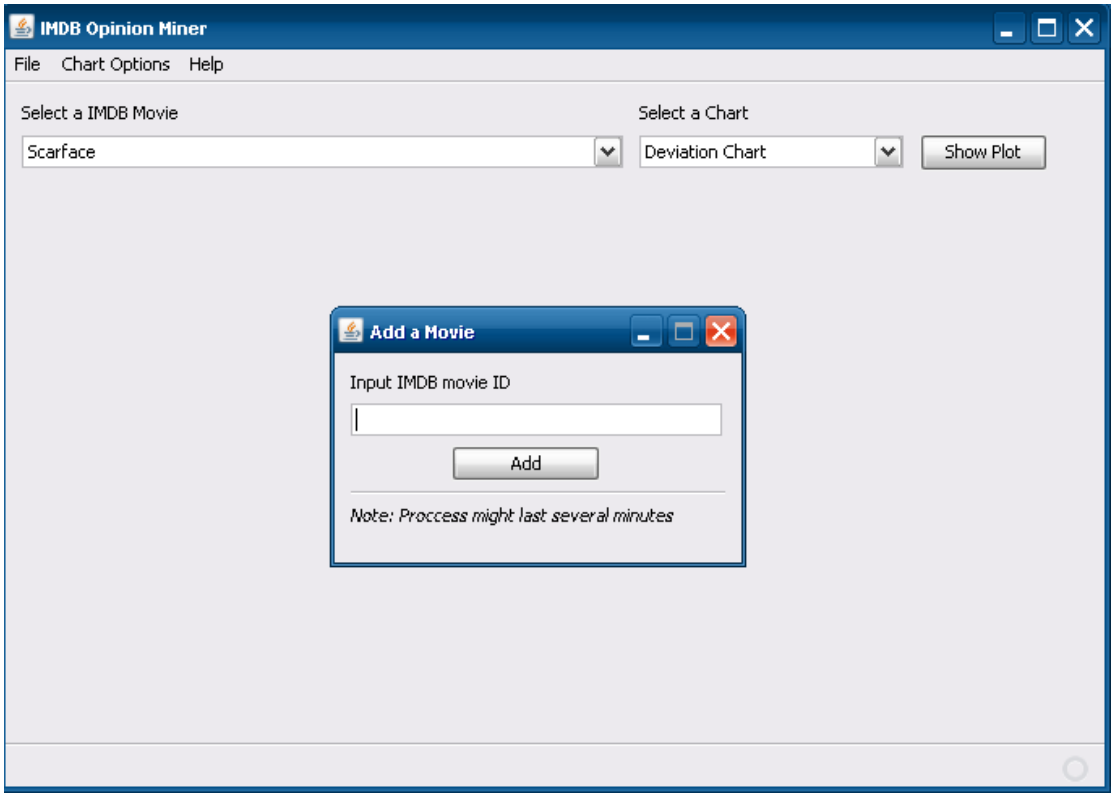
- Επιλογή “Add a Movie”, η οποία εφόσον επιλεγεί προβάλει ένα νέο παράθυρο διεπαφής με το χρήστη. Μέσω της επιλογής αυτής ο χρήστης έχει τη δυνατότητα να προσθέση μια ταινία του ιστοτόπου IMDb στη λίστα επιλογής ταινίας, εφόσον γνωρίζει το μοναδικό αναγνωριστικό αριθμό της ταινίας
  Στο σημείο αυτό πρέπει να διευκρινιστεί ότι κάθε ταινία στο IMDb διαθέτει ένα μοναδικό χαρακτηριστικό αριθμό, ο οποίος περιέχεται και στη διεύθυνση URL της κάθε ταινίας.

- Επιλογή “Exit”, η οποία τερματίζει την εκτέλεση της εφαρμογής.

Παρατίθενται εικόνες(screenshots) της Λίστας Επιλογών “File” :



Εικόνα 14. Λίστα επιλογών “File”



Εικόνα 15. Παράθυρο διεπαφής. Επιλογή "Add a Movie"

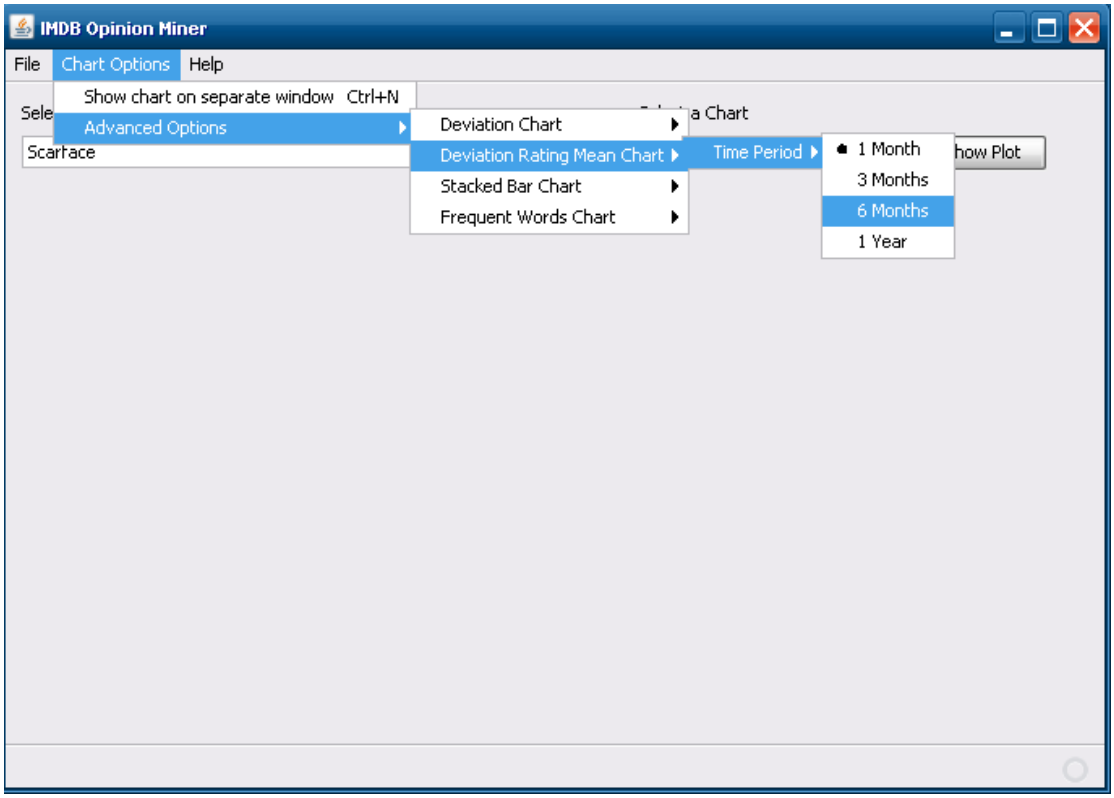
4.8.2.2 Λίστα Επιλογών "Chart Options"

Η λίστα επιλογών "*Chart Options*" διαθέτει δυο επιλογές:

- Επιλογή "Show chart on separate window". Ενεργοποιώντας την επιλογή αυτή ο χρήστης ουσιαστικά επιλέγει την εμφάνιση της οποιασδήποτε οπτικοποίησης σε ένα νέο παράθυρο διεπαφής και όχι στη κεντρική οθόνη. Οι οπτικοποιήσεις θα συνεχίζουν να εμφανίζονται σε ξεχωριστά παράθυρα, για όσο χρονικό διάστημα παραμένει ενεργή η επιλογή "Show chart on separate window". Η εν λόγω επιλογή είναι χρήσιμη για τη διενέργεια συγκρίσεων διαφορετικών οπτικοποιήσεων.

- Επιλογή "Advanced Options", η οποία αποτελεί μενού που περιέχει λίστα υπο-μενού, τα οποία περιέχουν επιλογές προς το χρήστη με τις οποίες μπορεί να τροποποιήσει το τελικό αποτέλεσμα των οπτικοποιήσεων. Ουσιαστικά επιτρέπουν στον χρήστη την παραμετροποίηση του τελικού αποτελέσματος. Κάθε επιλογή παραμετροποίησης παραμένει ενεργή και εφαρμόζεται σε όλες τις οπτικοποιήσης του ανάλογου τύπου, μέχρις ότου απενεργοποιηθεί από το χρήστη.

