ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εντοπισμός και οντολογική περιγραφή πλάνων κινηματογραφικών ταινιών με ειδικό περιεχόμενο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΕΥΑΓΓΕΛΙΑΣ ΘΑΝΟΥ

Επιβλέπων : Γιώργος Στάμου

Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΈΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εντοπισμός και οντολογική περιγραφή πλάνων κινηματογραφικών ταινιών με ειδικό περιεχόμενο

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΕΥΑΓΓΕΛΙΑΣ ΘΑΝΟΥ

Επιβλέπων : Γιώργος Στάμου
Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Ιουλίου 2016.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Γιώργος Στάμου	Στέφανος Κόλλιας	Ανδρέας-Γεώργιος Σταφυλοπάτης
Επίκουρος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2016

(Υπογραφή)

.....

ΕΥΑΓΓΕΛΙΑ ΘΑΝΟΥ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Γιώργο Στάμου για την ευκαιρία που μου έδωσε να ασχοληθώ με αυτή τη διπλωματική εργασία. Επίσης, τον κ.Αλέξανδρο Χορταρά για τη βοήθεια και τις συμβουλές του. Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για την υπομονή και την υποστήριξη τους.

Περίληψη

Ο σκοπός της διπλωματικής εργασίας ήταν η ανάπτυξη μίας εφαρμογής που θα συνδέει τις αδόμητες πληροφορίες που παρέχονται από τους χρήστες τη ιστοσελίδας IMDB ως "συμβουλές προς γονείς" με τις ίδιες τις "προβληματικές" σκηνές οι οποίες περιγράφονται, μέσω του σεναρίου της ταινίας.

Για να επιτευχθεί αυτή η διασύνδεση επεκτάθηκε η ήδη υπάρχουσα οντολογία αναπαράστασης σεναρίων ώστε να γίνει μοντελοποίηση για τις "προβληματικές" σκηνές. Με βάση την νέα αυτή οντολογία κατηγοριοποιήθηκαν αφενός οι σκηνές του σεναρίου και αφετέρου οι "προβληματικές" σκηνές που περιγράφονται στον οδηγό γονέων σε κλάσεις ανάλογα με το περιεχόμενο τους. Για να επιτευχθεί αυτό, έγινε επεξεργασία φυσικής γλώσσας και στην συνέχεια, εντοπίζοντας λέξεις-κλειδιά, δημιουργήθηκαν χαρακτηριστικά διανύσματα για τις σκηνές του σεναρίου και τις περιγραφές. Αντίστοιχα χαρακτηριστικά διανύσματα δημιουργήθηκαν για τις διάφορες κατηγορίες "προβληματικών" σκηνών με βάση τις λέξεις κλειδιά που αντιστοιχούν στην κάθε μία από αυτές. Τα χαρακτηριστικά αυτά διανύσματα συγκρίθηκαν με διάφορα μέτρα ομοιότητας, καταλήγοντας τελικά στην απόσταση Ochiai και οι σκηνές και οι περιγραφές κατηγοριοποιήθηκαν αναλόγως. Η τελική διασύνδεση έγινε με χρήση της μεθόδου tf-idf μεταξύ των περιγραφών και των σκηνών του σεναρίου που ανήκουν στην ίδια κλάση.

Χρησιμοποιώντας το σύστημα ο χρήστης μπορεί να δημιουργήσει μια τροποποιημένη σελίδα του οδηγού γονέων του IMDB που του επιτρέπει για κάθε περιγραφή να δει τα βίντεο των σκηνών που σε αυτά η περιγραφή αυτή αναφέρεται, καθώς και τα βίντεο των σκηνών που ανήκουν σε κάποια συγκεκριμένη κατηγορία που μπορεί να τον ενδιαφέρει. Τέλος, ο χρήστης έχει την δυνατότητα να δει και τις τροποποιημένες σελίδες άλλων ταινιών, εάν αυτές έχουν ήδη δημιουργηθεί.

Λέξεις Κλειδιά: <<Οντολογία, αναπαράσταση γνώσης, επεξεργασία φυσικής γλώσσας, μέτρα ομοιότητας, εξόρυξη πληροφορίας, RDF>>

Abstract

The goal of this diploma thesis was the development of an application that will connect the unstructured data provided by the users of the website IMDB as “parental guide” with the actual “problematic” scenes described by them, through the script of the movie.

In order to achieve that connection, the already existing ontology got expanded so that there could be modeling for the “problematic” scenes. Based on this new ontology, both the script scenes and the “problematic” scenes described at the parental guide got categorised into classes according to their content. To achieve this, there was natural language processing and thereafter, identifying keywords, characteristic vectors were created for the scenes of the scenario and the descriptions. Corresponding characteristics vectors were created for different categories of "problematic" scenes based on keywords that match each one of them. These characteristics vectors were compared with various similarity measures, ultimately choosing Ochiai distance, which brought about the best results, and the scenes and descriptions were categorized accordingly. The final connection was made using the method tf-idf between the descriptions of the parental guide and the script scenes belonging to the same class.

Using the system, the user can create a modified web page of the IMDB parental guide that will provide him for each description with the videos of the scenes that are referenced in that description, as well as the video of the scenes for each specific category that might interest him. Finally, the user has the option to visit the modified web pages of other movies, provided they have been created beforehand.

Keywords: <<ontology, knowledge representation, natural language processing, similarity measures, information retrieval, RDF>>

Πίνακας περιεχομένων

1

Εισαγωγή

Διάφορες βάσεις δεδομένων του διαδικτύου, όπως η ιστοσελίδα IMDB[1], αποτελούν ιδιαίτερα πλούσιες πηγές πληροφοριών όσον αφορά τις κινηματογραφικές ταινίες. Μεταξύ των πληροφοριών που παρέχονται για τις ταινίες, όπως τίτλος, συντελεστές, τεχνικά χαρακτηριστικά, είδος, κ.α., μία λειτουργία που προσφέρει στους χρήστες είναι ο οδηγός γονέων.

Ο οδηγός γονέων του IMDB έχει ως στόχο να προσφέρει επιπλέον πληροφορίες στους γονείς σχετικά με τις επιλήψιμες σκηνές της ταινίας, ώστε να γνωρίζουν εάν υπάρχει κάτι το οποίο θα προτιμούσαν να μην δουν τα παιδιά τους. Ο οδηγός αυτός αποτελείται από συνεισφορές των χρηστών του site που έχουν δει την ταινία και που προσθέτουν γενικές περιγραφές των "προβληματικών" σκηνών για διάφορες κατηγορίες.

Μία άλλη πηγή πληροφοριών για το περιεχόμενο των σκηνών μίας ταινίας σε έγγραφη μορφή είναι το post-production σενάριο. Το σενάριο αυτό, σε αντίθεση με το αρχικό σενάριο που συγγράφει ο σεναριογράφος και με βάση αυτό γυρίζεται η εκάστοτε ταινία, δημιουργείται μετά από την ολοκλήρωση των γυρισμάτων της ταινίας. Οπότε, περιέχει πολύ περισσότερες πληροφορίες, όπως αναλυτικές

περιγραφές της τοποθεσίας και των γεγονότων που λαβαίνουν χώρα σε κάθε σκηνή, το ακριβές κείμενο των διαλόγων των χαρακτήρων. Επίσης, περιλαμβάνει και πιο τεχνικές λεπτομέρειες όπως το είδος των πλάνων και την χρονική στιγμή παρουσίας τους μέσα στην ταινία. Αποτελεί επί της ουσίας μια αναπαράσταση της ταινίας σε κείμενο γραπτού λόγου.

Και για τις δύο περιπτώσεις, οι πηγές που έχουμε είναι αδόμητο κείμενο, χωρίς κάποιο ιδιαίτερο σημασιολογικό χαρακτηρισμό. Αν και για την περίπτωση του σεναρίου, έχουν αναπτυχθεί οντολογίες, αυτές είναι προσανατολισμένες βασικά προς τα δομικά συστατικά και όχι προς το περιεχόμενο.

1.1 Αντικείμενο διπλωματικής

Το αντικείμενο αυτής της διπλωματικής εργασίας ήταν η δημιουργία ενός συστήματος, το οποίο χρησιμοποιώντας πληροφορίες από τον οδηγό γονέων του IMDB καθώς και από το post-production σενάριο θα επιτύγχανε να εντοπίσει τις "προβληματικές" σκηνές μιας ταινίας και να τις αντιστοιχίσει με τις συγκεκριμένες περιγραφές του οδηγού γονέων, δίνοντας την δυνατότητα στον χρήστη να μπορεί να μεταβεί άμεσα από τις περιγραφές στις ίδιες τις "προβληματικές" σκηνές.

Η λειτουργία του συστήματος μπορεί να διαιρεθεί σε δύο βασικά στάδια. Το πρώτο αποτελείται από την μετατροπή των δεδομένων φυσικής γλώσσας που έχουμε σε μια πιο τυπική μορφή αναπαράστασης γνώσης και το δεύτερο αποτελείται από περαιτέρω επεξεργασία των αρχικών δεδομένων., χρησιμοποιώντας όμως ως βάση τα τυπικά δεδομένα που έχουν προκύψει από το πρώτο στάδιο.

Στο πρώτο στάδιο, το σύστημα κατηγοριοποιεί αφενός τις σκηνές του σεναρίου και αφετέρου τις περιγραφές του οδηγού γονέων σε οντολογικές κλάσεις σχετικές με το περιεχόμενο τους. Τέτοιες κλάσεις για παράδειγμα είναι η *SmokingEvent*, που αναφέρεται στις σκηνές εκείνες όπου κάποιος εμφανίζεται να καπνίζει, ή η

ReligiousExclamation, που αναφέρεται στις σκηνές εκείνες που περιέχουν επιφωνήματα θρησκευτικού περιεχομένου.

Για να επιτευχθεί αυτό, το σενάριο, το οποίο είναι ήδη δομημένο σε RDF τρίπλες, και το κείμενο των περιγραφών περνάνε από επεξεργασία φυσικής γλώσσας σε αναζήτηση συγκεκριμένων λέξεων-κλειδιών που αντιστοιχούν στις διάφορες κλάσεις της οντολογίας. Με βάση τις λέξεις-κλειδιά αυτές, δημιουργούνται χαρακτηριστικά διανύσματα για τις σκηνές του σεναρίου και για τις περιγραφές. Λόγω της διαφορετικής δομής των δύο περιπτώσεων, τα χαρακτηριστικά διανύσματα επίσης διαφέρουν. Για τις σκηνές του σεναρίου προκύπτουν δύο διανύσματα, ένα για το κομμάτι του διαλόγου και ένα για το κομμάτι της περιγραφής της δράσης. Για τις περιγραφές προκύπτει ένα χαρακτηριστικό διάνυσμα, το οποίο όμως εξαρτάται άμεσα από την κατηγορία του οδηγού γονέων στην οποία εμφανίζεται η περιγραφή. Τα διανύσματα αυτά συγκρίνονται με τα αντίστοιχα χαρακτηριστικά διανύσματα των κλάσεων χρησιμοποιώντας το μέτρο ομοιότητας Ochiai και ανάλογα με την απόσταση που προκύπτει κατηγοριοποιούνται ή όχι στις αντίστοιχες κλάσεις.

Ο στόχος του δεύτερου σταδίου είναι η τελική διασύνδεση μεταξύ των περιγραφών και των σκηνών της ταινίας. Μετά από την κατηγοριοποίηση τους, μια πρώτη άμεση αντιστοίχηση έχει ήδη επιτευχθεί μεταξύ των περιγραφών και των σκηνών του σεναρίου που ανήκουν στην ίδια κλάση. Χρησιμοποιώντας την μέθοδο tf-idf, επιλέγονται από το σύνολο αυτό των πρώτων αντιστοιχίσεων οι σκηνές που έχουν περισσότερη ομοιότητα με την εκάστοτε περιγραφή. Αφού έχει ολοκληρωθεί και αυτή η διαδικασία, χρησιμοποιώντας τις πληροφορίες του σεναρίου σχετικά με την χρονική στιγμή που συμβαίνει η κάθε σκηνή, γίνεται η διασύνδεση της περιγραφής με το οπτικό υλικό της σκηνής ή των σκηνών με τις οποίες έχει τις μεγαλύτερες ομοιότητες.

1.2 Οργάνωση κειμένου

Το κείμενο της διπλωματικής ακολουθεί την εξής δομή :

- Κεφάλαιο 1 : περιλαμβάνει την εισαγωγή ως προς τον στόχο και το αντικείμενο της διπλωματικής.
- Κεφάλαιο 2 : περιλαμβάνει το θεωρητικό υπόβαθρο που είναι απαραίτητο για την κατανόηση της διπλωματικής εργασίας.
- Κεφάλαιο 3 : περιλαμβάνει την περιγραφή της οντολογίας που αναπτύχθηκε
- Κεφάλαιο 4 : περιλαμβάνει την διαδικασία που ακολουθήθηκε ώστε να παραχθούν τα αποτελέσματα
- Κεφάλαιο 5 : περιλαμβάνει τα αποτελέσματα που προέκυψαν και την ανάλυση τους
- Κεφάλαιο 6 : περιλαμβάνει την περιγραφή του συστήματος
- Κεφάλαιο 7: περιλαμβάνει τον επίλογο της εργασίας.
- Κεφάλαιο 8: περιλαμβάνει την βιβλιογραφία που χρησιμοποιήθηκε για την συγγραφή της εργασίας.

2

Θεωρητικό υπόβαθρο

2.1 Σημασιολογικός ιστός και αναπαράσταση γνώσης

Ο Σημασιολογικός ιστός αποτελεί μια προσπάθεια για ένα δίκτυο δεδομένων, στο οποίο τα δεδομένα είναι τυποποιημένα και διασυνδεδεμένα μεταξύ τους. Κατά αυτόν τον τρόπο είναι εύκολα προσβάσιμα όχι μόνο από τον άνθρωπο, αλλά και από εφαρμογές.

Για να επιτευχθεί αυτό, πρέπει να υπάρχουν κοινά πρότυπα με βάση τα οποία θα περιγράφονται τα δεδομένα (RDF), καθώς και τεχνολογίες που θα κάνουν εύκολη την πρόσβαση με τυποποιημένα ερωτήματα στα δεδομένα αυτά(SPARQL).

2.1.1 RDF

Το πλαίσιο περιγραφής πόρων (Resource Description Framework-RDF)[2] είναι ένα πλαίσιο για την τυπική αναπαράσταση γνώσης στον Ιστό. Ορίζει ένα μοντέλο δεδομένων το οποίο χρησιμοποιείται ως κοινή βάση για πολλές γλώσσες αναπαράστασης γνώσης. Τα δεδομένα οργανώνονται σε τριάδες (τρίπλες) με δομή υποκείμενο (subject) - ιδιότητα (property) - αντικείμενο (object). Το υποκείμενο

μπορεί να είναι ένα IRI ή ένας κενός κόμβος, το αντικείμενο IRI, κενός κόμβος ή *literal*, ενώ η ιδιότητα μπορεί να είναι μόνο IRI.

Τα IRIs (Internationalized Resource Identifier) είναι μια συμβολοσειρά που χρησιμοποιείται ως αναγνωριστικό ενός πόρου. Αποτελεί μια γενίκευση των URIs, καθώς η συμβολοσειρά αυτή μπορεί να περιέχει χαρακτήρες Unicode σε αντίθεση με τα URIs που είναι περιορισμένα στους χαρακτήρες ASCII.
Τα *literals* χρησιμοποιούνται για να περιγράψουν αξίες αριθμών, ημερομηνιών ή συμβολοσειρών. Κάθε *literal* αποτελείται από δύο ή τρία μέρη : μια Unicode συμβολοσειρά που περιγράφει την τιμή της αξίας (“1”, “abc”), ένα IRI τύπου δεδομένων που περιγράφει τον τύπο της αξίας (String, Integer, Date) και τέλος, εάν το προηγούμενο IRI περιγράφει μια συμβολοσειρά, μια ετικέτα που διευκρινίζει σε ποια γλώσσα αντιστοιχεί.

2.1.2 RDF-Schema

Η RDF-Schema[3] είναι μια επέκταση της βασικής RDF γλώσσας και του λεξιλογίου που αυτή προσφέρει.

Η βασικότερη προσθήκη είναι η ύπαρξη κλάσεων. Οι κλάσεις είναι επίσης πόροι και ταυτοποιούνται από IRIs. Περιγράφονται χρησιμοποιώντας ιδιότητες RDF. Η ιδιότητα rdf:type χρησιμοποιείται για να δηλώσει ότι ένας πόρος είναι στιγμιότυπο μιας κλάσης. Επίσης, έχουμε και την προσθήκη περισσότερων προδιαγραφών για τις ιδιότητες, όπως την ιδιότητα rdfs:subPropertyOf.

2.1.3 SPARQL

Η SPARQL[4] είναι η γλώσσα που χρησιμοποιείται για την επικοινωνία με δεδομένα RDF. Επιτρέπει στον χρήστη να διατυπώσει ερωτήματα που περιλαμβάνουν υποχρεωτικά ή και προαιρετικά πρότυπα, καθώς και συζεύξεις ή διαζεύξεις μεταξύ

αυτών.

Τα ερωτήματα που μπορούν να τεθούν αποτελούνται από δύο μέρη. Το πρώτο προσδιορίζει τον τύπο της ερώτησης και τις μεταβλητές, εάν αυτές υπάρχουν, που θα εμφανιστούν στα αποτελέσματα. Το δεύτερο μέρος περιγράφει τις παραμέτρους που οφείλει να πληρεί η αναζήτηση. Οι μορφές ερωτημάτων που μπορούν να τεθούν είναι οι εξής :

- SELECT : επιστρέφει το σύνολο ή υποσύνολο των μεταβλητών που δεσμεύτηκαν ταιριάζοντας το πρότυπο της ερώτησης.
- CONSTRUCT : επιστρέφει ένα γράφημα RDF που δημιουργήθηκε αντικαθιστώντας τις μεταβλητές με ένα σύνολο προτύπων τριπλών.
- ASK : επιστρέφει μια τιμή αλήθειας ανάλογα με το εάν το πρότυπο του ερωτήματος πληρούται ή όχι.
- DESCRIBE : επιστρέφει ένα γράφημα RDF που περιγράφει τους πόρους που βρέθηκαν.

2.1.4 Οντολογία

Στον Σημασιολογικό Ιστό, η οντολογία αποτελεί έναν τυπικό ορισμό των τύπων, ιδιοτήτων και σχέσεων που αφορούν ένα συγκεκριμένο πεδίο ενδιαφέροντος, καθώς και τους περιορισμούς αυτών. Μπορεί να χρησιμοποιηθεί για την εξαγωγή συμπερασμάτων, καθώς και ως ένας δομημένος τρόπος οργάνωσης της πληροφορίας.

Τα κύρια χαρακτηριστικά οποιασδήποτε οντολογίας είναι τα άτομα, οι κλάσεις και οι σχέσεις.

Τα άτομα αποτελούν τα διακριτά αντικείμενα που συμπεριλαμβάνονται σε μία οντολογία. Μπορεί να είναι απτά αντικείμενα ή πρόσωπα, αφηρημένες έννοιες καθώς και τιμές αριθμών, συμβολοσειρών κ.α. Το βασικό τους χαρακτηριστικό είναι ότι είναι διακριτά και ταυτοποιήσιμα.

Οι κλάσεις αναπαριστούν έννοιες του πεδίου ενδιαφέροντος και είναι σύνολα ατόμων με κοινές ιδιότητες. Ένα άτομο μπορεί να ανήκει σε περισσότερες από μία κλάσεις, εκτός και αν οι κλάσεις αυτές έχουν οριστεί ως *disjoint*, μη έχοντας δηλαδή κοινά στοιχεία. Επίσης, μία κλάση μπορεί να υπάγεται σε άλλες κλάσεις, που αντιστοιχούν σε γενικότερες έννοιες, ή να συμπεριλαμβάνει υποκλάσεις, που να αντιστοιχούν σε κάποιες ειδικότερες έννοιες.
Οι σχέσεις αναπαριστούν τους τρόπους με τους οποίους τα διάφορα στοιχεία μιας οντολογίας σχετίζονται μεταξύ τους. Μπορεί να εκφράζουν σχέση ανάμεσα σε κλάσεις, άτομα, ή άτομο και κλάση.

2.1.5 OWL

Η OWL (Ontology Web Language)[5] είναι μία γλώσσα που έχει σχεδιαστεί για να χρησιμοποιείται από εφαρμογές που πρέπει να επεξεργαστούν πληροφορίες. Περιέχει μεγαλύτερη επεξηγηματικότητα από τις γλώσσες RDF και RDF-Schema, καθώς παρέχει επιπλέον λεξιλόγιο για την περιγραφή ιδιοτήτων και κλάσεων. Κάποιες από αυτές τις ιδιότητες είναι οι σχέσεις μεταξύ κλάσεων, αριθμητικές σχέσεις, ισότητα καθώς και χαρακτηριστικά ιδιοτήτων.

Η OWL περιλαμβάνει τρεις γλώσσες που προσφέρουν στους χρήστες διαφορετικές δυνατότητες :

- OWL Lite : υποστηρίζει τους χρήστες που χρειάζονται μία ιεραρχία ταξινόμησης και απλούς περιορισμούς. Είναι ευκολότερο να παρέχεις εργαλεία για την OWL Lite από τις υπόλοιπες γλώσσες και έχει μικρότερη πολυπλοκότητα από αυτές.
- OWL DL : υποστηρίζει τους χρήστες που θέλουν την μέγιστη εκφραστικότητα, εξασφαλίζοντας όμως συγχρόνως ότι όλα τα συμπεράσματα θα υπολογιστούν σε πεπερασμένο χρόνο. Περιλαμβάνει όλες τις δυνατότητες της γλώσσας, αλλά μπορούν να χρησιμοποιηθούν μόνο υπό ορισμένους

περιορισμούς.

- OWL Full : προορίζεται για χρήστες που θέλουν την μέγιστη εκφραστικότητα και συντακτική ελευθερία της RDF χωρίς εγγυήσεις στον τομέα των υπολογισμών.

Κάθε μία από τις γλώσσες αυτές είναι μία επέκταση του προκατόχου της, τόσο στον τομέα της έκφρασης, καθώς και στον τομέα του συμπερασμού.

2.2 Επεξεργασία φυσικής γλώσσας

Η επεξεργασία φυσικής γλώσσας είναι ένας τομέας της επιστήμης των υπολογιστών που ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και των φυσικών γλωσσών. Καθώς οι φυσικές γλώσσες είναι ιδιαίτερα πολύπλοκες, είναι δύσκολο να κατανοηθούν από τα συστήματα στην αρχική τους μορφή τους . Για τον λόγο αυτό, υποβάλλονται σε κάποιες διαδικασίες, ώστε να πάρουν μία πιο προσιτή μορφή που θα επιτρέψει την περαιτέρω επεξεργασία τους.

Παρακάτω θα παρουσιάσουμε κάποιες από τις βασικές διαδικασίες στις οποίες υποβάλλεται ένα κείμενο φυσικής γλώσσας :

-*sentence segmentation :* η διαδικασία διαίρεσης ενός κειμένου στις προτάσεις που το αποτελούν. Τα σημεία στίξης που συνήθως σηματοδοτούν το τέλος μιας πρότασης (“.”, “!”, “?”) μπορούν να χρησιμοποιηθούν και σε άλλες περιπτώσεις, για παράδειγμα μία τελεία σε ένα URL ιστοσελίδας. Οπότε, απλή εύρεση των σημείων αυτών δεν επαρκεί για να αποσαφηνίσουμε τα όρια των προτάσεων. Συνήθως είναι απαραίτητο να εξεταστεί το κείμενο που ακολουθεί, εάν ξεκινάει με κεφαλαίο ή όχι, ή ακόμα και του κειμένου που προηγείται του σημείου στίξης για να καταλήξουμε στα σωστά όρια της πρότασης.

-*tokenization* : η διαδικασία διαίρεσης ενός κειμένου στα κομμάτια που το αποτελούν.

Τα κομμάτια αυτά μπορεί να είναι λέξεις, αριθμοί ή και σύμβολα.[6] Η διαδικασία αυτή είναι απαραίτητη για να μπορέσουν να γίνουν πιο πολύπλοκες διεργασίες που γίνονται σε επίπεδο *token.*
Στην περίπτωση των αγγλικών, οι λέξεις διαχωρίζονται από κενό, εκτός από την περίπτωση των σημείων στίξης. Για να μπορέσει να γίνει διάκριση μεταξύ των περιπτώσεων που τα σημεία αυτά χωρίζουν διαφορετικά *tokens* ή όχι, χρησιμοποιείται συνήθως ένα μοντέλο μέγιστης εντροπίας, το οποίο εκπαιδεύεται καταλλήλως.

-*part-of-speech tagging* : η διαδικασία κατά την οποία το κείμενο επεξεργάζεται συντακτικά και κάθε λέξη μαρκάρεται με μία ετικέτα ανάλογα με το μέρος του λόγου στο οποίο ανήκει. Η ετικέτα αυτή προκύπτει και από την ίδια την λέξη και από την θέση της μέσα στην πρόταση, καθώς η ίδια λέξη μπορεί να έχει διαφορετικές θέσεις ανάλογα με την περίσταση. Συνήθως για την πολύπλοκη αυτή διαδικασία χρησιμοποιείται ένα μοντέλο πιθανότητας για να προβλέψει την σωστή ετικέτα.

-*stemming* : η διαδικασία αποκοπής των καταλήξεων των λέξεων ώστε να πάρουμε την ρίζα της λέξης. Η διαδικασία αυτή αποσκοπεί στο να μπορούμε να αναγνωρίσουμε τις διάφορες μορφές που μπορεί να πάρει μία λέξη, διαφορετικός αριθμός, γένος, χρόνος, κλίση κ.λ.π., ως την ίδια λέξη. Η ρίζα στην οποία μειώνονται οι λέξεις με αυτήν την αποκοπή των καταλήξεων δεν είναι απαραίτητα η ρίζα αυτή καθεαυτή της λέξης παρά μια τεχνητή ψευδο-ρίζα που προκύπτει ακολουθώντας τους κανόνες αποκοπής. Καθώς όμως δεν μας απασχολεί στην πραγματικότητα να βρούμε την ρίζα της λέξης, αλλά να την ταυτοποιήσουμε με άλλες της μορφές, η διαδικασία αυτή δουλεύει αρκετά καλά.
Ο πιο γνωστός αλγόριθμος αποκοπής είναι ο αλγόριθμος Porter[7]. Η λέξη περνάει από διάφορα στάδια αποκοπής (αφαιρώντας καταλήξεις πληθυντικού -es, -s, συνήθεις καταλήξεις ρηματικών μορφών όπως -ed και -ing και παραγωγικές καταλήξεις όπως

-ment, -able και -ism) καθώς και μετατροπής των καταλήξεων (η κατάληξη -y μετατρέπεται σε -i).

-*lemmatisation* : η διαδικασία αντιστοίχησης μίας λέξης με το λήμμα από το οποίο προκύπτει. Η διαδικασία αυτή έχει τον ίδιο στόχο με το *stemming* αλλά ακολουθεί διαφορετική πορεία. Ενώ στην περίπτωση του *stemming* η διαδικασία παίρνει υπόψιν της μόνο την λέξη, η διαδικασία του *lemmatisation* συμπεριλαμβάνει το τι μέρος του λόγου είναι η εξεταζόμενη λέξη.
Η διαδικασία αυτή είναι πιο αργή και πιο περίπλοκη από το *stemming*, αλλά γενικά επιφέρει καλύτερα αποτελέσματα.

Μια άλλη διαδικασία που δεν είναι τόσο τυπική όσο οι προηγούμενες, αλλά συνήθως είναι προτιμότερο να ακολουθηθεί, είναι η αφαίρεση κάποιων λέξεων, ώστε να μην επεξεργαστούν περαιτέρω. Οι λέξεις αυτές ονομάζονται *stop words* και είναι κοινές λέξεις που εμφανίζονται συχνά. Το σύνολο αυτό των λέξεων δεν είναι δεδομένο. Μπορεί να περιέχει μόνο γενικές κοινές λέξεις, όπως άρθρα, το ρήμα *be* ή προσωπικές αντωνυμίες,, ή αν τα εξεταζόμενα κείμενα έχουν κάποιο κοινό θέμα, μπορεί να διευρυνθεί και να περιλάβει και σχετικές θεματικές λέξεις που εμφανίζονται συχνά.

2.3 Μέτρα ομοιότητας

Η δυαδική ομοιότητα είναι ένα ιδιαίτερα βασικό στοιχείο σε προβλήματα ανάλυσης προτύπων όπως η ομαδοποίηση (*clustering*), η ταξινόμηση (*classification*) κ.α.[8] Καθώς τέτοια μέτρα ομοιότητας έχουν χρησιμοποιηθεί σε εφαρμογές διαφόρων πεδίων έρευνας, υπάρχουν πολυάριθμα μέτρα ομοιότητας που μπορεί να καλύψουν τις ανάγκες μας.

Θεωρούμε ότι έχουμε δυο δυαδικά διανύσματα χαρακτηριστικών *i* και *j*, τα οποία

επιθυμούμε να συγκρίνουμε. Για τα διανύσματα αυτά θα συμβολίσουμε ως εξής τα παρακάτω :

- α : ο αριθμός των χαρακτηριστικών του διανύσματος που και στα δύο διανύσματα οι αξίες είναι 1, χαρακτηριστικά δηλαδή που έχουν και τα δύο εξεταζόμενα διανύσματα. (*positive matches*)
- b : ο αριθμός των χαρακτηριστικών του διανύσματος που για το διάνυσμα *i* η αξία είναι 1 ενώ για το διάνυσμα *j* η αξία είναι 0. (*i absence mismatches*)
- c : ο αριθμός των χαρακτηριστικών του διανύσματος που για το διάνυσμα *i* η αξία είναι 0 ενώ για το διάνυσμα *j* η αξία είναι 1. (*j absence mismatches*)
- d : ο αριθμός των χαρακτηριστικών του διανύσματος που και στα δύο διανύσματα οι αξίες είναι 0, , χαρακτηριστικά δηλαδή που δεν έχει κανένα από τα δύο εξεταζόμενα διανύσματα. (*negative matches*)

Τα κριτήρια ομοιότητας χωρίζονται σε τρεις κατηγορίες, στα κριτήρια απόστασης, μη-συσχετισμού και συσχετισμού.

Τα κριτήρια απόστασης βασίζονται στην ομοιότητα μεταξύ των δύο εξεταζόμενων διανυσμάτων. Τέτοιες αποστάσεις είναι μεταξύ άλλων :

- δυαδική Ευκλείδεια : $D=\sqrt{b+c}$
- Hamming : $D=b+c$
- Bray & Curtis : $D=\frac{b+c}{2\mathrm{a}+b+c}$

Η κατηγορία των μέτρων ομοιότητας μη-συσχετισμού αποτελείται από σχετικά απλές μετρικές αθροισμάτων των 4 πληθών των όμοιων και ανόμοιων χαρακτηριστικών και των μεταξύ τους πηλίκων.

Στην κατηγορία αυτή ανήκουν τα παρακάτω μέτρα ομοιότητας :

- Jaccard : $S=\frac{a}{a+b+c}$
- Russel & Rao : $S=\frac{a}{a+b+c+d}$
- Dice & Sorenson : $S=\frac{2\mathrm{a}}{2\mathrm{a}+b+c}$
- Roger & Tanimoto : $S=\frac{a+d}{a+2(b+c)+d}$
- Sokal & Michener : $S=\frac{a+d}{a+b+c+d}$
- Faith : $S=\frac{a+0.5\mathrm{d}}{a+b+c+d}$
- Gower & Legendre : $S=\frac{a+d}{a+0.5(b+c)+d}$

Η τρίτη κατηγορία μέτρων ομοιότητας είναι η κατηγορία συσχετισμού. Στην κατηγορία αυτή συμπεριλαμβάνονται πιο περίπλοκα μέτρα, στα οποία τα πλήθη των όμοιων και ανόμοιων χαρακτηριστικών συσχετίζονται μέσω γινομένων και ριζών και όχι μόνο μέσω αθροισμάτων. Τέτοια μέτρα είναι τα παρακάτω :

- Sokal & Sneath I : $S=\frac{a}{a+2\mathrm{b}+2\mathrm{c}}$
- Sokal & Sneath II : $S=\frac{2(a+d)}{2\mathrm{a}+b+c+2\mathrm{d}}$
- Sokal & Sneath III : $S=\frac{a+d}{b+c}$
- Sokal & Sneath IV : $S=\frac{\frac{a}{a+b}+\frac{d}{b+d}+\frac{a}{a+c}+\frac{d}{c+d}}{4}$
- Baroni-Urbani I : $S=\frac{a+d}{b+c}$

- Baroni-Urbani II : $S=\frac{\sqrt{ad}+a-(b+c)}{\sqrt{ad}+a+b+c}$
- Sorgenfrei : $S=\frac{a^2}{(a+b)(a+c)}$
- Ochiai I : $S=\frac{a}{\sqrt{(a+b)(a+c)}}$
- Simpson : $S=\frac{a}{min(a+b)(a+c)}$
- Braun & Banquet : $S=\frac{a}{max(a+b)(a+c)}$
- Peirce : $S=\frac{ab+bc}{ab+2\mathrm{bc}+cd}$
- Yule : $S=\frac{ab-bc}{ad+bc}$
- Michael : $S=\frac{4(ab-bc)}{(a+d)^2+(b+c)^2}$
- McConnaughey : $S=\frac{a^2-bc}{(a+b)(a+c)}$

2.4 Μέθοδος tf-idf

Η μέθοδος *tf-idf* (*term frequency-inverse document frequency*) είναι μια μέθοδος που μας δίνει ένα μέτρο του πόσο σημαντικός είναι ένας όρος *t* σε ένα σύνολο εγγράφων *D*.[9] Το μέτρο αυξάνεται με την αύξηση του αριθμού των εμφανίσεων μιας λέξης σε ένα έγγραφο, αλλά μειώνεται με την συχνότητα εμφάνισης της λέξης στο σύνολο των εγγράφων, αντισταθμίζοντας έτσι την πιθανή επικράτηση των λέξεων που χρησιμοποιούνται συχνά εν γένει.

Η συχνότητα όρου (*tf(t,d)*) μας δίνει το πόσο συχνά εμφανίζεται ένας όρος *t* σε ένα έγγραφο *d*. Ο πιο συνήθης τρόπος υπολογισμού της συχνότητας όρου είναι με απλή απαρίθμηση των φορών που εμφανίζεται ο όρος *t* στο έγγραφο *d*. Το μέγεθος του

κειμένου των εγγράφων συχνά διαφέρει πολύ από περίπτωση σε περίπτωση, και η απλή απαρίθμηση θα ευνοούσε τα μεγαλύτερα έγγραφα, ενώ αντίθετα σημαντικές πληροφορίες των μικρότερων εγγράφων θα χάνονταν. Οπότε είναι προτιμότερο να χρησιμοποιηθεί μια κανονικοποιημένη εκδοχή υπολογισμού της συχνότητας όρου :

$$tf(t,d) = \frac{f_{t,d}}{n_d} \quad ,$$

όπου $f_{t,d}$ ο αριθμός εμφανίσεων του όρου *t* στο έγγραφο *d* και n_d το σύνολο των λέξεων του εγγράφου *d*.

Η αντίστροφη συχνότητα εγγράφου (*idf(t,D)*) μας δίνει ένα μέτρο για το μέγεθος της πληροφορίας που "περιέχει" ο όρος *t* στο σύνολο των εγγράφων *D*. Όσο πιο σπάνια εμφανίζεται ο όρος τόσο περισσότερη πληροφορία μας προσφέρει. Το μέτρο αυτό υπολογίζεται για κάθε όρο ως ο λογάριθμος του πηλίκου του αριθμού των εγγράφων *N* ως προς τον αριθμό των εγγράφων που ο όρος *t* εμφανίζεται :

$$idf(t,D) = \log \frac{N}{|d \in D : t \in D|}$$

Καθώς υπάρχει η πιθανότητα ο εξεταζόμενος όρος να μην εμφανίζεται σε κανένα από τα έγγραφα με αποτέλεσμα ο παρανομαστής να μηδενίζεται, η παραπάνω εξίσωση τροποποιείται ως εξής :

$$idf(t,D) = \log \frac{N}{1 + |d \in D : t \in D|}$$

Τέλος, το συνολικό χαρακτηριστικό βάρος *tf-idf(t,d,D)* κάθε εγγράφου *d* για τον όρο *t* με βάση το συγκεκριμένο σύνολο εγγράφων *D* υπολογίζεται από το γινόμενο της συχνότητας όρου *tf(t,d)* και την αντίστροφη συχνότητα εγγράφου *idf(t,D)* :

$$tf - idf(t,d,D) = tf(t,d)\, idf(t,D)$$

2.5 Ομοιότητα συνημιτόνου

Η ομοιότητα συνημιτόνου είναι ένα μέτρο της ομοιότητας μεταξύ δύο διανυσμάτων. Βασίζεται στην γωνία που δημιουργείται μεταξύ των δύο διανυσμάτων, εάν αυτά παρασταθούν σε ένα n-διάστατο χώρο. Όσο μικρότερη είναι η γωνία, τόσο πιο όμοια είναι τα διανύσματα.