Παρατίθεται εικόνα(screenshot) της Λίστας Επιλογών “*Chart Options*” :

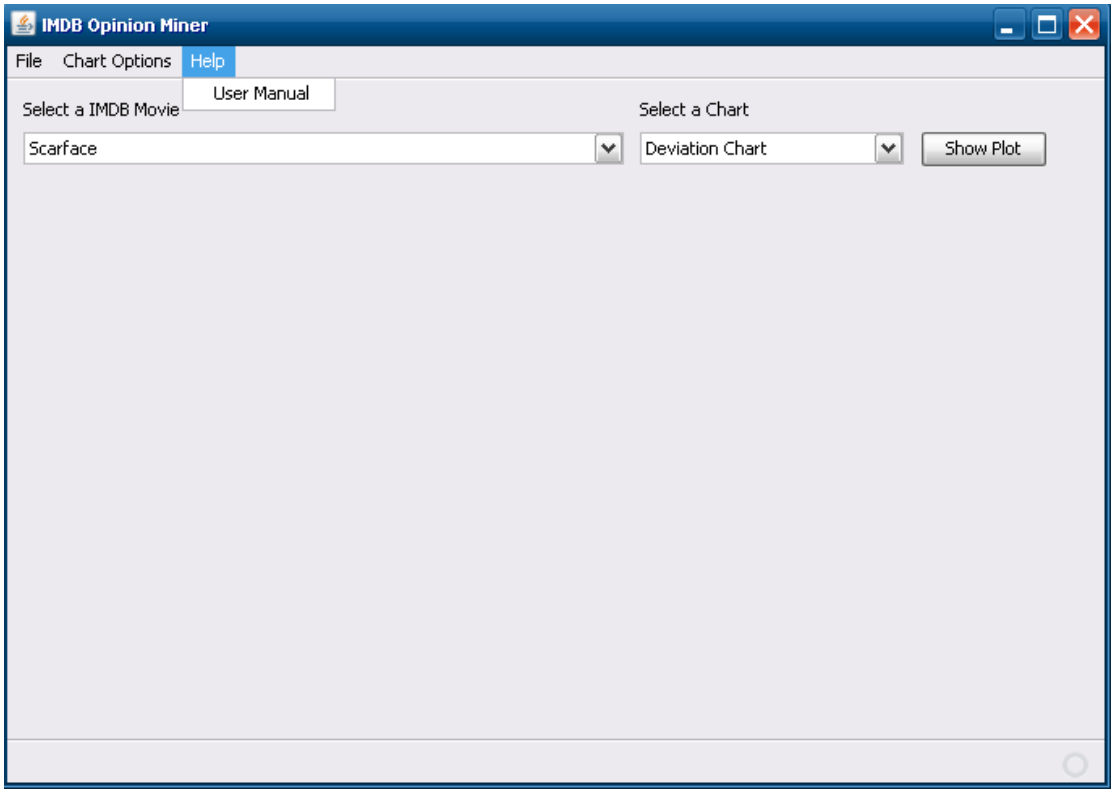


Εικόνα 16. Λίστας Επιλογών “Chart Options”

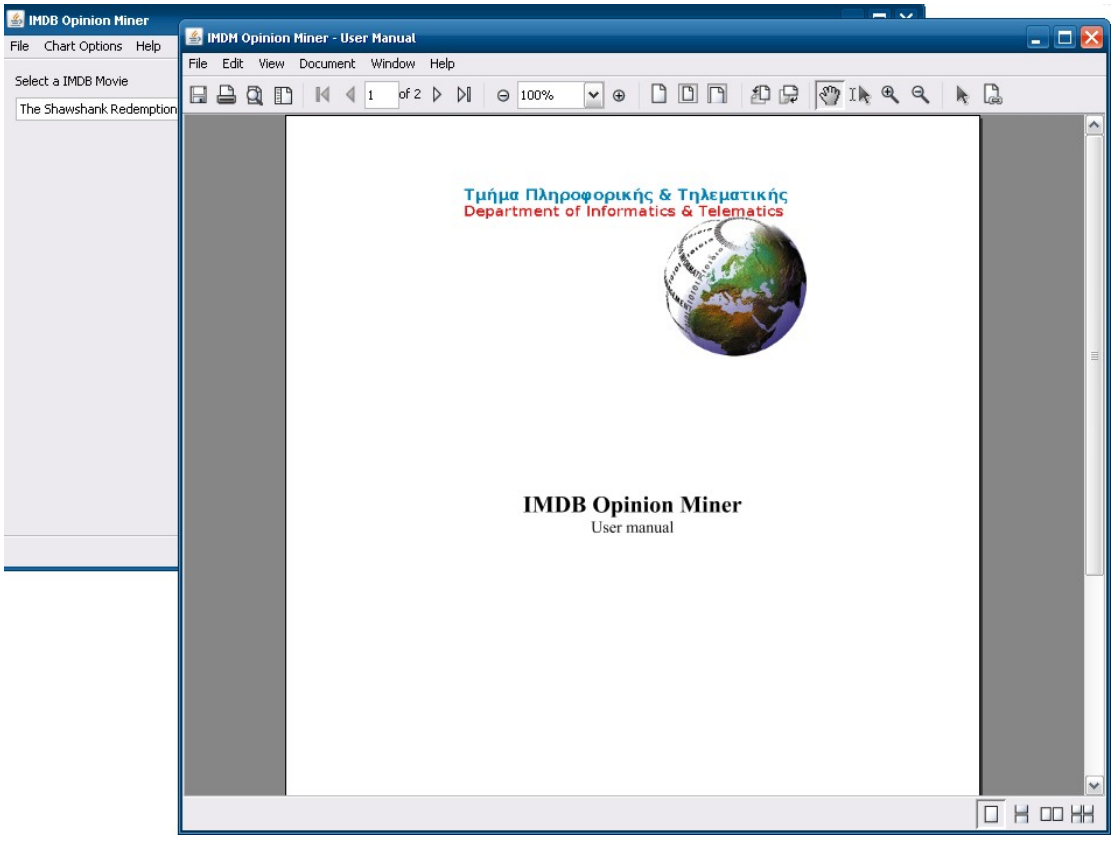
4.8.2.3 Λίστα Επιλογών “Help”

Η λίστα επιλογών “Help” διαθέτει την επιλογή “User Manual”, την οποία μπορεί να επιλέξει ο χρήστης ώστε να εμφανισθεί το Εγχειρίδιο Χρήσης της εφαρμογής, εφόσον χρείαζεται οποιαδήποτε διευκρίνιση για τον τρόπο χειρισμού της εφαρμογής. Το Εγχειρίδιο Χρήσης είναι γραμμένο στη Αγγλική γλώσσα και είναι έγγραφο μορφής PDF.

Στη συνέχεια παρατίθενται εικόνες (screenshots) της Λίστας Επιλογών “*Help*” :

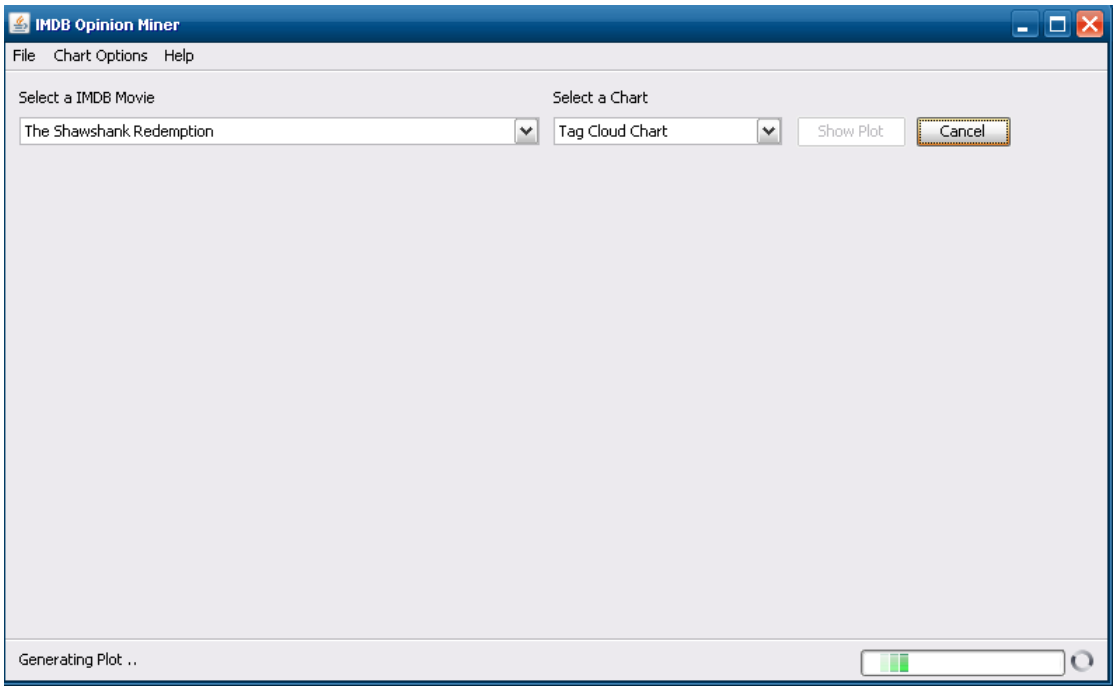


Εικόνα 17. Λίστας Επιλογών "Help"



Εικόνα 18. Εικόνα Εγχειρίδιου Χρήσης Εφαρμογής

Παρατίθεται εικόνα(screenshot) της εφαρμογής κατά τη διαδικασία επεξεργασίας των δεδομένων της ταινίας επιλογής του χρήστη και δημιουργίας του είδους οπτικοποίησης το οποίο έχει επιλεγεί. Εμφανής στο κάτω μέρος της οθόνης είναι και η ένδειξη αναμονής προς το χρήστη (Process Bar) .