Το συνημίτονο μεταξύ των δύο διανυσμάτων μπορεί να υπολογιστεί από το εσωτερικό γινόμενο μεταξύ των διανυσμάτων. Θεωρώντας δύο διανύσματα *d* και *q,* έχουμε :

$$dq = |d||q|\cos(\theta) \quad ,$$

όπου θ η μεταξύ τους γωνία.

Η ομοιότητα των διανυσμάτων ισούται με το συνημίτονο αυτό μεταξύ των διανυσμάτων και δίνεται από τον παρακάτω τύπο :

$$S = \frac{dq}{|d||q|} = \frac{\sum_{i=1}^{n} d_i q_i}{\sqrt{\sum_{i=1}^{n} d_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}}$$

3

Δημιουργία οντολογίας

Ο πρώτος στόχος της διπλωματικής ήταν να επιτευχθεί μία τυπική μοντελοποίηση της γνώσης που αφορά το περιεχόμενο των "προβληματικών" σκηνών. Για να επιτευχθεί αυτό, έγινε επέκταση της ήδη υπάρχουσας οντολογίας *scriptontology,* με έννοιες που αναπαριστούν αυτήν την γνώση.

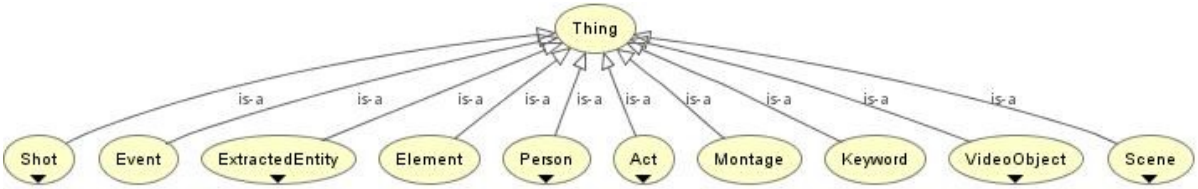
Στην συνέχεια, θα παρουσιάσουμε την ήδη υπάρχουσα οντολογία και την επέκτασή της.

3.1 Υπάρχουσα οντολογία

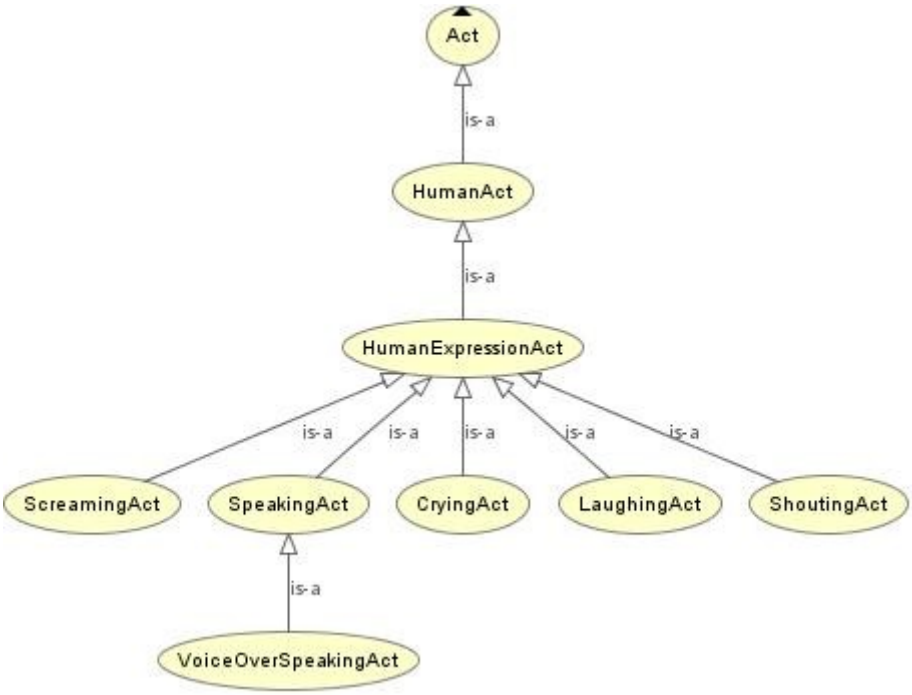
Η οντολογία *scriptontology* του εργαστηρίου image.ntua.gr περιέχει κλάσεις που ανταποκρίνονται στα βασικά δομικά χαρακτηριστικά μιας ταινίας και είναι προσανατολισμένη στο να καλύψει τις τεχνικές λεπτομέρειες, την σειρά των σκηνών, την αντιστοιχία των ρόλων με τους διαλόγους, τον τύπο κινηματογράφησης των πλάνων κλπ.

Τα σενάρια με τα οποία δουλέψαμε ήταν ήδη οργανωμένα σε τρίπλες σύμφωνα με την οντολογία αυτή.

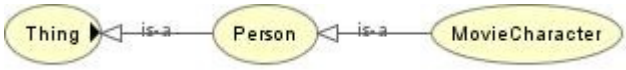
Στη συνέχεια βλέπουμε πιο αναλυτικά την οργάνωση της :



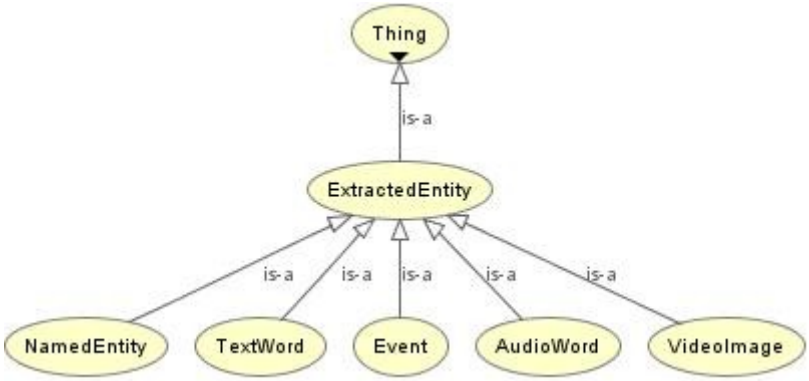
Σχήμα 3.1 : Βασικές κλάσεις της scriptontology



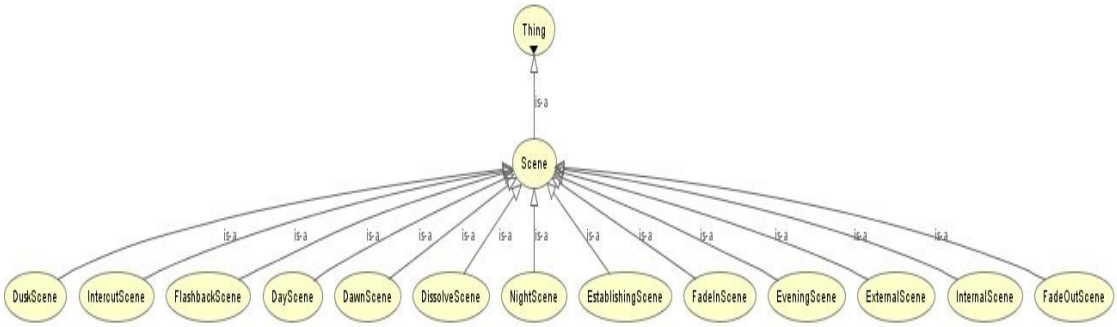
Σχήμα 3.2 : Υποκλάσεις της κλάσης Act



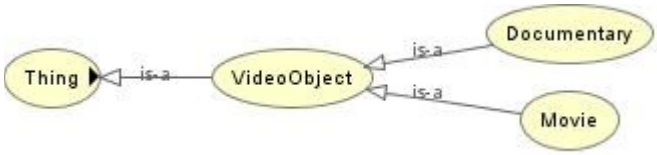
Σχήμα 3.3 : Υποκλάση της κλάσης Person



Σχήμα 3.4 : Υποκλάσεις της κλάσης ExctractedEntity



Σχήμα 3.5 : Υποκλάσεις της κλάσης Scene



Σχήμα 3.6 : Υποκλάσεις της κλάσης VideoObject

Σχήμα 3.7 : Υποκλάσεις της κλάσης Shot

3.2 Επέκταση οντολογίας

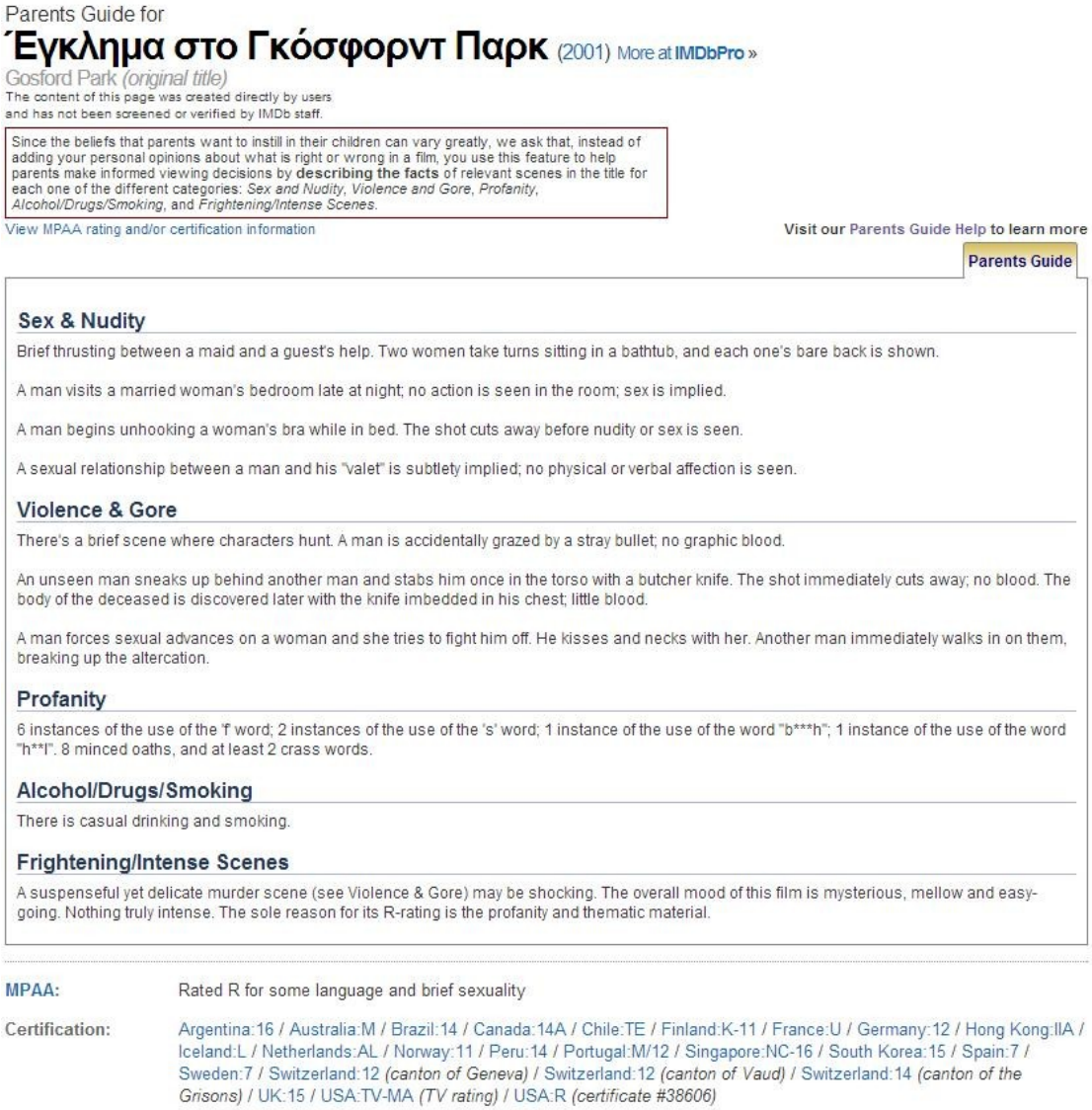
Για να καλυφθούν οι ανάγκες αναπαράστασης των επιλήψιμων σκηνών, η οντολογία επεκτάθηκε με νέες κλάσεις. Οι κλάσεις αυτές αντιστοιχούν στις κατηγορίες, στις οποίες οι σκηνές αυτές μπορεί να εμπίπτουν ανάλογα με το περιεχόμενο τους.

Η επέκταση αυτή έγινε κυρίως με βάση τις περιγραφές του οδηγού γονέων του imdb, και σε επίπεδο πληροφορίας για τις κατηγορίες των επιλήψιμων σκηνών και σε επίπεδο δομής της σελίδας, την οποία ακολουθήσαμε στην οντολογία.

3.2.1 Δομή σελίδας imdb parental guide

Ο οδηγός γονέων χωρίζεται σε 5 βασικές κατηγορίες, στις οποίες τοποθετούν οι χρήστες τις αντίστοιχες περιγραφές, έτσι ώστε να υπάρχει μια βασική δομή των δεδομένων παρά των ανθρώπινο παράγοντα που υπεισέρχεται. Οι κατηγορίες αυτές είναι οι εξής : Σεξ και Γυμνό (Sex and Nudity), Βία και Αίμα (Violence and Gore), Βλασφημία (Profanity), Αλκοόλ/Ναρκωτικά/Κάπνισμα (Alcohol/Drugs/Smoking) και Τρομακτικές/Εντονες Σκηνές (Frightening/Intense Scenes).

Παρακάτω μπορούμε να δούμε ένα παράδειγμα μίας τέτοιας σελίδας. Συγκεκριμένα τον οδηγό γονέων για την ταινία *Gosford Park*, η οποία αφού έχει χαρακτηριστεί ως R, έχει αρκετές επιλήψιμες σκηνές και αποτελεί ένα σχετικά χαρακτηριστικό παράδειγμα ενός τέτοιου οδηγού.


Parents Guide for

Έγκλημα στο Γκόσφορντ Παρκ (2001) More at IMDbPro »

Gosford Park *(original title)*

The content of this page was created directly by users and has not been screened or verified by IMDb staff.

Since the beliefs that parents want to instill in their children can vary greatly, we ask that, instead of adding your personal opinions about what is right or wrong in a film, you use this feature to help parents make informed viewing decisions by **describing the facts** of relevant scenes in the title for each one of the different categories: *Sex and Nudity*, *Violence and Gore*, *Profanity*, *Alcohol/Drugs/Smoking*, and *Frightening/Intense Scenes*.

View MPAA rating and/or certification information

Visit our Parents Guide Help to learn more

Parents Guide

Sex & Nudity

Brief thrusting between a maid and a guest's help. Two women take turns sitting in a bathtub, and each one's bare back is shown.

A man visits a married woman's bedroom late at night; no action is seen in the room; sex is implied.

A man begins unhooking a woman's bra while in bed. The shot cuts away before nudity or sex is seen.

A sexual relationship between a man and his "valet" is subtlety implied; no physical or verbal affection is seen.

Violence & Gore

There's a brief scene where characters hunt. A man is accidentally grazed by a stray bullet; no graphic blood.

An unseen man sneaks up behind another man and stabs him once in the torso with a butcher knife. The shot immediately cuts away; no blood. The body of the deceased is discovered later with the knife imbedded in his chest; little blood.

A man forces sexual advances on a woman and she tries to fight him off. He kisses and necks with her. Another man immediately walks in on them, breaking up the altercation.

Profanity

6 instances of the use of the 'f' word; 2 instances of the use of the 's' word; 1 instance of the use of the word "b***h"; 1 instance of the use of the word "h**l". 8 minced oaths, and at least 2 crass words.

Alcohol/Drugs/Smoking

There is casual drinking and smoking.

Frightening/Intense Scenes

A suspenseful yet delicate murder scene (see Violence & Gore) may be shocking. The overall mood of this film is mysterious, mellow and easy-going. Nothing truly intense. The sole reason for its R-rating is the profanity and thematic material.

MPAA: Rated R for some language and brief sexuality

Certification: Argentina:16 / Australia:M / Brazil:14 / Canada:14A / Chile:TE / Finland:K-11 / France:U / Germany:12 / Hong Kong:IIA / Iceland:L / Netherlands:AL / Norway:11 / Peru:14 / Portugal:M/12 / Singapore:NC-16 / South Korea:15 / Spain:7 / Sweden:7 / Switzerland:12 *(canton of Geneva)* / Switzerland:12 *(canton of Vaud)* / Switzerland:14 *(canton of the Grisons)* / UK:15 / USA:TV-MA *(TV rating)* / USA:R *(certificate #38606)*

Σχήμα 3.8 : Σελίδα οδηγού γονέων της ταινίας Gosford Park

Αν και οι περιγραφές, αφού προέρχονται από χρήστες, δεν μπορεί να είναι πλήρως τυποποιημένες, τείνουν ωστόσο να ακολουθούν κάποια πρότυπα.

Όπως βλέπουμε και παραπάνω, οι σκηνές που εμπίπτουν στις κατηγορίες *Sex & Nudity* και *Violence & Gore* περιγράφονται ιδιαιτέρως αναλυτικά.

Οι περιγραφές της κατηγορίας *Profanity* συνήθως αποτελούνται από μια απαρίθμηση των όρων που έχουν χρησιμοποιηθεί, πολλές φορές χρησιμοποιώντας παραφράσεις, όπως *f-word* και *sexual references*, και αντικαθιστώντας με αστερίσκους ή παύλες γράμματα των λέξεων, *s**t, f--k,* σε μια προσπάθεια να αποφύγουν μια ευθεία αναφορά.

Οι περιγραφές της κατηγορίας *Alcohol/Drugs/Smoking* τείνουν επίσης να είναι ιδιαίτερα επιγραμματικές. Αυτό οφείλεται στο ότι είναι μια πιο στενή κατηγορία , όσον αφορά την ποικιλία των γεγονότων, από τις δύο πρώτες, που επίσης περιγράφουν γεγονότα δράσης και όχι διαλόγου όπως η κατηγορία *Profanity*. Επίσης, όπως φαίνεται και από τα κριτήρια των χαρακτηρισμών της MPAA, με εξαίρεση την χρήση ναρκωτικών, οι σκηνές αυτές δεν θεωρούνται τόσο δριμείς και επόμενο είναι οι χρήστες να μην τους αφιερώνουν την ίδια προσοχή.

Τέλος, η κατηγορία *Frightening/Intense Scenes* δεν παρουσιάζει κάποιο ιδιαίτερο πρότυπο, εκτός από τις πιθανές παραπομπές σε κάποια άλλη κατηγορία, συνήθως την κατηγορία *Violence & Gore.*

3.2.2 Κλάσεις της οντολογίας

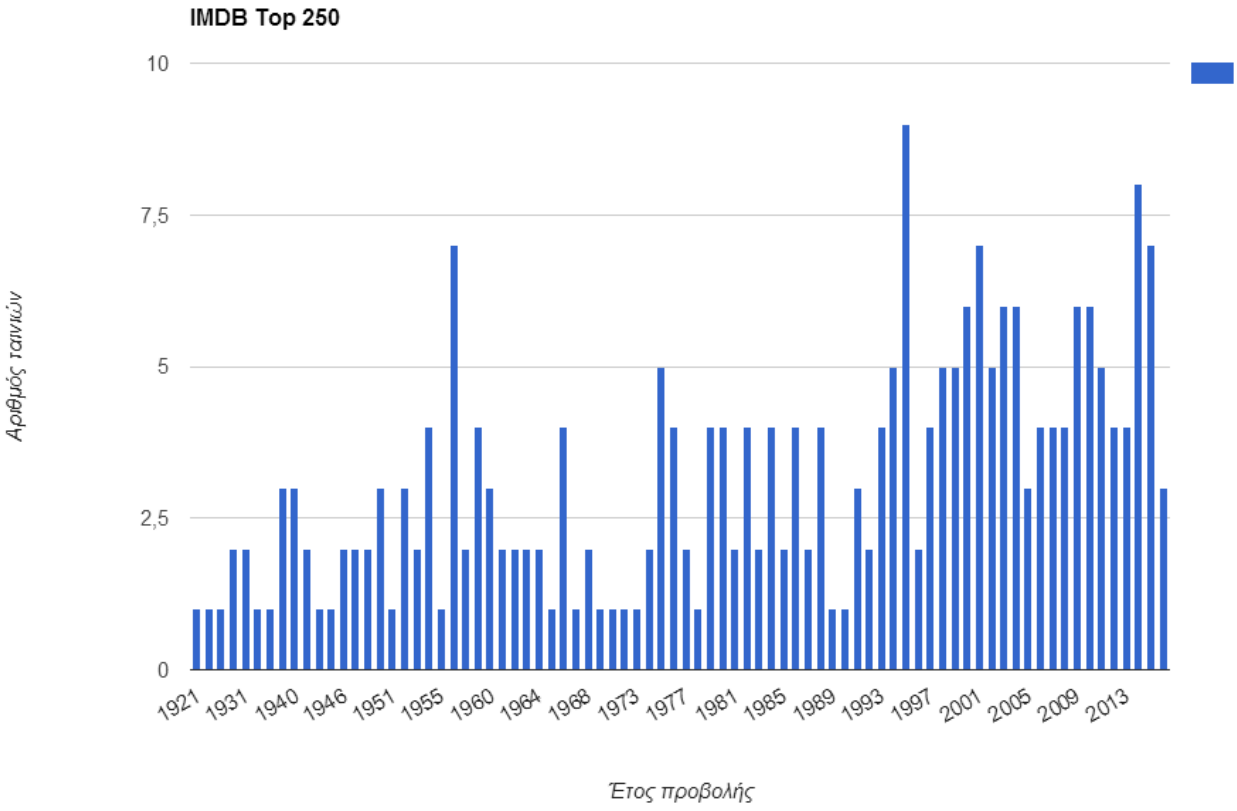
Για να καταλήξουμε στο τι είδους σκηνές περιέχονται συνήθως στις περιγραφές και να δημιουργήσουμε τις αντίστοιχες κλάσεις, εξετάστηκε ένα δείγμα ταινιών και των αντίστοιχων οδηγών γονέων τους.

Το δείγμα αυτό αποτελείται από τις 250 ταινίες με την υψηλότερη βαθμολογία στο imdb.[10] Επιλέχτηκε καθώς πληρεί τα παρακάτω κριτήρια και θα μας δώσει μια αρκετά αντιπροσωπευτική άποψη για το τι περιλαμβάνουν συνήθως οι περιγραφές. Πρώτον, είναι ένα αρκετά μεγάλο δείγμα ταινιών. Δεύτερον, εκτός από την βαθμολογία της ταινίας, για να συμπεριληφθεί στην λίστα αυτή μία ταινία πρέπει να έχει τουλάχιστον 25.000 ψήφους. Επομένως, το δείγμα απαρτίζεται από δημοφιλείς ταινίες οπότε είναι πολύ πιθανό ο οδηγός γονέων να είναι ενημερωμένος και σωστός.

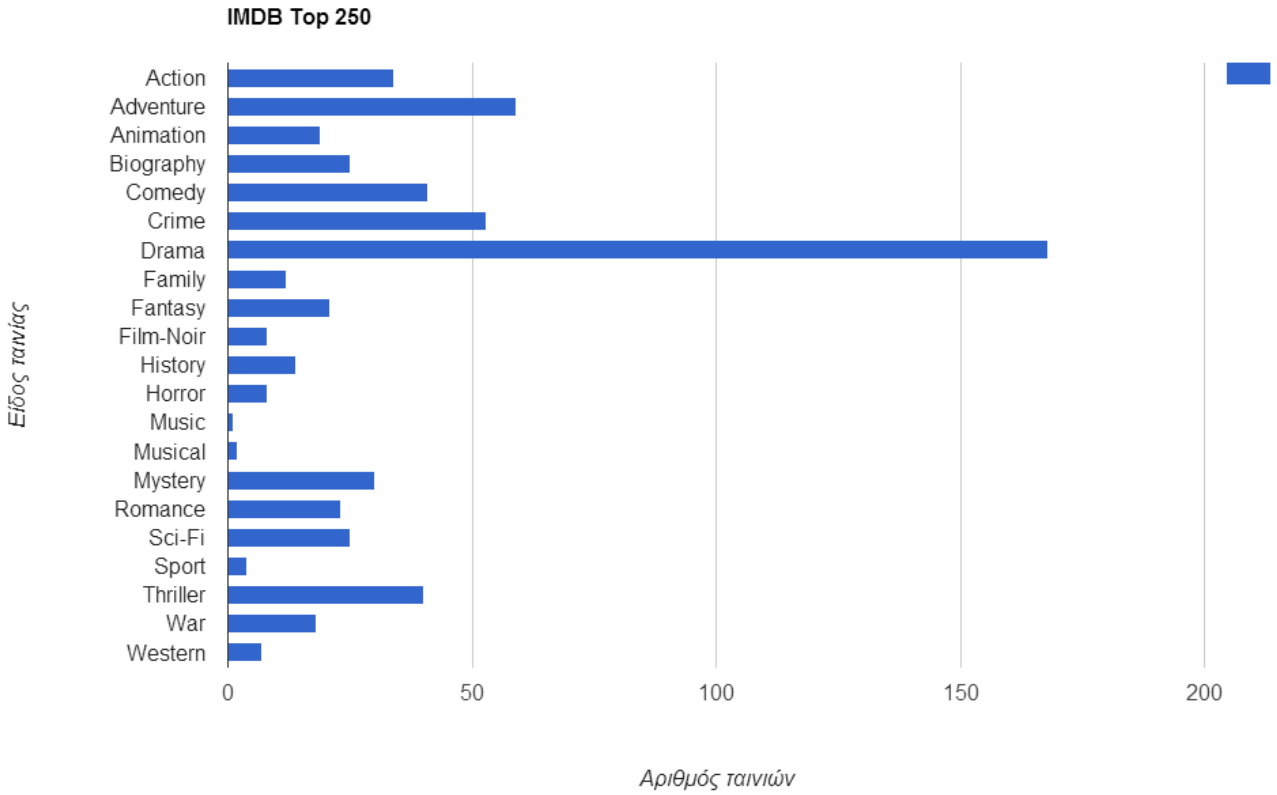
Τέλος, όπως φαίνεται και από τα διαγράμματα παρακάτω, το δείγμα είναι αρκετά ποικίλο και ως προς το είδος της ταινίας και ως προς το έτος παραγωγής, παράγοντες που επηρεάζουν σημαντικά το περιεχόμενο της ταινίας. Ακόμα και οι αποκλίσεις που παρατηρούνται και ως προς το έτος και ως προς το είδος, περισσότερο μας βοηθούν στην περίπτωση αυτή.

Ως προς το έτος, έχουμε περισσότερες ταινίες πρόσφατων ετών, που είναι αναμενόμενα πιο δημοφιλείς και γνωστές στο σημερινό κοινό που καταρτίζει τις λίστες αυτές. Ταινίες που είχαν παραχθεί μεταξύ των χρονολογιών 1934-1954 υποβάλλονταν στων κώδικα καταλληλότητας Hays, που πρακτικά απαγόρευε οποιαδήποτε από τις επιλήψιμες σκηνές με τις οποίες ασχολούμαστε.

Αντίστοιχα, ως προς το είδος, οι πιο "ελαφριές" ταινίες, χωρίς δύσκολα θέματα, σπανίως θεωρούνται ταινίες επιπέδου. Ως αποτέλεσμα, έχουμε πολύ μεγάλο δείγμα δραματικών ταινιών, όπως και περιπέτειας ή εγκλήματος, ταινίες δηλαδή που λόγω του περιεχομένου τους είναι πιθανό να έχουν αρκετές επιλήψιμες σκηνές με αναλυτικές περιγραφές που θα μας δώσουν αρκετές πληροφορίες. Τα είδη από τα οποία έχουμε πολύ μικρό δείγμα, είτε είναι είδη, μιούζικαλ, αθλητικές, οικογενειακού προσανατολισμού, που είναι εξαιρετικά απίθανο να περιείχαν επιλήψιμες σκηνές είτε είδη που αποτελούν μια δεύτερη κατηγορία κάποιας από τις ιδιαίτερα δημοφιλείς, όπως το νουάρ των ταινιών εγκλήματος.



Σχήμα 3.9 : Διάγραμμα ετών προβολής για τις 250 πιο δημοφιλείς ταινίες



Σχήμα 3.10 : Διάγραμμα ειδών ταινιών για τις 250 πιο δημοφιλείς ταινίες

Όπως αναφέραμε και πριν, οι κλάσεις της οντολογίας, ακολουθούν την δομή του parental guide του imdb, αφού θέλουμε να προσθέσουμε επιπλέον λειτουργίες σε αυτό.

Οι βασικές κατηγορίες που εμφανίζονται στον οδηγό είναι 5. Από αυτές, η πέμπτη κατηγορία του οδηγού (Frightening/Intense Scenes) δεν χρησιμοποιήθηκε καθόλου.

Μετά από εξέταση του δείγματος, διαπιστώθηκε ότι η πλειοψηφία των σκηνών που

περιγράφεται υπάγεται και σε κάποια άλλη κατηγορία, συνήθως στις σκηνές βίας, οπότε δεν θα ήταν απαραίτητο να επανεξεταστεί. Από την ίδια εξέταση προέκυψε επίσης ότι το κριτήριο για να ενταχθεί μια σκηνή στην κατηγορία αυτή εξαρτιόνταν έντονα από την εικόνα και τον ήχο της ταινίας. Ένα χαρακτηριστικό παράδειγμα είναι τα λεγόμενα "jump scares" όπου κάτι, όχι απαραίτητα τρομακτικό, εμφανίζεται αναπάντεχα, συνήθως με την συνοδεία κάποιου έντονου ήχου. Λαμβάνοντας υπόψιν ότι το βασικό μας υλικό από το οποίο θα πάρουμε στοιχεία για την ταινία είναι το σενάριο της ταινίας, το οποίο έχει μεν πλήρη περιγραφή των σκηνών αλλά δεν μπορεί να αποδώσει τις λεπτομέρειες αυτές, δεν έχουμε αρκετά δεδομένα για να εντοπίσουμε σωστά τα γεγονότα αυτά. Τέλος, όπως θα αναλυθεί αργότερα, για την κατηγοριοποίηση των σκηνών χρησιμοποιήθηκε ένα σύνολο με χαρακτηριστικές λέξεις κλειδιά. Οι σκηνές αυτής της κατηγορίας ήταν τόσο διάφορες που δεν ήταν δυνατό να βρεθούν κάποιες κοινές χαρακτηριστικές λέξεις.

Οι υπόλοιπες τέσσερις κατηγορίες του οδηγού (Sex and Nudity, Violence and Gore, Profanity, Alcohol/Drugs/Smoking) χρησιμοποιήθηκαν ως βασικές κλάσεις της επέκτασης της οντολογίας. Καθώς βέβαια οι κατηγορίες αυτές είναι ιδιαίτερα ευρείς, κάθε μία από αυτές έχει διάφορες υποκλάσεις που αντιστοιχούν σε πιο συγκεκριμένα γεγονότα που μπορεί να συμβαίνουν στην ταινία. Τα γεγονότα αυτά προέκυψαν επίσης από εξέταση του δείγματος των 250 ταινιών που αναφέραμε και των περιγραφών τους.

Πιο αναλυτικά έχουμε :

1. Κλάση SexNudityEvent : στην συγκεκριμένη κλάση συμπεριλαμβάνονται σκηνές που περιέχουν πράξεις σεξουαλικού τύπου και γυμνού.

Οι υποκλάσεις που υπάγονται στην κλάση αυτή είναι οι εξής :

- KissingEvent : σκηνές που περιέχουν δύο άτομα να φιλιούνται στο στόμα.
- SexEvent : σκηνές που περιέχουν σεξουαλικές πράξεις.

    - OralSexEvent : σκηνές που περιέχουν στοματικό σεξ.
    - MasturbationEvent : σκηνές που περιέχουν αυνανισμό.
    - RapeEvent : σκηνές που περιέχουν βιασμό.
- SexualReference : σκηνές που περιέχουν αναφορές σεξουαλικού περιεχομένου. (Η συγκεκριμένη κλάση εμφανίζεται και στην κατηγορία του υβριστικού περιεχομένου, καθώς οι χρήστες τοποθετούν τις αντίστοιχες περιγραφές των σκηνών αυτών και στις δύο κατηγορίες.)
- NudityEvent : σκηνές που περιέχουν γυμνό.
    - PartialNudityEvent : στις σκηνές αυτές έχουν αφαιρεθεί κάποια ή όλα τα ρούχα κάποιου εμπλεκόμενου, αλλά οι περιοχές που κανονικά καλύπτονται από εσώρουχα είναι καλυμμένες.
    - FullNudityEvent : στις σκηνές αυτές κάποια ή όλες οι περιοχές που κανονικά καλύπτονται από εσώρουχα είναι ακάλυπτες.

2. Κλάση ViolenceGoreEvent : στην συγκεκριμένη κλάση συμπεριλαμβάνονται όλες οι σκηνές που περιλαμβάνουν βία, είτε σωματική είτε ένοπλη, καθώς και οι απεικονίσεις των αποτελεσμάτων της. Στην κατηγορία αυτή επίσης συμπεριλαμβάνονται και οι γενικότερες σκηνές καταστροφής, ακόμα και αν δεν υπάρχουν ανθρώπινα θύματα.

Οι υποκλάσεις που υπάγονται στην κλάση αυτή είναι οι εξής :

- WeaponRelatedEvent : σκηνές που εμφανίζονται ή χρησιμοποιούνται όπλα.
    - MassiveWeaponryEvent : τα όπλα που εμφανίζονται ή χρησιμοποιούνται είναι μεγάλου βελησνεκούς, όπως χειροβομβίδες, φλογοβόλα, τανκς, κ.λ.π.
    - BladeEvent : τα όπλα που εμφανίζονται ή χρησιμοποιούνται είναι

λεπίδες κάποιου είδους.

    - MedievalWeaponryEvent : τα όπλα που εμφανίζονται ή χρησιμοποιούνται είναι προϋπάρχοντα της πυρίτιδας, σπαθιά, τόξα, κ.λ.π.
    - GunRelatedEvent : τα όπλα που εμφανίζονται ή χρησιμοποιούνται είναι πυροβόλα.

- ExplosionEvent : σκηνές που περιέχουν έκρηξη.
- FireEvent : σκηνές που περιέχουν φωτιά.
- Death : σκηνές που κάποιος απεικονίζεται να πεθαίνει.
    - Electrocution : ο τρόπος θανάτου είναι ηλεκτροπληξία.
    - Asphyxiation : ο τρόπος θανάτου είναι ασφυξία.
- PhysicalViolence : σκηνές που απεικονίζεται φυσική βία.
- GoreEvent : σκηνές που απεικονίζουν γραφικές απεικονίσεις τραυμάτων, αίματος, κ.λ.π.
    - InjuryEvent : σκηνές που απεικονίζουν κάποιο εμφανές τραύμα.
    - DeadBodyEvent : σκηνές που απεικονίζεται κάποιο πτώμα, ανεξάρτητα από το αν ο θάνατος του έχει επέλθει στην σκηνή αυτή.

3. Κλάση ProfanityEvent : στην συγκεκριμένη κλάση συμπεριλαμβάνονται όλες οι σκηνές που περιλαμβάνουν υβριστικό περιεχόμενο.

Οι υποκλάσεις που υπάγονται στην κλάση αυτή είναι οι εξής :

- RacialInsult : σκηνές που περιέχουν ρατσιστικές ή εθνικιστικές προσβολές.
- HomophobicInsult : σκηνές που περιέχουν ομοφοβικές προσβολές.

- GenderedInsult : σκηνές που περιέχουν έμφυλες προσβολές.
- HeavyInsult : σκηνές που περιέχουν “βαριές” προσβολές.
- CasualInsult : σκηνές που περιέχουν “ελαφριές” προσβολές.
- ExcrementReference : σκηνές που περιέχουν αναφορές σε περιττώματα.
- AnatomicalReference : σκηνές που περιέχουν αναφορές σε ακατάλληλα μέρη του ανθρώπινου σώματος, συνήθως γεννητικά όργανα.
- SexualReference : σκηνές που περιέχουν αναφορές σεξουαλικού περιεχομένου.
- ReligiousReference : σκηνές που περιέχουν αναφορές θρησκευτικού περιεχομένου.
- Exclamation : σκηνές που περιέχουν ακατάλληλα επιφωνήματα.
    - ReligiousExclamation : σκηνές που περιέχουν επιφωνήματα θρησκευτικού περιεχομένου.
- F-word : σκηνές που περιέχουν την λέξη “fuck”.

4. Κλάση AlcoholDrugSmokingEvent : στην συγκεκριμένη κλάση συμπεριλαμβάνονται όλες οι σκηνές που περιλαμβάνουν αναφορά ή χρήση αλκοόλ, ναρκωτικών ή καπνού. Οι υποκλάσεις που υπάγονται στην κλάση αυτή είναι οι εξής :

- SmokingEvent : σκηνές που κάποιος εμφανίζεται να καπνίζει.
- SmokingMention : σκηνές που αναφέρεται κάτι σχετικό με κάπνισμα, ανεξάρτητα με το αν βλέπουμε την πράξη ή όχι.
- AlcoholEvent : σκηνές που κάποιος εμφανίζεται να πίνει κάτι αλκοολούχο.