Εικόνα 19. Αναμονή δημιουργίας οπτικοποίησης

4.9 Σενάρια χρήσης

4.9.1 Σενάριο 1

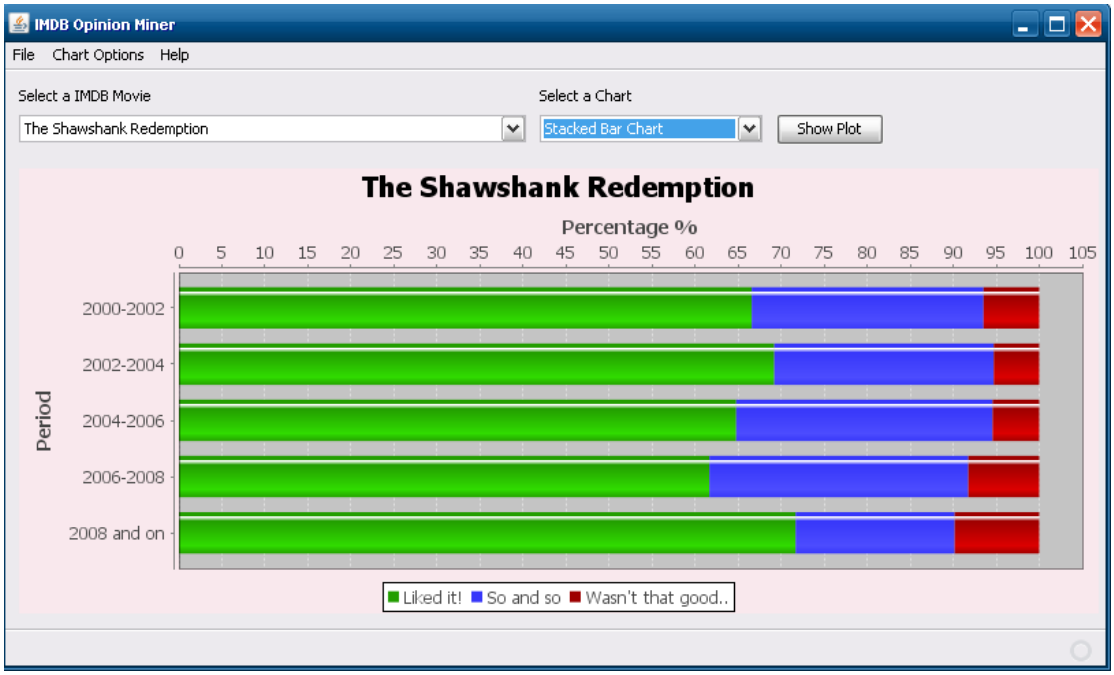
Οπτικοποίηση Stacked Bar Chart. Σύγκριση αποτελέσματος οπτικοποίησης με χρήση δεδομένων κατηγοριοποίησης, με το αποτέλεσμα οπτικοποίησης με χρήση των αρχικών δεδομένων ταινιών IMDb, βασισμένων στις βαθμολογίες αξιολόγησης ταινίας των χρηστών.

Βήμα 1

Ο χρήστης επιθυμεί να δει με τη μορφή Stacked Bar Chart την κατανομή των κριτικών χρηστών, της ταινίας με τίτλο “The Shawshank Redemption”, στο χρόνο.

Επιλέγει αρχικά την ταινία από τη λίστα, επιλέγει τον τύπο γραφήματος .

Αποτέλεσμα οπτικοποίησης με χρήση δεδομένων κατηγοριοποίησης:

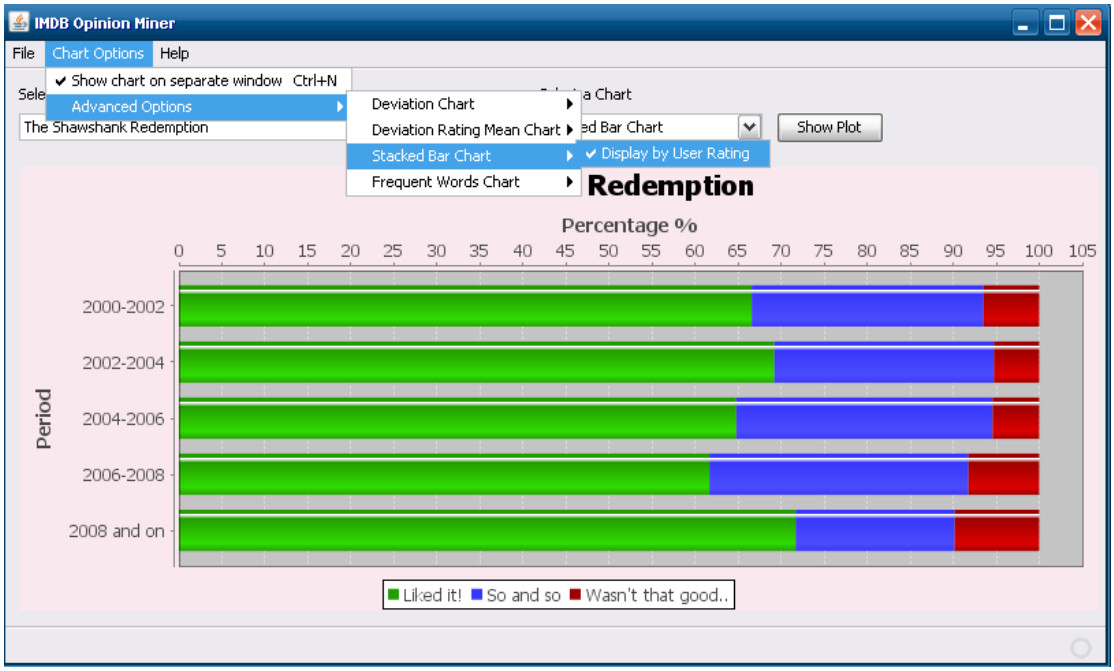


Εικόνα 20. Σενάριο 1, Οπτικοποίηση StackedBarChart 1

Όπως φαίνεται από την εικόνα, η ταινία που είναι η πρώτη στη λίστα με τις 250 καλύτερες του IMDB έχει μεγάλη διαφορά στην αναλογία μεταξύ θετικών (πράσινο χρώμα) και αρνητικών κριτικών (κόκκινο χρώμα). Οπότε η οπτικοποίηση παρουσιάζει μια αντικειμενικά αξιόπιστη εικόνα της πραγματικότητας.

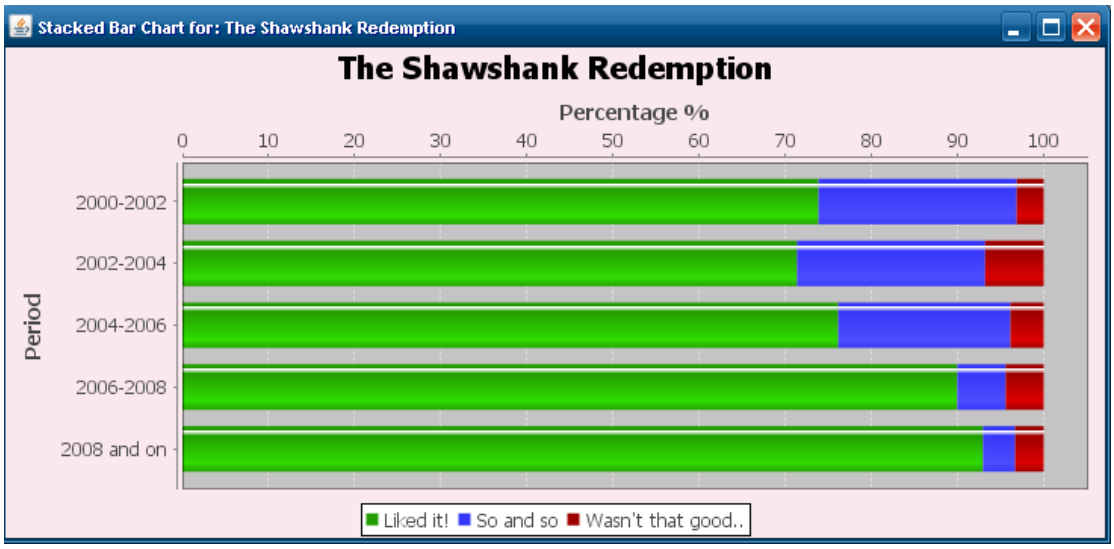
Βήμα 2

Επιλογή παραμέτρου “Show chart on separate window”.
Επιλογή παραμέτρου οπτικοποίησης “Display by User rating”



Εικόνα 21. Σενάριο 1, Οπτικοποίηση StackedBarChart 2

Αποτέλεσμα οπτικοποίησης, βασισμένο στις βαθμολογίες αξιολόγησης ταινίας των χρηστών του IMDb σε νέο παράθυρο:

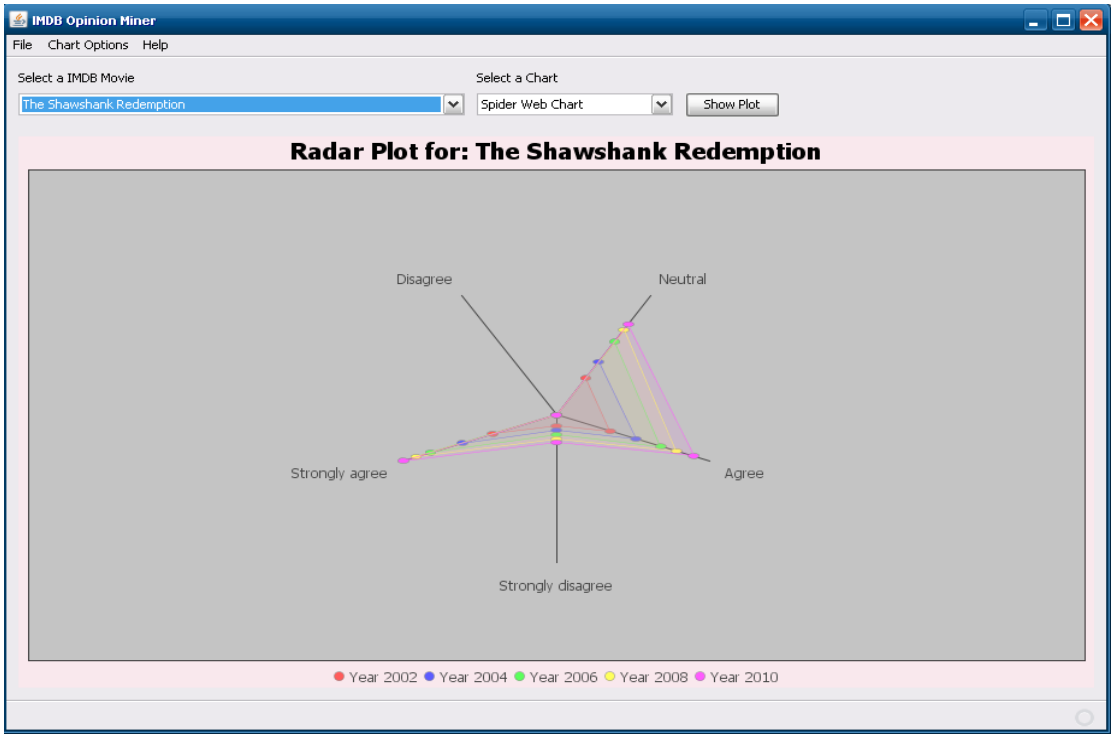


Εικόνα 22. Σενάριο 1, Οπτικοποίηση StackedBarChart 3

Συγκρίνοντας τις εικόνες 20 και 22, γίνεται φανερή μια μικρή απόκλιση μεταξύ των αποτελεσμάτων. Η οπτικοποίηση της εικόνας 21 (κατηγοριοποιημένα δεδομένα) δείχνει τάση του αλγορίθμου να κατηγοριοποιεί μεγαλύτερο αριθμό κριτικών, ως ουδέτερες.

Βήμα 3

Ο χρήστης αποφασίζει να χρησιμοποιήσει τον τύπο οπτικοποίησης Spider Web Chart, ώστε να δει με περισσότερη λεπτομέρεια τα συγκεντρωτικά αποτελέσματα άποψης.
Επιλέγει τον τύπο γραφήματος Spider Web Chart.



Εικόνα 23. Σενάριο 1, Οπτικοποίηση SpiderWebChart

Η οπτικοποίηση δίνει μια αναλυτικότερη εικόνα για τον τύπο συναισθήματος των κριτικών. Οι θετικές κριτικές βρίσκονται πλέον χωρισμένες σε δύο άξονες ανάλογα με την ένταση συναισθήματος των χρηστών(Strongly Agree, Agree). Το ίδιο συμβαίνει και για τις αρνητικές κριτικές. Γίνεται επίσης φανερό ότι το μεγαλύτερο μέρος των αρνητικών κριτικών για τη συγκεκριμένη ταινία είναι άκρως αρνητικές, ενώ υπάρχει σχετική ισορροπία αναλογίας μεταξύ θετικών και πολύ θετικών κριτικών.

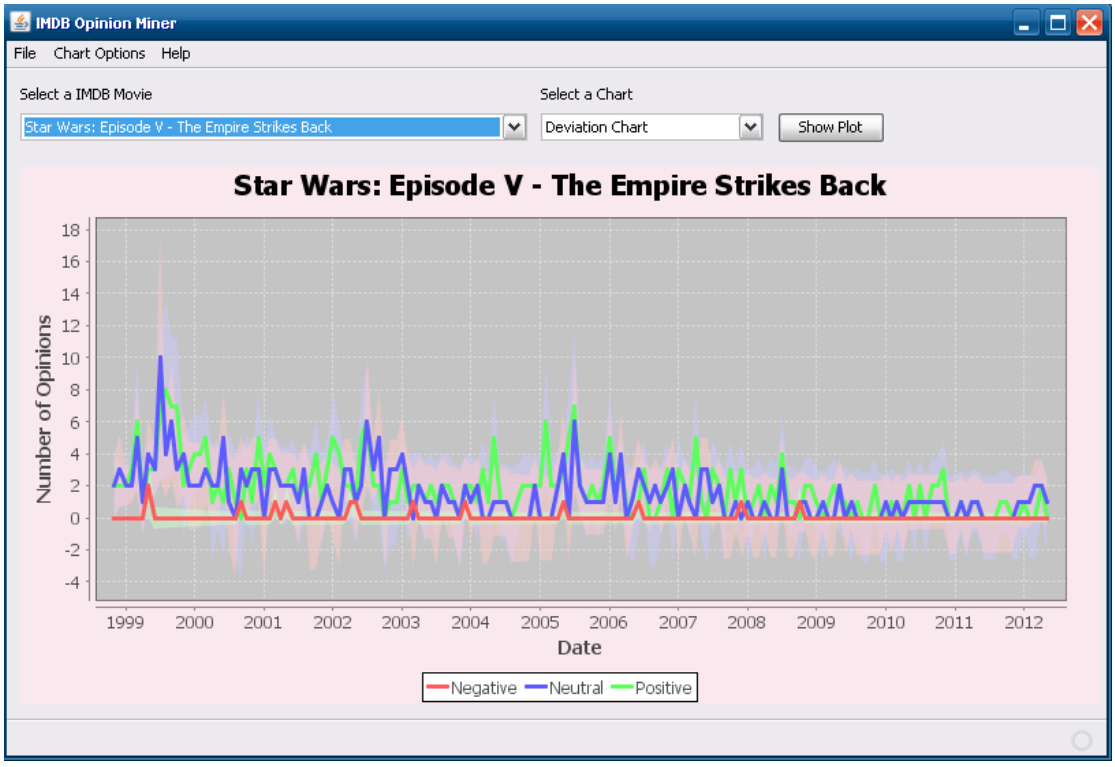
4.9.2 Σενάριο 2

Ο χρήστης επιθυμεί να δει με τη μορφή Deviation Chart την αυξομείωση και τη διασπορά των διαφόρων γνωμών, από τις κριτικές της ταινίας με τίτλο " Star Wars: Episode V – The Empire Strikes Back ", στο χρόνο.

Βήμα 1

Επιλογή ταινίας με τίτλο "Star Wars: Episode V – The Empire Strikes Back", και οπτικοποίηση τύπου Deviation Chart.

Αποτέλεσμα οπτικοποίησης με χρήση δεδομένων κατηγοριοποίησης:

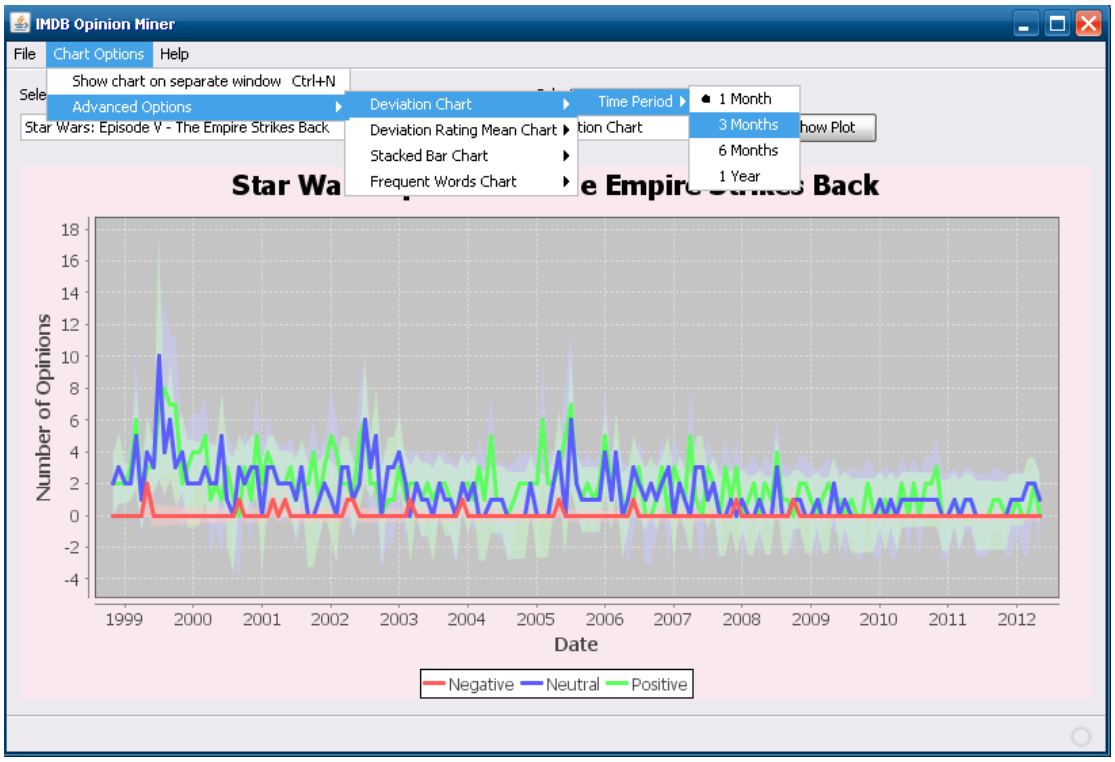


Εικόνα 24. Σενάριο 2, Οπτικοποίηση DeviationChart 1

Από την παραπάνω εικόνα γίνεται φανερό ότι υπήρξαν τρείς χρονικά περίοδοι, κατά τη διάρκεια των οποίων γινόταν μεγάλος αριθμός υποβολών κριτικών στο IMDb για τη συγκεκριμένη ταινία: (1999-2000), (2002-2003) και (2004-2006). Υπάρχει σταθερά μεγάλο πλήθος θετικών σχολίων, σχεδόν πάντοτε μεγαλύτερο από το πλήθος των ουδέτερων και σαφέστατα μεγαλύτερο από το πλήθος των αρνητικών.

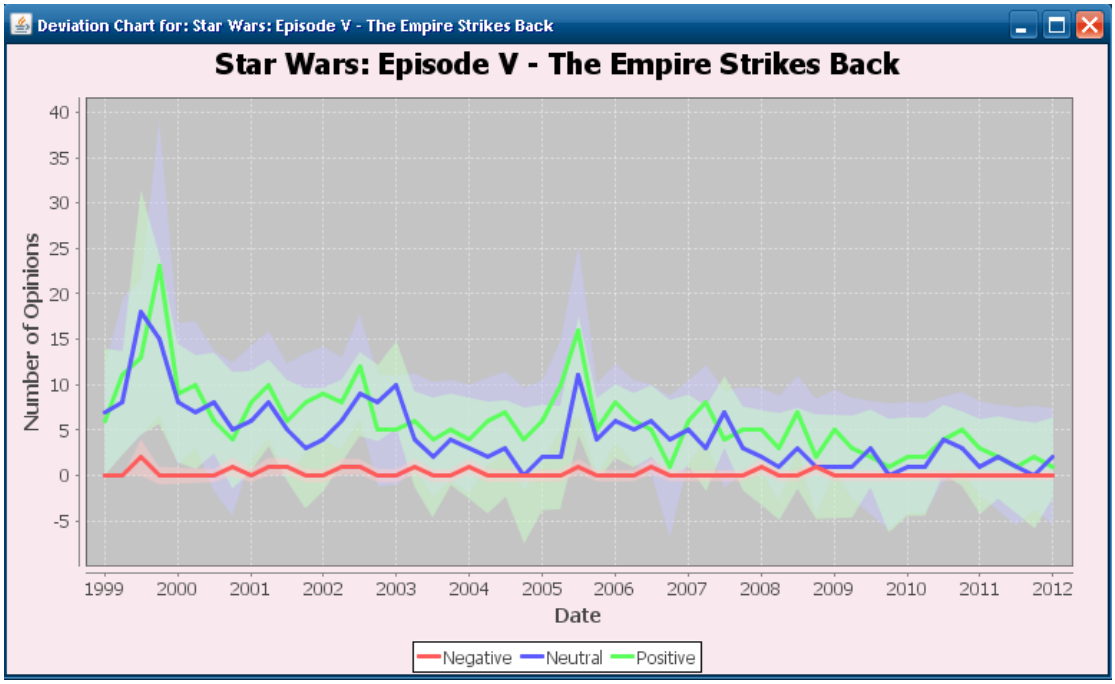
Βήμα 2

Επιλογή παραμέτρου οπτικοποίησης “Time Period -> 3 Months”



Εικόνα 25. Σενάριο 2, Οπτικοποίηση DeviationChart 2

Αποτέλεσμα οπτικοποίησης:



Εικόνα 26. Σενάριο 2, Οπτικοποίηση DeviationChart 3

Η επιλογή αύξησης του εύρους χρονικής περιόδου, διαμορφώνει την οπτικοποίηση ώστε να υπάρχει μια εξομάλυνση των άκρων των καμπυλών(spikes) , καθώς και ελάττωση του αριθμού αυτών, αποτέλεσμα αναμενόμενο. Η γενικότερη συνολική εικόνα της ταινίας στο χρόνο παραμένει ίδια με αυτή που φαίνεται στην εικόνα 23.

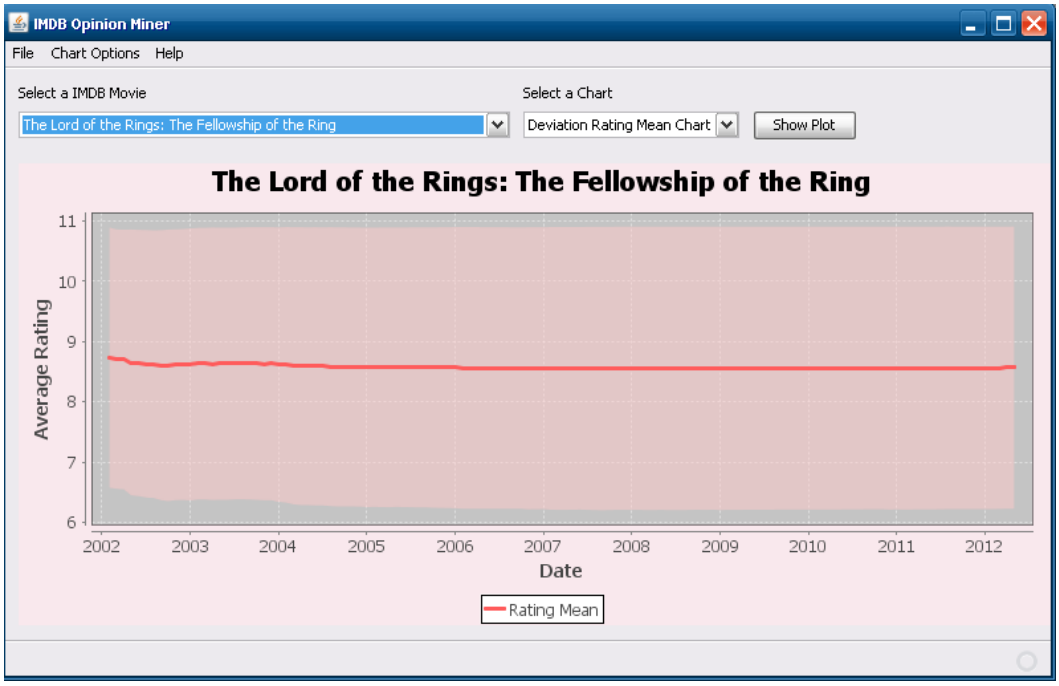
4.9.3 Σενάριο 3

Ο χρήστης επιθυμεί να δει με τη μορφή Deviation Rating Mean Chart την αυξομείωση των βαθμών αξιολόγησης ταινίας από τους χρήστες του IMDb, από τις κριτικές της ταινίας με τίτλο " The Lord of the Rings: The Fellowship of the Ring ", στο χρόνο, και να συγκρίνει τα αποτελέσματα, με τα αντίστοιχα αποτελέσματα για την ταινία με τίτλο "Apocalypse Now".

Βήμα 1

Επιλογή ταινίας με τίτλο "The Lord of the Rings: The Fellowship of the Ring", και οπτικοποίηση τύπου Deviation Chart Mean Chart.

Αποτέλεσμα οπτικοποίησης:

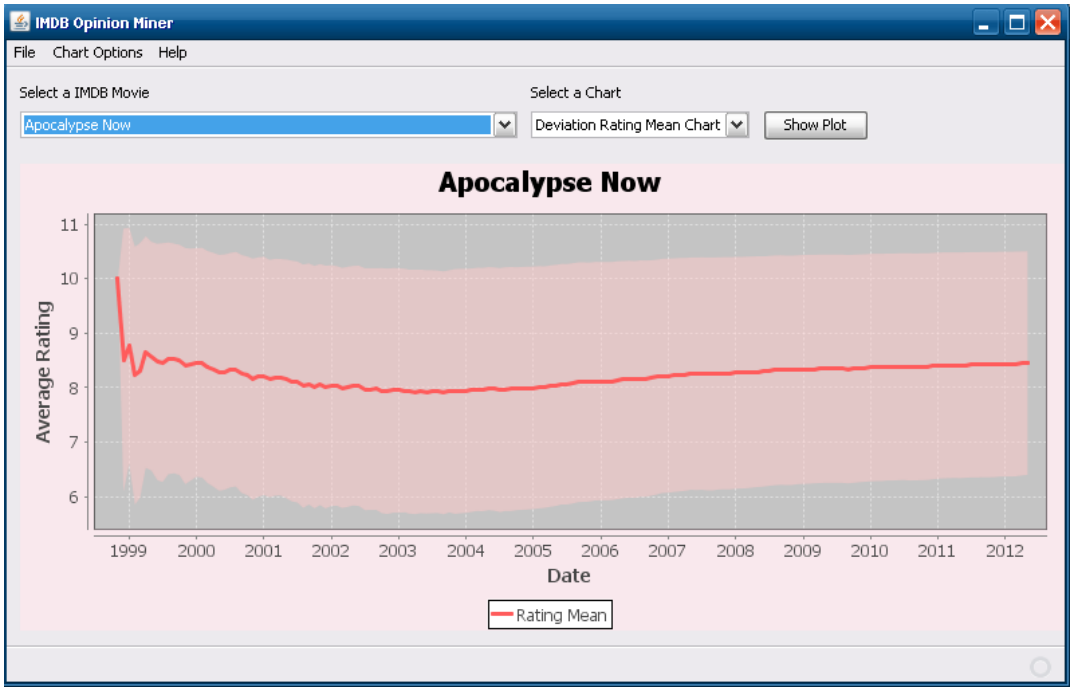


Εικόνα 27. Σενάριο 3, Οπτικοποίηση DeviationChartMean 1

Η ταινία "The Lord of the Rings: The Fellowship of the Ring", φαίνεται να λαμβάνει σταθερά υψηλή μέση βαθμολογία(~8.7/10) από το 2002 μέχρι και το 2010, με απόκλιση (+/-) 2 βαθμών.

Βήμα 2

Επιλογή ταινίας με τίτλο "Apocalypse Now", και οπτικοποίηση τύπου Deviation Chart Mean Chart. Αποτέλεσμα οπτικοποίησης:



Εικόνα 28. Σενάριο 3, Οπτικοποίηση DeviationChartMean 2

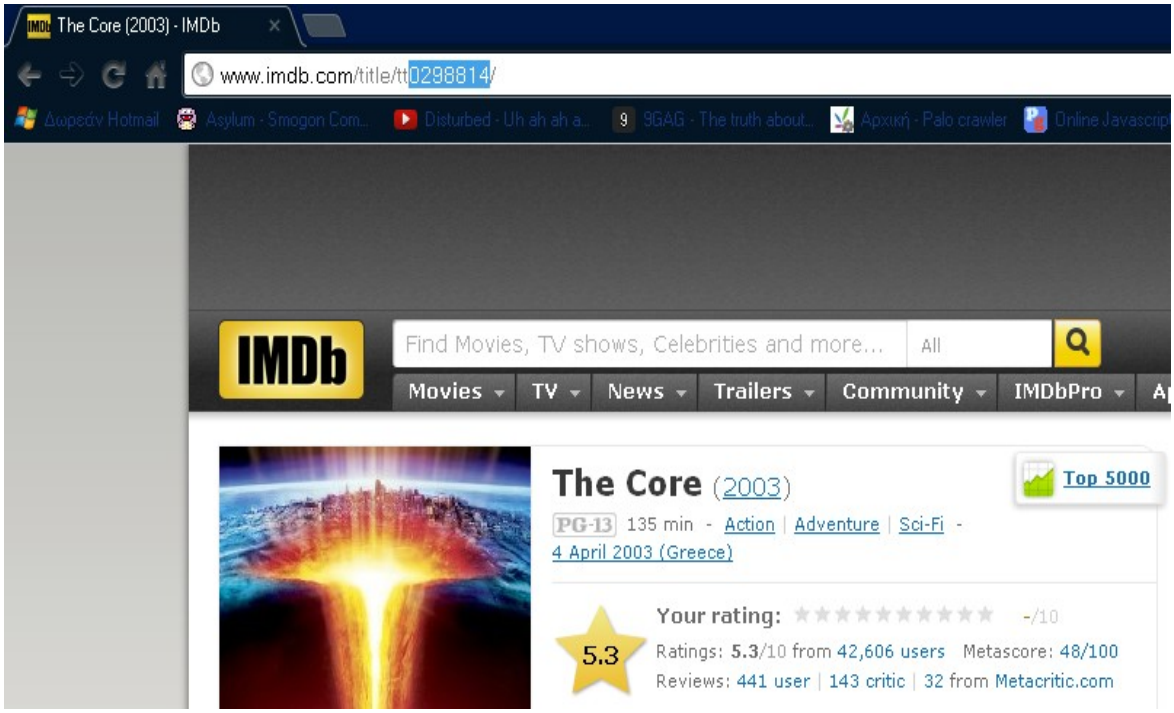
Η ταινία “Apocalypse Now”, φαίνεται να λαμβάνει διαφορετική μέση βαθμολογία στο πέρασμα του χρόνου. Είναι χαρακτηριστικά φανερό ότι αρχικά λάμβανε ιδιαίτερα υψηλές βαθμολογίες, μέχρι και 10/10, στη συνέχεια η μέση βαθμολογίας παρουσιάζει πτώση και από το 2007 και ύστερα παρουσιάζει μικρή άνοδο, με τάση σταθεροποίησης κοντά στη τιμή ~8.4, με σταθερή διασπορά εύρους (+/-) 2.

4.9.4 Σενάριο 4

Ο χρήστης επιθυμεί να εισάγει μία νέα ταινία στη εφαρμογή ώστε να δει τους όρους με τη μεγαλύτερη συχνότητα εμφάνισης στις κριτικές. Επιλέγεται η ταινία “The Core”.

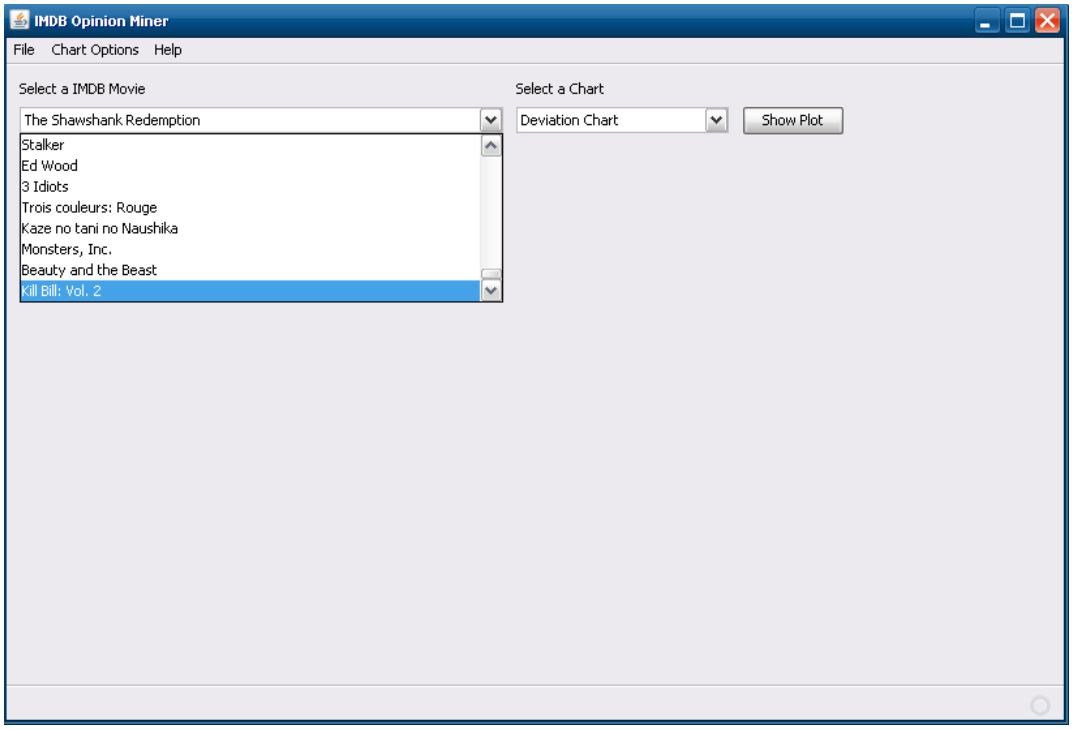
Βήμα 1

Επιλογή ταινίας του IMDb, μέσω web browser.
Καταγραφή ID ταινίας, από τη γραμμή διευθύνσεων.