- DrunkEvent : σκηνές που εμφανίζεται κάποιος να ενεργεί υπό την επήρεια αλκοόλ.
- AlcoholMention : σκηνές που αναφέρεται κάτι σχετικό με αλκοόλ, ανεξάρτητα με το αν εμφανίζεται στην σκηνή ή όχι.
- DrugEvent : σκηνές που εμφανίζονται ναρκωτικά. Είναι πιθανόν να γίνεται και χρήση.
    - SoftDrugEvent : το ναρκωτικό που εμφανίζεται ή χρησιμοποιείται είναι ινδική κάνναβη.
    - HardDrugEvent : το ναρκωτικό που εμφανίζεται ή χρησιμοποιείται είναι ηρωίνη ή κοκαΐνη.
- DrugMention : σκηνές που αναφέρεται κάτι σχετικό με ναρκωτικά, ανεξάρτητα με το αν εμφανίζεται στην σκηνή ή όχι.

4

Δημιουργία διανυσμάτων χαρακτηριστικών

Οι σκηνές των ταινιών θα κατηγοριοποιηθούν στις κλάσεις της οντολογίας που φτιάξαμε χρησιμοποιώντας διανύσματα χαρακτηριστικών. Τα χαρακτηριστικά αυτά που θα αποτελούν τα διανύσματα θα αντιστοιχούν με συγκεκριμένες λέξεις. Οι λέξεις αυτές θα επιλεγούν ανάλογα με το αν η εμφάνιση τους σε κάποια σκηνή υποδεικνύει έντονα ότι η σκηνή αυτή ανήκει σε κάποια συγκεκριμένη κλάση.

Στη συνέχεια, θα εξετάσουμε πως επιλέχτηκαν αυτές οι λέξεις-κλειδιά, καθώς και τα διανύσματα που προέκυψαν.

4.1 Επιλογή λέξεων-κλειδιών

Για την επιλογή των λέξεων-κλειδιών, εξετάσαμε αρχικά ποιες ήταν οι λέξεις που εμφανίζονταν συχνά για κάθε κατηγορία και στη συνέχεια, ελέγξαμε και επεκτείναμε το λεξιλόγιο αυτό, χρησιμοποιώντας την λεξιλογική βάση Wordnet.

4.1.1 Συχνά χρησιμοποιούμενες λέξεις

Για να συγκεντρωθούν οι συχνά χρησιμοποιούμενες λέξεις, χρησιμοποιήθηκε πάλι το δείγμα των 250 ταινιών. Για κάθε ένα από τους 4 τομείς του οδηγού γονέων, συγκεντρώθηκε το σύνολο των λέξεων που εμφανίζονται σε όλες τις περιγραφές των

250 ταινιών αυτών και έγινε εξαγωγή των πιο συχνά χρησιμοποιημένων.

Από τις λέξεις αυτές αφαιρέθηκαν εκείνες που είχαν πολλαπλές εμφανίσεις γιατί είναι κοινές εν γένει ή κοινές στην διαδικασία περιγραφής σκηνών, αφήνοντας τις υπόλοιπες που είναι υποψήφιες λέξεις-κλειδιά. Παρακάτω παραθέτονται αυτές μαζί με τον αριθμό των εμφανίσεων τους στο δείγμα μας. Το κατώτερο όριο ήταν οι 5 εμφανίσεις για όλες τις κατηγορίες, εκτός της κατηγορίας *Profanity*, που το όριο κατέβηκε στο 2, λόγω των σύντομων και περιφραστικών περιγραφών της κατηγορίας αυτής.

1. Sex & Nudity

adultery::6	buttock::66	lover::8	panties::16	shirtless::13
affair::11	chest::11	low-cut::6	penis::21	shower::20
bare::98	cleavage::25	masturbate::14	porn::6	skimpy::6
bare-chest::7	crotch::7	moan::10	prostitute::34	strip::14
bikini::9	explicit::23	naked::39	pubic::17	testicle::6
boxer::6	frontal::13	nipple::25	rape::45	topless::20
bra::15	genital::27	non-explicit::6	seduction::6	underwear::20
breast::117	innuendo::7	non-sexual::15	sensual::8	undress::8
brothel::8	intercourse::6	nude::68	sex::157	vagina::6
butt::27	kiss::116	orgasm::6	sexual::105	

Πίνακας 4.1: Υποψήφιες λέξεις-κλειδιά της κατηγορίας Sex & Nudity

2. Gore & Violence

abuse::9	cavalry::6	grenade::8	murder::31	spurt::29
amputate::6	choke::21	fight::107	mutilate::7	stomach::25
arrow::22	combat::16	fire::66	nuclear::9	strangle::17
attack::58	corpse::28	fist::10	ooze::11	strike::9
ax::9	crash::32	fistfight::8	pain::38	struck::11
bash::10	crime::6	flame::12	pistol::12	struggle::24
battle::50	criminal::11	flesh::13	punch::89	suicide::29
beaten::41	crush::16	groin::7	push::24	survive::11
bite::36	cut::113	gun::89	rifle::11	sword::43
blade::9	dead::130	gunfight::7	shoot::143	tank::10
bleed::25	decapitate::14	gunfire::7	shooter::6	throw::41
blood::526	destroy::13	gunshot::14	shootout::9	thrown::27
bloody::219	drown::6	handgun::7	shot::329	torture::30
bloodless::15	duel::6	hang::27	shotgun::13	vicious::6
bloodlessly::6	enemy::9	hit::114	shove::13	violence::129
bomb::27	execute::12	impale::10	slap::31	violent::67
bone::12	explode::28	injure::26	slash::19	war::24
broken::21	explosion::30	injury::20	slice::9	weapon::16
bruise::29	gang::12	interrogate::7	slit::12	whip::17
brutal::18	gangster::8	kill::169	smash::12	wound::138
bullet::62	gore::28	killer::8	smoke::11	
bully::8	gory::11	knife::40	sniper::9	
burn::45	graphic::103	martial::9	splatter::31	

Πίνακας 4.2: Υποψήφιες λέξεις-κλειδιά της κατηγορίας Violence & Gore

3. Profanity

ass::8	damn::17	goddamn::9	faggot::2	sexual::5
bastard::9	dick::4	hell::14	nigger::2	shit, sh*t::22
bitch, b*tch::20	f-word::3	insult::3	profanity::7	slur::3
Christ::8	fuck, f**k::17	Jesus::13	racial::2	swear::4
crap::2	God::12	tit::2	religious::3	

Πίνακας 4.3: Υποψήφιες λέξεις-κλειδιά της κατηγορίας Profanity

4. Alcohol/Drugs/Smoking

alcohol::79	cigar::25	drunken::7	marijuana::22	tobacco::10
bar::33	cigarette::57	heroin::10	pill::12	whiskey::11
beer::31	cocain::25	inject::8	pipe::15	wine::24
beverage::11	drink::224	intoxicant::7	rum::6	
bottle::19	drug::47	joint::7	smoke::209	
champagne::7	drunk::52	liquor::6	snort::11	

Πίνακας 4.4: Υποψήφιες λέξεις-κλειδιά της κατηγορίας Alcohol/Drugs/Smoking

4.1.2 Λειτουργίες Wordnet

Το Wordnet είναι μια βάση δεδομένων λεξιλογίου της αγγλικής γλώσσας.[11] Οι λέξεις που περιλαμβάνει είναι οργανωμένες σε σύνολα συνωνύμων, τα οποία ονομάζονται *synsets*. Κάθε ένα από αυτά τα σύνολα περιέχει τα συνώνυμα τα οποία

το απαρτίζουν, έναν ορισμό της έννοιας που εκφράζει και σε αρκετές περιπτώσεις κάποιες προτάσεις-παραδείγματα της χρήσης των λέξεων. Επίσης, έχουν την δυνατότητα να συνδέονται με άλλα τέτοια σύνολα μέσω ορισμένων σημασιολογικών και λεξιλογικών σχέσεων, που θα εξετάσουμε πιο αναλυτικά στη συνέχεια. Σε περίπτωση που κάποια λέξη έχει περισσότερες από μια ερμηνείες, εμφανίζεται σε όλα τα πιθανά *synsets*, τα οποία όμως κατατάσσονται με βάση την λέξη αυτή σύμφωνα με την συχνότερη ερμηνεία της.

Οι σημασιολογικές σχέσεις που συνδέουν τα *synsets* διαφέρουν ανάλογα με το μέρος του λόγου των συνώνυμων λέξεων που τα αποτελούν. Για τα ουσιαστικά και τα ρήματα, τα *synsets* οργανώνονται σε ιεραρχίες με βάση την σχέση *hypernym/hyponym* και συνδέονται αναλόγως και με άλλες σχέσεις. Στην περίπτωση των επιθέτων, η οργάνωση των *synsets* γίνεται με βάση την σχέση αντιθέτων μεταξύ τους, εκτός από ορισμένα που δεν έχουν αντίθετες έννοιες. Το ίδιο συμβαίνει και με τα επιρρήματα.

Για όλα τα *synsets* έχουμε τις εξής σημασιολογικές σχέσεις[12] :

- *domain category* : χρησιμοποιείται ως ετικέτα για τα *synsets* που η έννοια τους είναι άμεσα συνδεδεμένη με κάποια συγκεκριμένο πεδίο αναφοράς. Ο όρος *domain term category* χρησιμοποιείται για τα *synsets* που η έννοια τους είναι το πεδίο στο οποίο υπάγονται με την παραπάνω σχέση άλλα *synsets*.

- *domain region* : χρησιμοποιείται ως ετικέτα για τα *synsets* που η έννοια τους είναι άμεσα συνδεδεμένη με κάποια συγκεκριμένη περιοχή. Ο όρος *domain term region* χρησιμοποιείται για τα *synsets* που η έννοια τους είναι η περιοχή στην οποία υπάγονται με την παραπάνω σχέση άλλα *synsets*.

- *domain usage* : χρησιμοποιείται ως ετικέτα για τα *synsets* που η έννοια τους είναι άμεσα συνδεδεμένη με τον τρόπο με τον οποίο χρησιμοποιούνται. Ο όρος *domain term usage* χρησιμοποιείται για τα *synsets* που η έννοια τους

είναι ο τρόπος χρήσης στον οποία υπάγονται με την παραπάνω σχέση άλλα *synsets*.

Για τα *synsets* ουσιαστικών και ρημάτων έχουμε τις εξής σημασιολογικές σχέσεις :

- *hypernym* : χρησιμοποιείται για να ορίσει μια κατηγορία ειδικών περιπτώσεων. Το Υ είναι *hypernym* του Χ εάν το Χ είναι ένα είδος του Υ. Είναι μεταβατική ιδιότητα, επομένως το κάθε *synset* έχει το *direct hypernym* που αναφέρεται στην άμεση κατηγοριοποίηση του και το *inherited hypernym* που αναφέρεται σε όλες τις κατηγοριοποιήσεις που έχει κληρονομήσει. Η αντίστροφη σχέση του είναι το *hyponym.*

- *sister term* : χρησιμοποιείται για να ορίσει τα *synsets* που έχουν το ίδιο *direct hypernym*.

- *hyponym* : χρησιμοποιείται για να ορίσει ένα μέλος μιας κατηγορίας. Το Χ είναι *hyponym* του Υ εάν το Χ είναι ένα είδος του Υ. Είναι μεταβατική ιδιότητα, επομένως το κάθε *synset* έχει το *direct hyponym* που αναφέρεται στα *synsets* εκείνα που υπάγονται άμεσα στην κατηγορία του και το *full hyponym* που αναφέρεται σε όλα τα πιθανά *synsets* που ανήκουν στην κατηγορία του. Η αντίστροφη σχέση του είναι το *hypernym.*

Για τα *synsets* ουσιαστικών και επιθέτων έχουμε την παρακάτω σημασιολογική σχέση:

- *attribute* : χρησιμοποιείται για να ορίσει τα ουσιαστικά και τα επίθετα, όπου για το ουσιαστικό τα αντίστοιχα επίθετα παίρνουν διάφορες τιμές.

Μόνο για τα *synsets* ουσιαστικών έχουμε τις εξής σημασιολογικές σχέσεις :

- *meronym* : χρησιμοποιείται για να ορίσει ένα συστατικό ή ένα μέρος κάποιου *synset.* Το Χ είναι *meronym* του Υ εάν το Χ κομμάτι του Υ.

- *holonym* : χρησιμοποιείται για να ορίσει το σύνολο, τα μέρη του οποίου έχουν

οριστεί ως *meronym.* Το Υ είναι *holonym* του Χ εάν το Χ κομμάτι του Υ.

Μόνο για τα *synsets* ρημάτων έχουμε τις εξής σημασιολογικές σχέσεις :

- *troponym* : χρησιμοποιείται για να ορίσει έναν συγκεκριμένο τρόπο εκπόνησης κάποιου άλλου *synset.* Το Χ είναι *troponym* του Υ, εάν το να γίνει το Χ συνεπάγεται να γίνει το Υ με κάποιο τρόπο.
- *entailment* : χρησιμοποιείται για να ορίσει μια εξάρτηση μεταξύ *synsets*. Το Χ είναι *entailment* του Υ, εάν το Χ δεν μπορεί να γίνει, χωρίς να γίνεται ή να έχει ήδη γίνει το Υ.
- *cause* : χρησιμοποιείται για το διαφορετικό *synset* του ίδιου ρήματος, η έννοια του οποίου είναι η παθητική φωνή του ρήματος αυτού.
- *verb group* : χρησιμοποιείται για να ορίσει ένα σύνολο *synsets* ρημάτων με παρόμοιες έννοιες.

Εκτός από τις σημασιολογικές σχέσεις, το Wordnet συνδέει τα *synsets* και με λεξιλογικές σχέσεις :

- *antonym* : χρησιμοποιείται για το αντίθετο *synset*. Η σχέση αυτή υπάρχει για όλα τα *synsets,* αν και εμφανίζεται πολύ συχνά στα *synsets* επιθέτων.
- *pertainym* : χρησιμοποιείται για να υποδείξει το ουσιαστικό στο οποίο αναφέρεται το τρέχον επίθετο.
- *participle of verb* : στην περίπτωση μετοχών που χρησιμοποιούνται ως επίθετα υποδεικνύει το ρήμα από το οποίο παράγεται η μετοχή.
- *see also*
- *similar to*
- *derivationally related form* : χρησιμοποιείται για λέξεις διαφορετικών

συντακτικών κατηγοριών, που έχουν την ίδια ρίζα και έχουν σημασιολογικές ομοιότητες.

Από τις σχέσεις που περιγράφηκαν παραπάνω, η διαδικασία της επέκτασης θα μπορούσε να γίνει αυτόματα, ξεκινώντας από κάποιες βασικές έννοιες και ακολουθώντας τις σχέσεις που τις συνέδεαν με συγγενικές έννοιες. Όμως, λόγω των πολλαπλών ερμηνειών που μπορεί να υπάρχουν για κάθε λέξη, έγινε χειροκίνητα, έτσι ώστε να υπάρχει έλεγχος για την ερμηνεία των υποψηφίων λέξεων.

Συμπεριλήφθηκαν λέξεις που έχουν είτε ως μοναδική είτε ως πρώτη την όποια έννοια ενδιαφέροντος. Οι μόνες περιπτώσεις που παραβιάστηκε ο κανόνας αυτός ήταν σε περιπτώσεις όπου η έννοια ενδιαφέροντος δεν ήταν μεν η πρώτη, αλλά η λέξη έχει ιδιαίτερη σύνδεση με κάποια άλλη έννοια που θα ήταν προτιμότερο να συμπεριλάβουμε πιθανά λάθη από το να την αγνοήσουμε.

Ένα χαρακτηριστικό παράδειγμα είναι η λέξη *weed*, που χρησιμοποιείται ως διαφορετική ονομασία της μαριχουάνας. Η συγκεκριμένη έννοια είναι η τρίτη έννοια που εμφανίζεται στο Wordnet, αλλά ανάμεσα στα συνώνυμα της μαριχουάνας είναι ίσως το πιο συχνά χρησιμοποιούμενο οπότε δε θα μπορούσαμε να το αγνοήσουμε.

Επίσης, το Wordnet περιλαμβάνει εκτός από λέξεις και φράσεις. Καθώς εξετάζουμε το κείμενο μας λέξη προς λέξη, οι φράσεις δεν ήταν δυνατό να συμπεριληφθούν στο λεξιλόγιο μας. Σε αρκετές περιπτώσεις όμως οι φράσεις αυτές περιλάμβαναν κάποια λέξη με συναφή έννοια, οπότε συμπεριλήφθηκε η λέξη αυτή. (π.χ. *sweet Fanny Adams-fanny*).

Για να επεκταθεί το λεξιλόγιο, χρησιμοποιήσαμε κάποιες από τις παραπάνω σχέσεις των λέξεων του λεξιλογίου για να βρούμε και άλλες με συγγενικά νοήματα. Είναι εμφανές ότι όλες οι σχέσεις δεν είναι κατάλληλες για την συγκεκριμένη δουλειά. Επίσης, αν και υπάρχει η δυνατότητα, φυσικά ανάλογα με το μέρος του λόγου, δεν έχει απαραίτητα το κάθε *synset* όλες αυτές τις σχέσεις.

Η συχνότερα εμφανιζόμενη σχέση είναι η *hypernym/hyponym*. Καθώς με αυτόν τον τρόπο είναι οργανωμένο το Wordnet, κάθε *synset* εμφανίζει τουλάχιστον μία από τις δύο σχέσεις, αν είναι στην κορυφή ή στην βάση της ταξονομίας, ή και τις δύο, σε οποιαδήποτε άλλη περίπτωση.

Οι σχέσεις αυτές μαζί με την *sister terms*, που πηγάζει άμεσα από την *hypernym*, ήταν τα βασικά μας εργαλεία στην διαδικασία της επέκτασης.

Τέλος, μία ιδιαιτέρως χρήσιμη σχέση ήταν η *domain usage.* Η συγκεκριμένη σχέση είναι αρκετά σπάνια, καθώς τα πεδία της είναι μόλις 29, αλλά ορισμένα που περιέγραφαν σχεδόν απόλυτα κάποιους από τους τρόπους χρήσης που μας απασχολούσαν. Τα πεδία αυτά, που όλα εμπίπτουν στο πλαίσιο των ύβρεων, ήταν τα εξής :

- **ethnic slur** : περιλαμβάνει λέξεις που χρησιμοποιούνται προσβλητικά ως προς την φυλή ή την γλώσσα κάποιου. Στην περίπτωση αυτή, καθώς έχουμε ένα πολύ συγκεκριμένο πλαίσιο, όλες οι λέξεις συμπεριλήφθηκαν στο λεξιλόγιο μας και πιο συγκεκριμένα ως λέξεις-κλειδιά της κατηγορίας RacialInsult, με την οποία αντιστοιχεί απόλυτα.

- **obscenity** : περιλαμβάνει προσβλητικές ή άπρεπες λέξεις ή φράσεις. Εδώ το πεδίο είναι πιο ευρύ, οπότε οι λέξεις περιγράφουν διάφορες κατηγορίες (SexualReference, AnatomicalReference, ExcrementReference κλπ), αλλά επίσης χρησιμοποιήθηκαν όλες.

- **disparagement :** περιλαμβάνει υποτιμητικές λέξεις ή φράσεις. Προφανώς όλες οι λέξεις που περιλαμβάνονται εδώ έχουν υποτιμητικό χαρακτήρα, αλλά σε διαφορετικό βαθμό. Κατ' επέκταση δεν είναι όλες το ίδιο, ή και σχεδόν καθόλου, προσβλητικές οπότε κάποιες δεν χρησιμοποιήθηκαν.

4.2 Διανύσματα χαρακτηριστικών για τις σκηνές του σεναρίου

Έχοντας βρει τις λέξεις-κλειδιά που αντιστοιχούν σε κάθε κλάση, το διάνυσμα χαρακτηριστικών θα συγκέντρωνε το σύνολο των λέξεων αυτών σε ένα διάνυσμα, όπου κάθε λέξη θα αντιστοιχούσε με μία θέση του διανύσματος. Το διάνυσμα αυτό θα συμπληρωνόταν με 1, εάν η λέξη εμφανιζόταν στην σκηνή, και με 0 σε αντίθετη περίπτωση.

Λόγω όμως των ειδικών συνθηκών του σεναρίου, χρειάστηκε να δημιουργηθούν δύο διανύσματα χαρακτηριστικών για κάθε σκηνή σεναρίου. Αυτό συνέβη διότι για τις περισσότερες από τις λέξεις-κλειδιά που χρησιμοποιήθηκαν, έχει τεράστια διαφορά η θέση στην οποία εμφανίζονται στην σκηνή.

Για παράδειγμα, η λέξη-κλειδί της κλάσης *Death*, *kill,* δεν έχει την ίδια σημασία εάν εμφανίζεται σε κάποιον διάλογο. Στον διάλογο, μπορεί να είναι μία απειλή που πραγματοποιείται ή και όχι στην ίδια σκηνή, ή ακόμα και να εμφανίζεται χάριν αστεϊσμού. Στην περίπτωση αυτή, δεν αποτελεί ενδεικτική λέξη μιας σκηνής θανάτου, οπότε δεν μπορεί να υπολογιστεί ως λέξη-κλειδί.

Αντίθετα, υπάρχουν περιπτώσεις, όπως στην περίπτωση των κοσμητικών επιθέτων, που μας ενδιαφέρει η εμφάνιση τους στον διάλογο μόνο. Αν και οι πιθανότητες να εμφανιστούν στην περιγραφή, που είναι λιτή και απλά απεικονίζει τα τεκταινόμενα, χωρίς σχόλια, είναι ελάχιστες, δεν υπάρχει ανάγκη να ελέγξουμε την περιγραφή για εμφάνιση των συγκεκριμένων λέξεων. Τ

Τέλος, υπάρχει η πιθανότητα η "θέση" της λέξης να μεταβάλει την κατηγορία στην οποία ανήκε η σκηνή : π.χ. λέξεις σεξουαλικού περιεχομένου στην περιγραφή υποδεικνύουν ότι γίνονται οι αντίστοιχες πράξεις, οπότε πρέπει να κατηγοριοποιηθεί σε κάποια υποκλάση του SexNudityEvent, ενώ αν εμφανίζονται στον διάλογο η σκηνή ανήκει στην κλάση SexualReference.

Επομένως, οι κλάσεις της οντολογίας διαχωρίστηκαν σε δύο βασικές κατηγορίες ανάλογα με το που στο σενάριο θα περιγράφεται το γεγονός ενδιαφέροντος, στον διάλογο ή στην περιγραφή, και επομένως ποια κομμάτια του σεναρίου θα εξεταστούν για τις αντίστοιχες λέξεις-κλειδιά.

Σύμφωνα με αυτόν τον διαχωρισμό, οι κλάσεις διαλόγου είναι οι παρακάτω :

- RacialInsult
- HomophobicInsult
- GenderedInsult
- HeavyInsult
- CasualInsult
- ExcrementReference
- AnatomicalReference
- DrugMention
- SexualReference
- ReligiousReference
- Exclamation
- ReligiousExclamation
- F-word
- SmokingMention
- AlcoholMention

Αντίστοιχα, στο κομμάτι της περιγραφής αντιστοιχούν οι εξής κλάσεις :

- KissingEvent
- OralSexEvent
- MasturbationEvent
- RapeEvent
- NudityEvent
- PartialNudityEvent
- FullNudityEvent
- WeaponRelatedEvent
- MassiveWeaponryEvent
- BladeEvent
- MedievalWeaponry
- GunRelated
- ExplosionEvent
- FireEvent
- PhysicalViolence
- GoreEvent
- InjuryEvent
- DeadBodyEvent
- SmokingEvent
- AlcoholEvent
- DrunkEvent
- DrugEvent
- SoftDrugEvent
- HardDrugEvent
- Death
- Electrocution
- Asphyxiation

Συγκεντρώνοντας όλες τις λέξεις-κλειδιά που υποδεικνύουν αυτές τις κλάσεις, δημιουργούμε ένα διάνυσμα χαρακτηριστικών για κάθε μία από τις δύο αυτές κατηγορίες.

Για να περιορισθεί το μέγεθος του διανύσματος, οι συνώνυμες λέξεις, λέξεις που συμπεριλαμβάνονταν στο ίδιο *synset* του Wordnet, τοποθετήθηκαν στην ίδια θέση του διανύσματος.

Στη συνέχεια, για εξοικονόμηση χώρου, αντί να παραθέσουμε ολόκληρα τα διανύσματα, θα παρουσιάσουμε τις κλάσεις και μόνο τις λέξεις-κλειδιά που αντιστοιχούν στην κλάση αυτή, τις θέσεις δηλαδή που θα έχουν 1 στο χαρακτηριστικό διάνυσμα της κλάσης. Όλες οι υπόλοιπες λέξεις-κλειδιά των άλλων κλάσεων που εμπίπτουν στην ίδια κατηγορία (περιγραφής ή διαλόγου) συμπληρώνονται με 0 στο χαρακτηριστικό διάνυσμα.

-Διάνυσμα διαλόγου

HomophobicInsult	**GenderedInsult**	**ExcrementReference**
fagot, faggot, fag, nance, queer, poof, poove	floozy, floozie, slattern	bullshit, horseshit, shit, shite, dogshit
dyke, dike, butch	bitch	piss, pee
	whore, harlot, bawd, hooker	crap, poop, turd
	slut, trollop, strumpet, hussy	

CasualNameCalling	**AlcoholMention**
dork, jerk	drunk
airhead	inebriated
dimwit, nitwit, doofus	tipsy
dummy	besotted, fuddled, slopped, sloshed
fool, chump, gull, patsy, sucker	intoxicated
fathead, goof, goofball, bozo, jackass	alcohol
idiot, imbecile, cretin, moron, half-wit, retard	aperitif
stupid	beer, lager, stout, ale
dunce, dunderhead, blockhead, bonehead, knucklehead, shithead, dumbass, fuckhead	mead
twerp	wine, champagne, sherry
	brandy
	gin
	rum
	tequila
	vodka
	whiskey, whisky, bourbon, firewater
	liqueur, liquor
	absinth, absinthe
	highball
	cocktail
	bar, saloon, speakeasy, pub

SmokingMention	DrugMention	RacialInsult
smoke	reefer, spliff, joint	coolie, cooly
cigarette, cigaret	pot, dope, weed	kike, hymie, sheeny, yid
cigar, cigarillo	cannabis, marijuana, hashish, hasheesh, marihuana	chink, chinaman
pipe	cocaine	Paddy, Mick
hookah, nargileh, narghile	heroin, skag, scag	wop, dago, ginzo, greaseball, guinea
tobacco	drug	spic, spik, spick, greaser, wetback
ashtray	LSD	pickaninny, piccaninny, picaninny, nigger
	methamphetamine, meth	Redskin, Injun
	opium	Jap
	methadone	Kraut, Krauthead, Boche
	codeine	wog
	morphine	gypo, gypsy, gipsy
	amphetamine	

Exclamation	ReligiousExclamation	ReligiousReference	F-word
damn, darn	Goddamn, goddam	God	fuck
Hell		Jesus, Christ	

AnatomicalReference	**SexualReference**	**HeavyInsult**
dick, cock, prick, pecker	screw, shag	bastard, dickhead
ass, arse	sex, sexual	asshole
pussy, cunt, puss, twat	sodomy, anal	motherfucker, fucker
tits, boobs, titties	oral, blowjob	cocksucker
	handjob	
	masturbate, wank, jerk-off	
	cum	
	orgasm	

Πίνακες 4.5: Διανύσματα χαρακτηριστικών διαλόγου

-Διάνυσμα περιγραφής

FullNudityEvent	**InjuryEvent**	**PartialNudityEvent**
naked, nude, bare	stab	scantily, skimpily
frontal	injury, trauma	shirtless
topless, braless, bare-breasted, bare-chested	nosebleed	chest
breasts	beaten	bra, panties, underwear, bikini
butt, buttocks	bruise, shiner	cleavage,low-cut, skimpy
nipples	dislocation	undress, strip, unclothe, disrobe
genitals, vagina	frostbite	
penis, crotch, testicle	sunburn, scorch, singe, scald	
pubic	blister	
nudity	wound	
	abrasion, scratch, graze, excoriation	
	gash, slash, slice	

Electrocution	**FireEvent**	**ExplosionEvent**
electrocute	fire, flame	explosion, explode, detonation, blowup
	burn	

SmokingEvent	**AlcoholEvent**	**DrunkEvent**	**DrugEvent**
smoke	alcohol	drunk, drunken	drug
cigarette, cigaret	aperitif	inebriated	LSD
cigar, cigarillo	beer, lager, stout, ale	tipsy	methamphetamine, meth
pipe	mead	besotted, fuddled, slopped, sloshed	opium
hookah, nargileh, narghile	wine, champagne, sherry	intoxicated	methadone
tobacco	brandy		codeine
ashtray	gin		morphine
	rum		amphetamine
	tequila		
	vodka		
	whiskey, whisky, bourbon, firewater		
	liqueur, liquor		
	absinth, absinthe		
	Highball, cocktail		

MassiveWeaponryEvent	MedievalWeaponryEvent	GunRelatedEvent
flamethrower	bow, crossbow, longbow	bullet, slug, pellet
cannon	arrow	ammunition, ammo
mortar	pike	cartridge, canister
launcher, bazooka	spear, lance, javelin, trident	gun, gunfight
torpedo	tomahawk, hatchet	shot, shoot, gunshot, gunfire
missile	ax, broadax, battle-ax, poleax	barrel, muzzle, trigger
bomb	catapult, trebuchet	firearm, pistol, handgun
grenade	sword, blade	submachine, Kalashnikov, Uzi
tank, panzer	backsword, broadsword	semiautomatic, M-1, Luger, Garand
Ack-ack	saber, sabre	musket
	rapier	shotgun, scattergun
		rifle, carbine, Dragunov, Winchester
		derringer, forty-five
		revolver, six-gun, six-shooter, Colt
		sniper

Death	**BladeEvent**	**Asphyxiation**	**PhysicalViolence**
kill	knife	drown	punch, biff, pummel, pommel
die, decease, perish	pocketknife, switchblade	smother, asphyxiate, suffocate, stifle	kick
murder, slay	bayonet	strangle, strangulate, throttle	shove, push
assassinate	dagger, poniard, stiletto	choke	slap, smack, thwack
execute	shiv		whack, wham, whop, wallop
behead, decapitate	razor		smite
lynch, slaughter, massacre			beat, abuse, bash
suicide			fight, fistfight

SoftDrugEvent	**HardDrugEvent**
reefer, spliff, joint	cocaine
pot, dope, weed	heroin, skag, scag
cannabis, marijuana, hasheesh, hashish marihuana	

OralSexEvent	**MasturbationEvent**	**RapeEvent**
fellate, fellatio	masturbate	rape, ravish
blowjob, oral	wank, jerk-off	
cunnilingus, cunnilinctus		

GoreEvent	**DeadBodyEvent**	**KissingEvent**	**SexEvent**
torture	dead, deceased	kiss, snog	sex, sexual
amputate, mutilate	corpse, cadaver	smacker, smooch	intercourse, coitus, coition, copulation
blood, bleed		peck	screw, shag
flesh, bone		lips	sodomy, anal
intestine, gut			handjob
ooze			cum
gore			brothel, prostitute, whore
			porn, pornstar
			orgasm

Πίνακες 4.6: Διανύσματα χαρακτηριστικών περιγραφής

4.3 Διανύσματα χαρακτηριστικών για τις σκηνές όπως προκύπτουν από τις περιγραφές

Τα διανύσματα χαρακτηριστικών των περιγραφών θα προκύψουν ακολουθώντας παρόμοια βήματα.

Στις περιγραφές, δεν υπάρχει ο διαχωρισμός μεταξύ διανυσμάτων διαλόγου - περιγραφής. Υπάρχει όμως διαφορετικό διάνυσμα ανάλογα με το κομμάτι του οδηγού στο οποίο εμφανίζεται η πρόταση. Καθώς οι κλάσεις της οντολογίας δημιουργήθηκαν με βάση τον οδηγό, κάθε κλάση εμπίπτει πολύ συγκεκριμένα σε μία από τις 4 κατηγορίες του οδηγού. Οπότε, μπορούμε εύκολα να τις διαχωρίσουμε με βάση τις κατηγορίες αυτές, μειώνοντας έτσι το μέγεθος του εκάστοτε διανύσματος χαρακτηριστικών και τον χρόνο της όλης διαδικασίας.

Επίσης, αν και οι περισσότερες λέξεις - κλειδιά των κλάσεων παραμένουν ίδιες, έχουν προστεθεί κάποιες που μόνο η εμφάνιση τους στο συγκεκριμένο πλαίσιο του οδηγού τις κάνει τέτοιες. Τέτοιες περιπτώσεις είναι διάφορες παραφράσεις ύβρεων (*f**k, f-*word), γενικές περιγραφικές λέξεις (*violence, curse*), όπως και λέξεις οι οποίες δεν θα μπορούσαμε να θεωρήσουμε ασφαλώς ως λέξεις-κλειδιά σε ένα ευρύτερο πλαίσιο, αλλά εδώ υποδεικνύουν κάτι συγκεκριμένο (*drink*).

-Sex & Nudity

MasturbationEvent	**KissingEvent**	**RapeEvent**	**OralSexEvent**
masturbate	kiss, snog	rape, ravish	fellate, fellatio
wank, jerk-off	smacker, smooch		blowjob, oral
	peck		cunnilingus, cunnilinctus
	lips		

SexualReference	FullNudityEvent	PartialNudityEvent	SexEvent
masturbate	naked, nude, bare	scantily, skimpily	sex, sexual
wank, jerk-off	frontal	shirtless	intercourse, coitus, coition, copulation
sex, sexual	topless, braless, bare-breasted, bare-chested	chest	sodomy, anal
fellate, fellatio	breasts	bra, panties, underwear, bikini	handjob
blowjob, oral	butt, buttocks	cleavage, low-cut, skimpy	orgasm
cunnilingus, cunnilinctus	nipples	undress, strip, unclothe, disrobe	brothel, prostitute, whore
genitals, vagina	genitals, vagina		porn, pornstar
penis, crotch, testicle	penis, crotch, testicle		
screw, shag	pubic		
sodomy, anal	nudity		
handjob			
cum			
orgasm			
brothel, prostitute, whore			
porn, pornstar			
dialogue, discussion			
say, talk, tell			

Πίνακες 4.7: Διανύσματα χαρακτηριστικών των περιγραφών για την κατηγορία Sex&Nudity

-Violence & Gore

MassiveWeaponryEvent	MedievalWeaponryEvent	GunRelatedEvent
flamethrower	bow, crossbow, longbow	bullet, slug, pellet
cannon	arrow	ammunition, ammo
mortar	pike	cartridge, canister
launcher, bazooka	spear, lance, javelin, trident	gun, gunfight
torpedo	tomahawk, hatchet	shot, shoot, gunshot, gunfire
missile	ax, broadax, battle-ax, poleax	barrel, muzzle, trigger
bomb	catapult, trebuchet	firearm, pistol, handgun
grenade	sword, blade	submachine, Kalashnikov, Uzi
tank, panzer	backsword, broadsword	semiautomatic, M-1, Luger, Garand
Ack-ack	saber, sabre	musket
	rapier	shotgun, scattergun
		rifle, carbine, Dragunov, Winchester
		derringer, forty-five
		revolver, six-gun, six-shooter, Colt
		sniper

Death	**PhysicalViolence**	**InjuryEvent**	**BladeEvent**	**GoreEvent**
kill	punch, biff, pummel, pommel	stab	knife	torture
die, decease, perish	kick	injury, trauma	pocketknife, switchblade	amputate, mutilate
murder, slay	shove, push	nosebleed	bayonet	blood, bleed
assassinate	slap, smack, thwack	beaten	dagger, poniard, stiletto	flesh, bone
execute	whack, wham, whop, wallop	bruise, shiner	shiv	intestine, gut
behead, decapitate	smite	dislocation	razor	ooze
lynch, slaughter, massacre	beat, abuse, bash	frostbite		gore
suicide	fight, fistfight	sunburn, scorch, singe, scald		
		blister		
		wound		
		abrasion, scratch, graze, excoriation		
		gash, slash, slice		
		blood, bleed		

ExplosionEvent	Electrocution	Asphyxiation	FireEvent	DeadBodyEvent
explosion, explode, detonation, blowup	electrocute	drown	fire, flame	dead, deceased
		smother, asphyxiate, suffocate, stifle	burn	corpse, cadaver
		strangle, strangulate, throttle		
		choke		

Πίνακες 4.8: Διανύσματα χαρακτηριστικών των περιγραφών για την κατηγορία Violence&Gore

-Profanity

Exclamation	ReligiousReference	ReligiousExclamation	HomophobicInsult
exclamation	religious	Goddamn, goddam	insult
damn, darn	God		slur, derogatory
Hell	Jesus, Christ		homophobic
profanity, curse, swear			fagot, faggot, fag, nance, queer, poof, poove
			dyke, dike, butch