Εικόνα 29. Σενάριο 4, Πλοήγηση σελίδας ταινίας IMDB

Εικόνα λίστας ταινιών πριν την εκτέλεση της λειτουργίας προσθήκης:

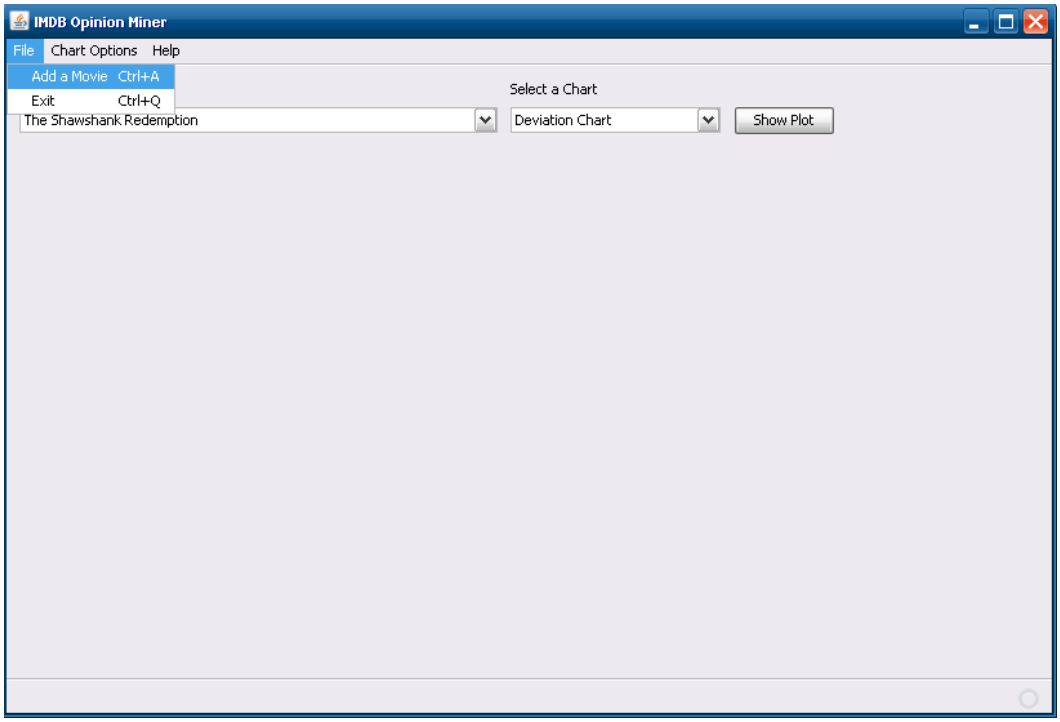


Εικόνα 30. Σενάριο 4, Προσθήκη ταινίας 1

Τελευταία στη λίστα η ταινία με τίτλο “Kill Bill: Vol. 2”

Βήμα 2

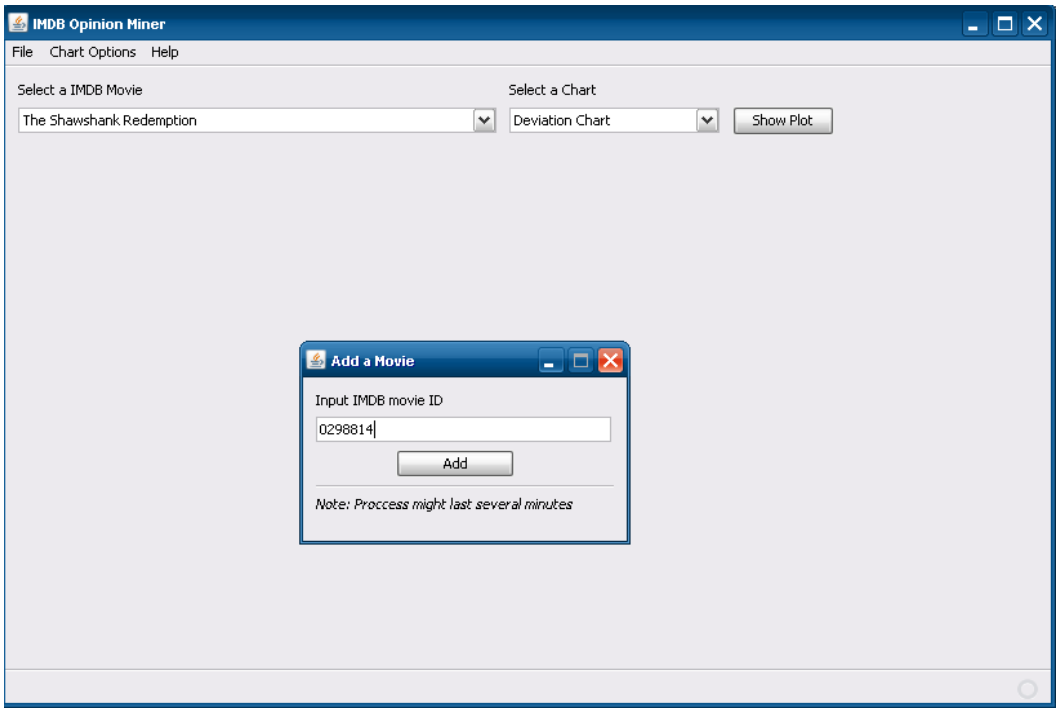
Επιλογή της λειτουργίας προσθήκης από το κεντρικό μενού της εφαρμογής:



Εικόνα 31. Σενάριο 4, Προσθήκη ταινίας 2

Βήμα 3

Εισαγωγή του ID της ταινίας στην εφαρμογή και έναρξη της διαδικασίας:

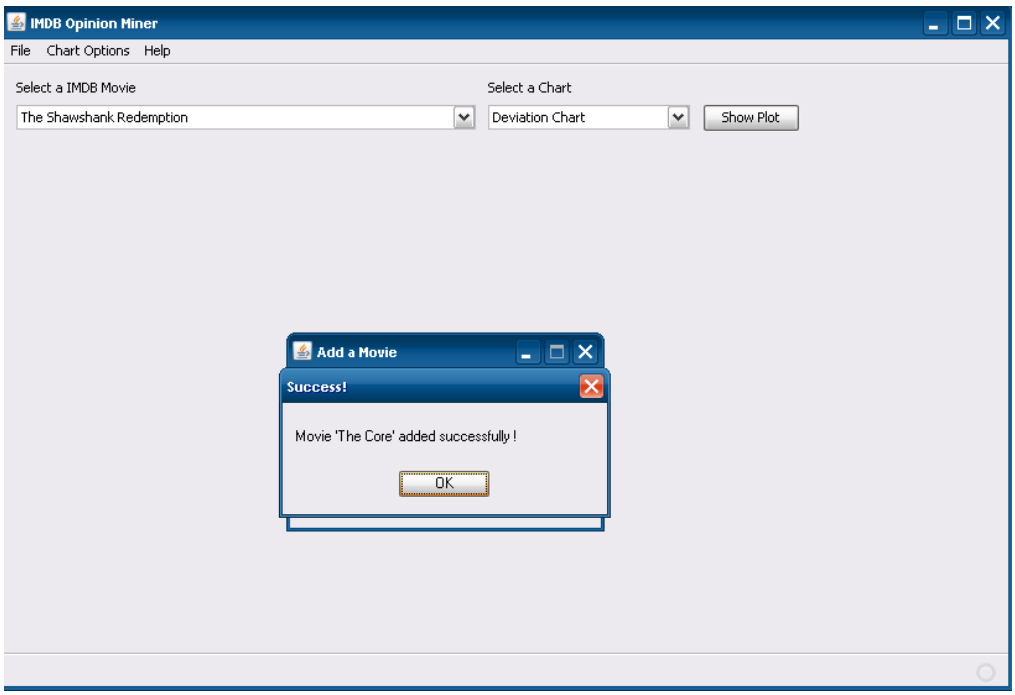


Εικόνα 32. Σενάριο 4, Προσθήκη ταινίας 3

Εφαρμογή δέχεται μόνον αριθμητικές τιμές για είσοδο. Εάν ο χρήστης για παράδειγμα εισάγει κατά λάθος επιπλέον μέρος του URL, εκτός του ID, η εφαρμογή θα επισημάνει το λάθος με κατάλληλο μήνυμα.

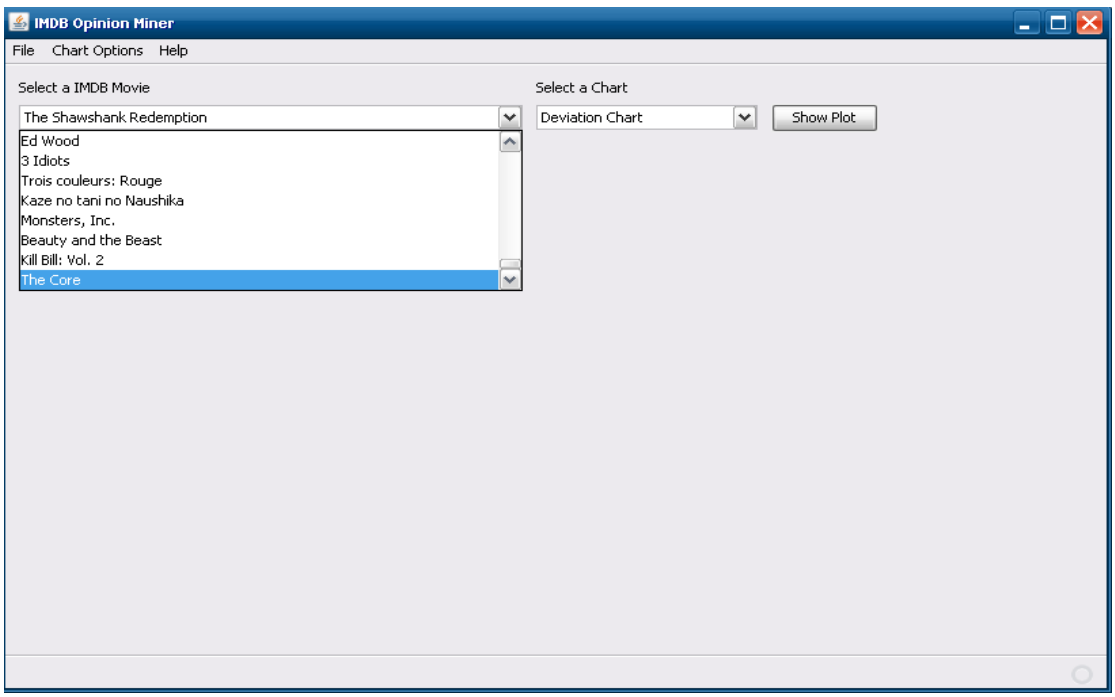
Βήμα 4

Μήνυμα επιτυχημένης εισαγωγής ταινίας:



Εικόνα 33. Σενάριο 4, Προσθήκη ταινίας 4

Εικόνα λίστας ταινιών μετά την ολοκλήρωση εκτέλεσης της λειτουργίας προσθήκης:



Εικόνα 34. Σενάριο 4, Προσθήκη ταινίας 5

Τελευταία στη λίστα βρίσκεται πλέον η ταινία “The Core” που επιλέχθηκε σε αυτό το σενάριο.

Βήμα 5

Επιλογή οπτικοποίησης τύπου Frequent Words και εμφάνιση αυτής, για τα δεδομένα της νέας ταινίας.

Αποτέλεσμα οπτικοποίησης:

Εικόνα 35. Σενάριο 4, Οπτικοποίηση Frequent Words

Όροι με υποκειμενικά θετικό συναίσθημα παρουσιάζονται με μεγαλύτερη συχνότητα στις θετικές, σε σύγκριση με της αρνητικές κριτικές, για παράδειγμα ο όρος ”fun”.

5

Συμπεράσματα

Βάσει της οπτικοποίησης, γίνεται φανερή ορισμένη απόκλιση μεταξύ των αποτελεσμάτων βασισμένων σε κατηγοριοποιημένα δεδομένα και αυτών βασισμένων στο βαθμό αξιολόγησης ταινιών από τους χρήστες του ιστότοπου IMDb (rating). Λόγω του γεγονότος αυτού θα πρέπει μελλοντικά να δοκιμαστούν επιπλέον αλγόρυθμοι κατηγοριοποίησης με στόχο την επίτευξη βέλτιστων αποτελεσμάτων.

Από άποψης βελτίωσης των αποτελεσμάτων των τύπων οπτικοποίησης Tag Cloud και FrequentWords, θα μπορούσε ενδεχομένως να χρησιμοποιηθεί μια τεχνική Lemmatization.

Η καταγραφή, σε αρχείο τύπου PDF, οπτικού υλικού ως μορφή ιστορικού ενεργειών στην εφαρμογή, αλλά και για το σκοπό της συγκεντρωτικής παρουσίασης θεματικών ομάδων οπτικοποιήσεων, φαντάζει ως μια πολύ χρήσιμη μελλοντική βελτίωση.

Ο Web Crawler μπορεί να υλοποιηθεί ως πολυνηματικός, ώστε να συγκεντρώνει τα απαραίτητα δεδομένα με πολλή μεγαλύτερη ταχύτητα. Η βελτίωση αυτή θα αύξανε σημαντικά την ευχρηστία και την αποτελεσματικότητα της εφαρμογής

Στα πλαίσια γενικότερων μελλοντικών βελτιώσεων, η βελτίωση της δομής της γραφικής διεπαφής χρήστη και η προσθήκη επιπλέον τύπων οπτικοποιήσεων, θα συνέβαλαν στην σαφέστατη αύξηση της χρησιμότητας της εφαρμογής και της «ευελιξίας» αυτής, ως προς τη παρουσίαση οπτικοποιήσεων, για δεδομένα μεγαλύτερου εύρους θεματολογιών.

Οι τεχνικές οπτικοποίησης ανέδειξαν χρήσιμα συμπεράσματα για την πορεία των ταινιών στο χρόνο. Η δοκιμή στο IMDb φάνηκε επιτυχής και θα είχε μεγάλο ενδιαφέρον η δοκιμή της των τεχνικών της παρούσας εφαρμογής (πολιτική, ειδήσεις κα).

Βιβλιογραφικές Αναφορές

Friedman, V. (2008) "Data Visualization and Infographics", Smashing Magazine (http://www.smashingmagazine.com)

Liu, B., (2012), "*Sentiment Analysis and Opinion Mining*" , Morgan & Claypool Publishers

Thomas, J. J. & Cook, K. A. (2005), "*Illuminating the Path: The Research and Development Agenda for Visual Analytics*" , National Visualization and Analytics Ctr.

Viıgas,F., Wattenberg, M., van Ham,F., Kriss,J., McKeon, M. (2007), "*Many Eyes: A Site for Visualization at Internet Scale*" IEEE Transactions on Visualization and Computer Graphics 13, 6 (November 2007), 1121-1128.

Annett, M., Kondrak, G, (2008) *"A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs"* In Proceedings of the Canadian Society for computational studies of intelligence, 21st conference on Advances in artificial intelligence (Canadian AI'08), Sabine Bergler (Ed.). Springer-Verlag, Berlin, Heidelberg, 25-35.

Chen, C.; Ibekwe-Sanjuan, F.; SanJuan, E. & Weaver, C. (2006), "*Visual Analysis of Conflicting Opinions*"., *in* Pak Chung Wong & Daniel A. Keim, ed., 'IEEE VAST' , IEEE, , pp. 59-66 .

Radovanović,M., Ivanović,M. (2008) "*Text Mining Approaches and Applications*", Novi Sad J. Math. Vol. 38, No. 3, 2008, 227-234

D. Maynard and K. Bontcheva and D. Rout. "*Challenges in developing opinion mining tools for social media*". In Proceedings of @NLP can u tag #user*generated*content?! Workshop at LREC 2012, May 2012, Istanbul, Turkey.

Oelke, D.; Hao, M. C.; Rohrdantz, C.; Keim, D. A.; Dayal, U.; Haug, L.-E. & Janetzko, H. (2009), "*Visual opinion analysis of customer feedback data*"., *in* 'IEEE VAST' , IEEE, , pp. 187-194 .

Potthast, M. & Becker, S. (2010), "*Opinion Summarization of Web Comments*"., *in* Cathal Gurrin; Yulan He; Gabriella Kazai; Udo Kruschwitz; Suzanne Little; Thomas Roelleke; Stefan M. Rüger & Keith van Rijsbergen, ed., 'ECIR' , Springer, , pp. 668-669.

Χρήσιμοι Σύνδεσμοι

1. http://www.imdb.com/
2. http://www.jfree.org/jfreechart/
3. http://opencloud.mcavallo.org/
4. http://www.cs.waikato.ac.nz/ml/weka/
5. http://icesoft.org

Παραπομπές στο Διαδίκτυο

1. http://docs.oracle.com/javase/1.4.2/docs/api/java/io/Serializable.html

2. http://docs.oracle.com/javase/1.4.2/docs/api/java/util/Date.html

3. http://docs.oracle.com/javase/1.4.2/docs/api/java/lang/Exception.html

4. http://docs.oracle.com/javase/1.4.2/docs/api/java/net/URLConnection.html

5. http://docs.oracle.com/javase/1.4.2/docs/api/java/util/HashMap.html

6. http://opencloud.mcavallo.org/documentation/api/org/mcavallo/opencloud/Cloud.html

7. http://docs.oracle.com/javase/1.4.2/docs/api/java/lang/String.html#split(java.lang.String, int)

8. http://www.jfree.org/jfreechart/api/javadoc/org/jfree/data/xy/YIntervalSeries.html

9. http://www.jfree.org/jfreechart/api/javadoc/org/jfree/data/time/RegularTimePeriod.html

10. http://www.jfree.org/jfreechart/api/javadoc/org/jfree/chart/JFreeChart.html

11.http://docs.oracle.com/javase/1.4.2/docs/api/java/util/Set.html

12. http://manual.openestate.org/extern/appframework-1.0.3/org/jdesktop/application/SingleFrameApplication.html

13. http://manual.openestate.org/extern/appframework-1.0.3/org/jdesktop/application/FrameView.html

14. http://docs.oracle.com/javase/1.5.0/docs/api/javax/swing/JComboBox.html

15. http://docs.oracle.com/javase/1.4.2/docs/api/javax/swing/JButton.html

16. http://manual.openestate.org/extern/appframework-1.0.3/org/jdesktop/application/Task.html

17. http://www.jarvana.com/jarvana/view/net/java/dev/appframework/appframework/1.03/appframework-1.03-javadoc.jar!/org/jdesktop/application/TaskMonitor.html

18. http://docs.oracle.com/javase/1.4.2/docs/api/javax/swing/JFrame.html

19. http://docs.oracle.com/javase/1.4.2/docs/api/javax/swing/JPanel.html