CasualNameCalling	**RacialInsult**	**SexualReference**	**F-word**
insult	insult	obscene, obscenity, crude	f-word
casual, mild, moderate, tame	slur, derogatory	sexual	fuck, f**k, f*ck
dork, jerk	racial, race	screw, shag	
airhead	coolie, cooly	sex	
dimwit, nitwit, doofus	kike, hymie, sheeny, yid	sodomy, anal	
dummy	chink, chinaman	oral, blowjob	
fool, chump, gull, patsy, sucker	Paddy, Mick	handjob	
fathead, goof, goofball, bozo, jackass	wop, dago, ginzo, greaseball, guinea	masturbate, wank, jerk-off	
idiot, imbecile, cretin, moron, half-wit, retard	spic, spik, spick, greaser, wetback	cum	
stupid	pickaninny, piccaninny, picaninny, nigger, n*gger, n*gg*r	orgasm	
dunce, dunderhead, blockhead, bonehead, knucklehead, shithead, dumbass, fuckhead	Redskin, Injun		
twerp	Jap		
	Kraut, Krauthead, Boche		
	wog		
	gypo, gypsy, gipsy		

HeavyInsult	GenderedInsult	ExcrementReference	AnatomicalReference
insult	insult	scatological	obscene, obscenity, crude
bastard, dickhead	slur, derogatory	bullshit, horseshit, shit, shite, dogshit, sh*t, s**t	anatomical
asshole	floozy, floozie, slattern	piss, pee	dick, cock, prick, pecker
profanity, curse, swear	bitch, b*tch	crap, poop, turd	ass, arse, a**
cocksucker	whore, harlot, bawd, hooker		pussy, cunt, puss, twat
motherfucker, fucker, motherf**ker, f**ker, motherf*cker, f*cker	slut, trollop, strumpet, hussy	profanity, curse, swear	tits, boobs, titties
	profanity, curse, swear		profanity, curse, swear

Πίνακες 4.9: Διανύσματα χαρακτηριστικών των περιγραφών για την κατηγορία Profanity

-Alcohol/Drugs/Smoking

DrugMention	DrugEvent	SmokingMention	SmokingEvent
reefer, spliff, joint	drug	smoke	smoke
pot, dope, weed	LSD	cigarette, cigaret	cigarette, cigaret
cannabis, marijuana, hashish, hasheesh, marihuana	methampheta mine, meth	cigar, cigarillo	cigar, cigarillo
heroin, skag, scag	opium	pipe	pipe
cocaine	methadone	hookah, nargileh, narghile	hookah, nargileh, narghile
drug	codeine	ashtray	ashtray
LSD	morphine		
methamphetamine, meth	amphetamine		
opium			
methadone			
codeine			
morphine			
amphetamine			

AlcoholEvent	**AlcoholMention**
drink	drunk, drunken
alcohol	inebriated
aperitif	tipsy
beer, lager, stout, ale	besotted, fuddled, slopped, sloshed
mead	intoxicated
wine, champagne, sherry	drink
brandy	alcohol
gin	aperitif
rum	beer, lager, stout, ale
tequila	mead
vodka	wine, champagne, sherry
whiskey, whisky, bourbon, firewater	brandy
liqueur, liquor	gin
absinth, absinthe	rum
highball	tequila
cocktail	vodka
bar, saloon, speakeasy, pub	whiskey, whisky, bourbon, firewater
	liqueur, liquor
	absinth, absinthe
	highball
	cocktail
	bar, saloon, speakeasy, pub

DrunkEvent	**SoftDrugEvent**	**HardDrugEvent**
drunk, drunken	reefer, spliff, joint	heroin, skag, scag
inebriated	pot, dope, weed	cocaine
tipsy	cannabis, marijuana, hasheesh, hashish, marihuana	
besotted, fuddled, slopped, sloshed		
intoxicated		

Πίνακες 4.10: Διανύσματα χαρακτηριστικών των περιγραφών για την κατηγορία Alcohol/Drugs/Smoking

5

Επεξεργασία δεδομένων

Στο κεφάλαιο αυτό θα εξετάσουμε την διαδικασία με την οποία επεξεργαζόμαστε τα δεδομένα μας, το σενάριο και τις περιγραφές, έτσι ώστε να φτάσουμε στο ζητούμενο αποτέλεσμα, που είναι η τελική αντιστοίχηση μεταξύ τους.

Η διαδικασία που ακολουθήθηκε ήταν η εξής :

- το κείμενο του σεναρίου περνάει από επεξεργασία φυσικής γλώσσας, δημιουργούνται τα δύο διανύσματα χαρακτηριστικών, με βάση τα οποία κατηγοριοποιείται στις κλάσεις της οντολογίας
- οι περιγραφές περνάνε από επεξεργασία φυσικής γλώσσας, δημιουργούνται τα τέσσερα διανύσματα χαρακτηριστικών, με βάση τα οποία κατηγοριοποιούνται στις κλάσεις της οντολογίας
- γίνεται η τελική αντιστοίχηση μεταξύ των σκηνών του σεναρίου και των περιγραφών.

5.1 Επεξεργασία φυσικής γλώσσας

Τα δεδομένα μας και στις δύο περιπτώσεις είναι κείμενα φυσικής γλώσσας και η επεξεργασία η οποία τους γίνεται είναι η ίδια, την οποία θα δούμε αναλυτικά

παρακάτω.

Η μόνη διαφορά είναι ότι στην περίπτωση των περιγραφών το κείμενο επεξεργάζεται κατευθείαν όπως το παίρνουμε από την ιστοσελίδα, ενώ στην περίπτωση του σεναρίου, πρέπει να συγκεντρώσουμε τα διάφορα κομμάτια που υπάγονται στην σκηνή.

Το σενάριο έχει ήδη επεξεργαστεί σε τρίπλες σύμφωνα με την οντολογία του scriptontology. Η δομή που ακολουθείται για κάθε σκηνή είναι η εξής :

- Η κάθε σκηνή συνδέεται με την σχέση hasPart με όλα τα πλάνα που την απαρτίζουν.
- Το κάθε πλάνο έχει την ιδιότητα description η οποία είναι μια πλήρης περιγραφή των τεκταινόμενων σε αυτό το πλάνο : π.χ. *SECURITY GUARD WALKS PAST PARKED LINCOLN TOWN CAR AND EXITS F/G R. EXTREME L3-S LYNN, CARTER AND DOORMAN WALK DOWN STEPS FROM BUILDING R. DOORMAN OPENS CAR DOOR FOR LYNN, THEN RETURNS TO HIS POSITION BY MAIN DOOR. CARTER GETS INTO DRIVER'S SIDE. CAR PULLS AWAY DOWN DRIVE, INTO ROAD AND DRIVES AWAY*
- Υπάρχουν άτομα της οντολογίας με τύπο SpeakingAct τα οποία αντιστοιχούν στα ομιλούντα κομμάτια της ταινίας. Τα άτομα αυτά συνδέονται με την σχέση happensIn με το πλάνο στο οποίο συμβαίνουν.
- Το καθένα από αυτά έχει την ιδιότητα text η οποία είναι μια αυτολεξεί παράθεση του ομιλούντος κομματιού : π.χ. *I'll never understand you, Car. You luxuriate in the company of these...ugh! You know the damage those women...well, no, their husbands*

Το σύνολο λοιπόν της πληροφορίας για μια σκηνή είναι οι περιγραφές (description)

όλων των πλάνων που είναι μέρη της και το κείμενο (text) όλων των ομιλούντων κομματιών που συμβαίνουν στα πλάνα αυτά και από αυτό θα προκύψει το χαρακτηριστικό διάνυσμα της σκηνής αυτής.

Αφού έχουμε συγκεντρώσει τα κείμενα προς επεξεργασία, μπορούμε να περάσουμε στην ίδια την διαδικασία της επεξεργασίας.

Αρχικά, το όποιο κείμενο χωρίζεται στα κομμάτια που το αποτελούν.

Στη συνέχεια, η κάθε λέξη του κειμένου παίρνει μια ετικέτα αναλόγως με τι μέρος του λόγου είναι. Οι ετικέτες αυτές μπορεί να είναι οποιαδήποτε από τις παρακάτω 36 κατηγορίες[13] :

1. Συντονιστικός σύνδεσμος (Coordinating conjunction). Οι σύνδεσμοι αυτοί είναι οι λέξεις *for, and, nor, but, or, yet,so.*

2. Αριθμητικό (Cardinal number).

3. Τροποποιητής (Determiner). Ως τέτοιες χαρακτηρίζονται οι λέξεις που υπάρχουν πριν από κάποιο ουσιαστικό και συγκεκριμενοποιούν το ουσιαστικό αυτό. Οι λέξεις αυτές μπορεί να είναι άρθρα, κτητικές αντωνυμίες, δεικτικές αντωνυμίες ή ποσοδείκτες.

4. Υπαρξιακό *εκεί* (Existential *there*). Η λέξη *there* όταν χρησιμοποιείται μπροστά από ρήμα για να βεβαιώσει ότι κάποιος ή κάτι υπάρχει.

5. Ξένη λέξη (Foreign word).

6. Πρόθεση ή δευτερεύων σύνδεσμος (Preposition or subordinating conjunction)

7. - 9. Επίθετο, στον θετικό, συγκριτικό ή υπερθετικό βαθμό (Adjective, Adjective comparative, Adjective superlative)

10. Δείκτης στοιχείου λίστας (List item marker).

11. Βοηθητικό ρήμα (Modal).

12. - 15. Ουσιαστικό ή κύριο όνομα, στον ενικό ή πληθυντικό αριθμό (Noun singular or mass,Noun plural,Proper noun singular,Proper noun plural).

16. Προσδιοριστικό επίθετο (Predeterminer). Ως τέτοιες χαρακτηρίζονται οι λέξεις που προηγούνται των τροποποιητών. Συνήθως χρησιμοποιείται για να εκφράσει τμήμα του όλου που δηλώνει η υπόλοιπη φράση.

17. Κτητική κατάληξη (Possessive ending).

18. Προσωπική αντωνυμία (Personal pronoun).

19. Κτητική αντωνυμία(Possessive pronoun).

20. - 22. Επίρρημα, στον θετικό, συγκριτικό ή υπερθετικό βαθμό (Adverb, Adverb comparative, Adverb superlative).

23. Μόριο (Particle).

24. Σύμβολο (Symbol).

25. Η πρόθεση *to*.

26. Επιφώνημα (Interjection).

27. - 32. Ρήμα, σε βασική μορφή, παρελθοντικό χρόνο, γερούνδιο, μετοχή αορίστου ή τρίτο ενικό ενεστώτα (Verb base form, Verb past tense, Verb gerund or present participle, Verb past participle, Verb non-3rd person singular present, Verb 3rd person singular present).

33. - 36. Λέξη που αρχίζει από wh-, όπως *when, where, who, what, which* καθώς και η λέξη *how* (Wh-determiner, Wh-pronoun, Possessive wh-pronoun, Wh-adverb). Οι περισσότερες από τις λέξεις αυτές χρησιμοποιούνται ως επιρρήματα, αλλά κάποιες μπορεί να έχουν διαφορετική χρήση από πρόταση σε πρόταση, τροποποιητής και

αντωνυμία για παράδειγμα.

Παρατηρώντας τις κατηγορίες αυτές είναι εμφανές ότι δεν μας προσφέρουν όλες οι λέξεις του κειμένου χρήσιμες πληροφορίες ως προς το περιεχόμενου του. Οι λέξεις που κρατάμε για περαιτέρω επεξεργασία είναι όσες έχουν ετικέτα επιθέτου, ουσιαστικού, επιρρήματος ή ρήματος.

Το επόμενο βήμα της επεξεργασίας είναι η αποκοπή καταλήξεων. Η διαδικασία αυτή αποκόπτει τις διάφορες καταλήξεις των λέξεων, όπως *-ing, -ed* και *-s,* μειώνοντας την λέξη σε μια θεματική ρίζα. Κατ' αυτόν τον τρόπο, διάφορες μορφές της ίδιας λέξης, ουσιαστικά διαφορετικού αριθμού ή ρήματα διαφορετικών χρόνων, προσλαμβάνονται ως η ίδια λέξη. Αντίστοιχα, το ίδιο συμβαίνει για συγγενικά παράγωγα της ίδιας λέξης, που αντιστοιχούνται στο κοινό τους θέμα. Οπότε, όταν θα γίνεται ο έλεγχος για λέξεις-κλειδιά, δεν θα χρειαστεί να συμπεριλάβουμε όλες τις πιθανές μορφές και τα παράγωγα μιας λέξης για να είμαστε σίγουροι ότι όλες οι εμφανίσεις της λέξης-κλειδί θα προσμετρηθούν.

Μία εναλλακτική διαδικασία που θα μπορούσε να ακολουθηθεί ήταν η λημματοποίηση. Στην διαδικασία αυτή η κάθε λέξη αντιστοιχείται με το λήμμα του λεξικού από το οποίο προέρχεται. Η διαδικασία αυτή είναι πιο περίπλοκη και πιο χρονοβόρα από την αποκοπή, αλλά είναι και πιο αποτελεσματική καθώς καθώς καταφέρνει να συμπεριλάβει και περιπτώσεις στις οποίες η αποκοπή αποτυγχάνει. Χαρακτηριστικό παράδειγμα ο συγκριτικός βαθμός του επιθέτου *bad, worse,* που δεν παράγουν το ίδιο απόκομμα, αλλά έχουν το ίδιο λήμμα.

Η ιδιαίτερη μορφή των σεναρίων όμως, της κύριας πηγής κειμένου, δεν κάνει απαραίτητη την χρήση της πιο περίπλοκης αυτής διαδικασίας. Στο σενάριο, οι σκηνές περιγράφονται σαν να συμβαίνουν αυτή την στιγμή και σχετικά επιγραμματικά. Είναι σχετικά σπάνιο να έχουμε χρήση επιθέτων και ακόμα περισσότερο σε οποιονδήποτε βαθμό πλην του θετικού, όπως και το να έχουμε ρηματικές μορφές σε άλλον χρόνο εκτός του ενεστώτα. Αντίστοιχα, οι περιγραφές είναι πολύ λιτές και σύντομες και

επίσης δεν περιέχουν συνήθως περίπλοκους τύπους. Επομένως, δύο από τους βασικότερους λόγους για τους οποίους θα θέλαμε να επιλέξουμε λημματοποίηση, οι ανώμαλοι τύποι επιθέτων και ρημάτων πρακτικά εξαλείφονται.

Επιπλέον, η ίδια αυτή μορφή των σεναρίων κάνει την χρήση της λημματοποίησης δυσκολότερη. Στους περισσότερους αλγορίθμους λημματοποίησης, ένα βασικό στοιχείο είναι η ταυτοποίηση του μέρους του λόγου της λέξης. Στην περίπτωση αυτή όμως, πολλές προτάσεις δεν έχουν την καθιερωμένη σύνταξη, με το πιο σύνηθες πρόβλημα να είναι η παντελής έλλειψη ρήματος από την πρόταση. Συνεπώς, είναι συχνό το φαινόμενο η ετικέτα μέρους του λόγου που ανατίθεται σε μία λέξη να μην είναι η σωστή. Το γεγονός αυτό δεν μας δημιουργεί προβλήματα στην προηγούμενη διαδικασία, καθώς οι λάθος αναθέσεις γίνονται μεταξύ των επιτρεπόμενων ετικετών, οπότε οι λέξεις περνάνε τον έλεγχο έτσι και αλλιώς.

5.2 Κατηγοριοποίηση σκηνών

Για κάθε σκηνή του σεναρίου, αφού το κείμενο περάσει από την επεξεργασία φυσικής γλώσσας που έχουμε αναφέρει προηγουμένως, δημιουργούνται δύο διανύσματα χαρακτηριστικών που αποτελούνται από 0 και 1. Το ένα προκύπτει από όλα τα κομμάτια διαλόγου της σκηνής και με βάση το αντίστοιχο διάνυσμα λέξεων περιέχει 1 για όποια λέξη εμφανίζεται στο σύνολο του διαλόγου και 0 για εκείνες που δεν εμφανίζονται. Αντίστοιχα, το δεύτερο προκύπτει από όλα τα κομμάτια περιγραφών της σκηνής και με βάση το διάνυσμα λέξεων περιγραφής περιέχει 1 για τις λέξεις που εμφανίζονται και 0 για τις υπόλοιπες.

Αντίστοιχα, για κάθε μία από τις υποψήφιες κλάσεις δημιουργείται ένα διάνυσμα χαρακτηριστικών, σύμφωνα με το διάνυσμα διαλόγου ή περιγραφής ανάλογα με το σε ποια κατηγορία ανήκει η κλάση. Το διάνυσμα αυτό συμπληρώνεται με 1 για τις λέξεις-κλειδιά της συγκεκριμένης κλάσης και με 0 για τις υπόλοιπες.

Το διάνυσμα αυτό της κλάσης συγκρίνεται με το αντίστοιχο από τα διανύσματα της

εξεταζόμενης σκηνής και η σκηνή εντάσσεται στην κλάση ή όχι σύμφωνα με το αποτέλεσμα της σύγκρισης αυτής.

Στην περίπτωση των περιγραφών, ένα πρόβλημα που αντιμετωπίσαμε ήταν ποια θα ήταν η μονάδα εξέτασης.

Σε αντίθεση με το σενάριο, όπου ήταν απόλυτα σαφές που ξεκινάει και που τελειώνει μια σκηνή, στις περιγραφές ήταν πολύ πιο ασαφές. Έπρεπε να αποφασιστεί ποια θα ήταν η μονάδα την οποία θα θεωρούσαμε προς εξέταση. Οι μόνες σχετικά τυποποιημένες επιλογές που υπήρχαν ήταν να θεωρήσουμε ως μονάδα κάθε μία από τις 4 κατηγορίες ξεχωριστά ή κάθε πρόταση. Η πρώτη επιλογή περιείχε σίγουρα πάνω από μία σκηνές οπότε η αντιστοίχηση θα ανάγκαζε τον χρήστη να εξετάσει ένα μεγάλο σύνολο σκηνών για να βρει την επιθυμητή σκηνή. Επιπλέον, η υπερβολική πληροφορία πιθανόν να δυσκόλευε την αντιστοίχηση. Η δεύτερη επιλογή επίσης παρουσιάζει προβλήματα καθώς αρκετές φορές σε μία πρόταση περιγράφονται πάνω από μία σκηνές, σχετικού όμως περιεχομένου. Σπανιότερα επίσης η περιγραφή μίας μοναδικής σκηνής επεκτείνεται σε περισσότερες των μία προτάσεων. Γενικά όμως, αυτή είναι η πιο κοντινή αντιστοίχηση με μία σκηνή ταινίας, οπότε ως μονάδα επεξεργασίας για τον οδηγό επιλέχθηκε η πρόταση.

Στη συνέχεια η διαδικασία είναι παρόμοια. Για την κάθε σκηνή της περιγραφής, αφού περάσει από επεξεργασία φυσικής γλώσσας, δημιουργείται ένα διάνυσμα χαρακτηριστικών, ανάλογα με τον τομέα του οδηγού στον οποίον εμφανίζεται η πρόταση. Το διάνυσμα συμπληρώνεται με 0, για τις λέξεις που δεν εμφανίζονται στην πρόταση, και με 1, για αυτές που εμφανίζονται.

Για τις κλάσεις, δημιουργείται πάλι ένα διάνυσμα χαρακτηριστικών, σύμφωνα με το διάνυσμα της κατηγορίας στην οποία ανήκει η κλάση. Το διάνυσμα αυτό συμπληρώνεται με 1 για τις λέξεις-κλειδιά της συγκεκριμένης κλάσης και με 0 για τις υπόλοιπες.

5.3 Αντιστοίχηση περιγραφών με σκηνές σεναρίου

Πλέον έχουμε κατηγοριοποιημένες τις σκηνές του σεναρίου από την μία πλευρά και τις προτάσεις του οδηγού από την άλλη. Κατ' αυτόν τον τρόπο έχει πραγματοποιηθεί η πρώτη αντιστοίχηση μεταξύ των σκηνών και των προτάσεων που υπάγονται στην ίδια κατηγορία.

Σε περιπτώσεις αρκετά ασαφών περιγραφών, αυτή η πρώτη αντιστοίχηση είναι και η τελειωτική καθώς δεν μας δίνονται περισσότερα στοιχεία από τον οδηγό ώστε να προχωρήσουμε περαιτέρω. Αρκετές είναι οι περιπτώσεις όπου έχοντας πολλές σκηνές παρομοίου είδους χωρίς κάποια να έχει κάποιο ιδιαίτερο χαρακτηριστικό, οι χρήστες συνοψίζουν το σύνολο των σκηνών αυτών συγκεντρωτικά με μία γενική περιγραφή. (πχ *Numerous scenes of gun related violence*).

Υπάρχουν όμως και περιπτώσεις που ο χρήστης έχει γράψει μια πιο αναλυτική περιγραφή της σκηνής. Για τις περιπτώσεις αυτές, μετά την αρχική αυτή αντιστοίχηση συνεχίζουμε και σε μια περαιτέρω διαδικασία ώστε να επιτύχουμε και μια πιο ειδική αντιστοίχηση.

Για να πετύχουμε την αντιστοίχηση αυτή, χρησιμοποιούμε την μέθοδο *tf-idf*. Για κάθε μία από τις λέξεις της εξεταζόμενης πρότασης βρίσκουμε το χαρακτηριστικό βάρος *tf-idf* για κάθε μία από τις σκηνές που ανήκουν στην ίδια κλάση με την πρόταση. Χρησιμοποιώντας τα χαρακτηριστικά βάρη αυτά, δημιουργούμε ένα διάνυσμα για κάθε μία από τις σκηνές αυτές, καθώς και για την ίδια την πρόταση. Το διάνυσμα της πρότασης συγκρίνεται με τα αντίστοιχα διανύσματα των σκηνών, χρησιμοποιώντας την ομοιότητα συνημιτόνου.

5.3.1 Λίστα stop word

Στην πραγματικότητα, δεν χρησιμοποιείται το σύνολο των λέξεων κάθε πρότασης για την δημιουργία των διανυσμάτων. Όπως και στην επεξεργασία φυσικής γλώσσας του

σεναρίου, αφαιρούνται φυσικά όσα μέρη του λόγου έχουμε θεωρήσει ότι δεν μας προσφέρουν κάποια πληροφορία όπως άρθρα, αντωνυμίες κ.λ.π.

Παρατηρούμε ότι υπάρχουν και άλλες λέξεις που δεν περιλαμβάνονται στην προηγούμενη κατηγορία, αλλά δεν μας προσφέρουν κάποια χρήσιμη πληροφορία. Στο πλαίσιο αυτό των περιγραφών των σκηνών, οι προτάσεις περιέχουν μαζί με το περιεχόμενο της σκηνής και την διαδικασία περιγραφής μέσω φυσικής γλώσσας ενός οπτικού συμβάντος, καθώς και σε αρκετές περιπτώσεις μια προσπάθεια συγκέντρωσης διαφόρων σκηνών πιο περιληπτικά.

Τέτοιες λέξεις εμφανίζονται αρκετά συχνά, όπως μπορούμε να δούμε και από τις πιο χρησιμοποιημένες λέξεις των περιγραφών του Top250 του imdb. Εδώ συγκεντρώθηκαν οι πιο χρησιμοποιημένες λέξεις στο σύνολο των περιγραφών, ανεξαρτήτως κατηγορία. Ως κατώτερο όριο χρησιμοποιήθηκαν οι 50 εμφανίσεις της λέξης, δηλαδή τουλάχιστον 1 χρήση της λέξης στο σύνολο των περιγραφών 5 ταινιών. Με έντονη γραφή είναι σημειωμένες οι λέξεις που φαίνεται να αποτελούν υποψήφιες λέξεις εξαίρεσης.

breast::120	clothes::52	drug::62	fish::65	head::220
brief::122	**couple::57**	drunk::57	floor::52	hear::64
briefly::166	cover::77	end::80	get::163	hit::116
bullet::66	cut::127	eye::54	girl::66	**however::68**
buttock::69	dead::146	face::222	graphic::126	**is::2263**
camera::52	death::71	fall::105	ground::69	**imply::106**
car::94	die::55	**few::151**	gun::92	**include::62**
character::311	disturbing::56	fight::109	**has::161**	**intense::61**
chest::83	**do::71**	**film::195**	hand::93	kill::172
cigarette::60	drink::228	fire::70	**have::135**	kiss::120

large::59	**most::58**	**refer::97**	**show::90**	**visible::80**
later::134	mouth::72	run::60	**shown::464**	**was::73**
leg::79	**movie::98**	say::61	smoke::230	wall::61
little::57	**not::221**	**scene::587**	soldier::83	wear::63
look::51	nothing::64	scream::56	stab::83	**when::127**
lot::58	nude::69	second::118	take::93	**where::120**
main::60	nudity::112	**see::468**	talk::59	wife::51
make::56	**only::84**	**seen::416**	**then::221**	woman::530
man::1374	open::59	severe::254	time::220	women::90
many::118	**other::162**	sex::166	**use::209**	wound::141
men::210	**people::193**	sexual::113	**very::178**	young::92
mild::59	punch::90	shoot::145	violence::147	
more::55	rape::66	shot::367	**violent::74**	

Πίνακες 5.1: Υποψήφιες λέξεις εξαίρεσης

Παρατηρούμε ότι περιλαμβάνονται διάφορες λέξεις σχετικές με το τι είναι εμφανές ή όχι στην σκηνή (shown, imply, visible), με τον αριθμό των σκηνών (many, few), με την περιγραφή των γεγονότων (when, where, later) καθώς και με την ίδια την ταινία και τα συστατικά της (movie, scene, character). Επίσης, συμπεριλαμβάνονται και κάποιες λέξεις που χρησιμοποιούνται γενικότερα συχνά και όχι μόνο στο συγκεκριμένο πλαίσιο (is, have, do).

Η πλήρης λίστα με τις λέξεις εξαίρεσης προέκυψε όπως και οι λέξεις-κλειδιά των εκάστοτε σκηνών, με αρχική βάση τις συχνά χρησιμοποιούμενες και επεκτείνοντας με την βοήθεια του Wordnet.

brief, briefly	occur	camera	other
character	repeatedly	couple, few	refer
implied	scene	do	then
intense	several	have, had	use
include	sequence	is, was	very
multiple, numerous, many, lot, more, most	term	however	when
off-screen, offscreen	translate	later	where
see, seen	word	more, most	visible
show, shown	people	not	
suggestive	movie, film	only	

Πίνακες 5.2: Τελική stop word list

5.3.2 Αντιστοιχήσεις κατηγορίας Frightening/Intense Scenes

Για την 5η κατηγορία του οδηγού γονέων, τις τρομακτικές και έντονες σκηνές, ακολουθήθηκε μια ελαφρώς διαφορετική διαδικασία για την αντιστοίχηση των προτάσεων του οδηγού και τις σκηνές του σεναρίου. Όπως αναφέραμε και στην επέκταση της οντολογίας, λόγω των ειδικών συνθηκών, δεν δημιουργήθηκαν κλάσεις που να αντιστοιχούν σε αυτήν την κατηγορία. Οπότε, δεν μπορεί να γίνει, όπως στις υπόλοιπες κατηγορίες, η πρώτη αντιστοίχηση μεταξύ των προτάσεων και των σκηνών με βάση το αν ανήκουν στην ίδια κλάση της οντολογίας.

Έγινε μια προσπάθεια να γίνει αντιστοίχηση χρησιμοποιώντας κατευθείαν το δεύτερο

βήμα της διαδικασίας που χρησιμοποιήθηκε για τις υπόλοιπες κατηγορίες, την μέθοδο συχνότητας όρου - αντίστροφης συχνότητας εγγράφου. Σε αυτήν την περίπτωση, στο σύνολο των εγγράφων που εξετάστηκαν συμπεριλήφθηκαν όλες οι σκηνές του σεναρίου. Επίσης, δεν υπήρχε κάποια δικλείδα για την περίπτωση που καμία από τις αποστάσεις μεταξύ των σκηνών και της εξεταζόμενης πρότασης δεν ξεπερνούσε το όποιο κατώφλι. Αφού δεν έχουμε κάποια σχετική κατηγορία, θα έπρεπε να συμπεριλάβουμε όλες τις σκηνές της ταινίας. Αν και γενικά θεωρήθηκε προτιμότερο να συμπεριλάβουμε περισσότερες λάθος προτάσεις απ' το να αφήσουμε κάποια πρόταση χωρίς ούτε μία σωστή αντιστοίχηση, αυτό ίσχυε για μικρότερο όγκο σκηνών. Δεν προσφέρει τίποτα στον χρήστη να παρουσιάζεται το σύνολο των σκηνών της ταινίας ως αντιστοίχηση σε ένα γεγονός, καθώς ο στόχος μας είναι να προσφέρουμε την δυνατότητα στον χρήστη να εξετάσει μόνο τις επισφαλείς σκηνές χωρίς να χρειαστεί να δει όλη την ταινία.

6

Μεταβολή παραμέτρων και αποτελέσματα

Στις δύο βασικές διαδικασίες που εκτελεί το σύστημα, την κατηγοριοποίηση και την αντιστοίχηση, υπήρχαν παράμετροι οι οποίοι μπορούσαν να μεταβληθούν, αλλάζοντας τα αποτελέσματα τους συστήματος.

Για την διαδικασία της κατηγοριοποίησης των σκηνών, οι παράγοντες που έπαιζαν σημαντικό ρόλο στην ποιότητα των αποτελεσμάτων ήταν το μέτρο ομοιότητας που θα χρησιμοποιούσαμε για την σύγκριση των διανυσμάτων και το κατώφλι που θα θεωρούσαμε ως όριο για το μέτρο αυτό.

Αντίστοιχα, για την διαδικασία της αντιστοίχησης, ο βασικός παράγοντας ήταν το κατώφλι της ομοιότητας συνημιτόνου.

Η εξέταση έγινε ξεχωριστά για κάθε τμήμα, έτσι ώστε κάθε φορά να εμπλέκονται οι λιγότεροι το δυνατόν παράγοντες.

6.1 Αποτελέσματα κατηγοριοποίησης

Για να καταλήξουμε στο μέτρο και το κατώφλι που μας δίνει τα καλύτερα αποτελέσματα, εξετάσαμε διάφορα μέτρα ομοιότητας με αντίστοιχες μεταβολές κατωφλιών. Η εξέταση έγινε συνολικά, επιλέγοντας έναν συνδυασμό μέτρου ομοιότητας με το αντίστοιχο κατώφλι, που θα μας έδινε τα καλύτερα αποτελέσματα.

6.1.1 Υποψήφια μέτρα ομοιότητας

Υπάρχουν πολλά πιθανά μέτρα ομοιότητας για δυαδικά διανύσματα. Δεν είναι όμως όλα κατάλληλα για την περίπτωση μας. Οπότε, πριν ξεκινήσουμε τις δοκιμές, πρέπει να εξεταστούν οι συγκεκριμένες συνθήκες και ποια από τα μέτρα ταιριάζουν καλύτερα σε αυτές.

Ορίζοντας τις κλάσεις της οντολογίας, προσπαθήσαμε να καλύψουμε όσο πιο πολλές ειδικές περιπτώσεις γινόταν, έτσι ώστε η οντολογία μας να είναι πλήρης. Από την άλλη πλευρά, τα δεδομένα μας ήταν πάντα επαρκή για να ξεχωρίσουν υποπεριπτώσεις, οπότε έπρεπε να ορισθεί μια πιο ευρεία κλάση. Ως αποτέλεσμα, έχουμε μεγάλη ανισορροπία στο πόσο ευρύ είναι το περιεχόμενο στο οποίο αντιστοιχούν οι κλάσεις μας.

Το γεγονός αυτό αντανακλάται και στις λέξεις-κλειδιά που αντιστοιχούν σε κάθε κλάση και προφανώς και στα διανύσματα χαρακτηριστικών. Στη συνέχεια, θα δούμε πιο αναλυτικά ένα παράδειγμα.

Μία από τις κλάσεις της οντολογίας μας είναι η κλάση Electrocution, που περιλαμβάνει σκηνές που κάποιος απεικονίζεται να πεθαίνει από ηλεκτροπληξία. Δημιουργήθηκε μια ειδική υποκλάση για τον συγκεκριμένο τρόπο, αφού παρατηρήσαμε ότι στις περιγραφές οι χρήστες τον επισήμαναν ιδιαιτέρως. Μπορούσαμε επομένως, ακόμα και με την λιγότερη πληροφορία των περιγραφών, να κατηγοριοποιήσουμε τις σκηνές σε μια κλάση με πιο ειδικό ορισμό, κάτι που θα μας διευκολύνει στην συνέχεια. Όμως, για την κλάση αυτή, οι λέξεις-κλειδιά που μας επισημαίνουν ότι ένα τέτοιο γεγονός συμβαίνει δεν είναι πολλές, είναι μόνο η λέξη *electrocute*. Το διάνυσμα χαρακτηριστικών της κλάσης αυτής έχει μόνο ένα 1 και όλα τα υπόλοιπα είναι 0.

Αντίθετα, για την κλάση AlcoholEvent, που περιλαμβάνει σκηνές που κάποιος εμφανίζεται να πίνει κάτι αλκοολούχο, οι λέξεις-κλειδιά είναι πολύ περισσότερες.

Ακόμα και τοποθετώντας τις συνώνυμες λέξεις στην ίδια θέση του διανύσματος και μη περιλαμβάνοντας πολύ εξειδικευμένες κατηγορίες ποτών (γνωστές μάρκες π.χ.), το διάνυσμα χαρακτηριστικών περιλαμβάνει 17 άσσους.

Επομένως, η ζητούμενη απόσταση έπρεπε να μπορεί να καλύψει και τις δύο αυτές ακραίες περιπτώσεις βγάζοντας κατά το δυνατόν το ίδιο καλά αποτελέσματα.

Για τα κριτήρια απόστασης, η βαρύτητα πέφτει στα χαρακτηριστικά που διαφέρουν μεταξύ των διανυσμάτων. Η ιδιότητα αυτή καθιστά τις αποστάσεις αυτού του τύπου εντελώς ακατάλληλες για τον σκοπό μας.

Όταν συγκρίνουμε ένα διάνυσμα σκηνής με ένα διάνυσμα χαρακτηριστικών κλάσης, η ύπαρξη διαφοράς μεταξύ των χαρακτηριστικών σημαίνει είτε ότι, στην περίπτωση που το διάνυσμα σκηνής εμφανίζει 1 και το διάνυσμα κλάσης 0, η σκηνή περιέχει κάποιες λέξεις που είναι λέξεις-κλειδιά κάποιας άλλης κλάσης είτε ότι, στην περίπτωση που το διάνυσμα σκηνής εμφανίζει 0 και το διάνυσμα κλάσης 1, η κλάση έχει κάποιες λέξεις-κλειδιά που δεν εμφανίζονται στην σκηνή.

Σε καμία από τις δύο περιπτώσεις δεν μπορούμε να κρίνουμε για την ομοιότητα μεταξύ της σκηνής και της κλάσης.

Αντίστοιχα, δεν μπορούμε να έχουμε και καταληκτικά αποτελέσματα για την μη-ομοιότητα. Καθώς όλες οι κλάσεις έχουν τουλάχιστον μία ή περισσότερες λέξεις-κλειδιά, ο μηδενισμός των περιπτώσεων που το διάνυσμα σκηνής εμφανίζει 0 και το διάνυσμα κλάσης 1 συνεπάγεται άμεσα ότι η σκηνή περιέχει όλες τις λέξεις-κλειδιά της κλάσης. Αυτό όμως ισχύει μόνο στον μηδενισμό και όχι σε οποιαδήποτε άλλη περίπτωση, όσο μικρό και αν είναι το πλήθος των περιπτώσεων. Επίσης, η μη-ύπαρξη λέξεων-κλειδιών άλλων κλάσεων όπως και η ύπαρξη λιγοστών τέτοιων, δεν εξασφαλίζει ότι υπάρχουν λέξεις-κλειδιά της εξεταζόμενης κλάσης, κάτι που συμβαίνει όταν το διάνυσμα σκηνής εμφανίζει 1 και το διάνυσμα κλάσης 0.

Οι συγκεκριμένου τύπου αποστάσεις κατ’ επέκταση δεν εξετάστηκαν καθόλου ως

υποψήφιες.

Τα κριτήρια μη-συσχετισμού και συσχετισμού κινούνται μεταξύ του 0 και του 1, παίρνοντας την ελάχιστη τιμή τους όταν μηδενίζεται είτε ο αριθμός των κοινών 1 είτε ο αριθμός των κοινών 1 και ο αριθμός των κοινών 0 μεταξύ των διανυσμάτων. Η μέγιστη τιμή επιτυγχάνεται από έναν μηδενισμό των διαφορετικών χαρακτηριστικών και απαιτεί επιπλέον σε κάποιες περιπτώσεις μηδενισμό του αριθμού των κοινών μηδενικών (Russel&Rao, Faith).

Στα συγκεκριμένα μέτρα ξεκινάει να εμπλέκεται ο αριθμός των κοινών μηδενικών μεταξύ των δύο διανυσμάτων, που δεν είχε χρησιμοποιηθεί στην προηγούμενη κατηγορία αποστάσεων. Πρακτικά, τα κοινά μηδενικά μεταξύ ενός διανύσματος σκηνής και ενός διανύσματος κλάσης μας δείχνουν ότι κάποιες λέξεις-κλειδιά διαφορετικών κλάσεων, όχι της εξεταζόμενης, δεν εμφανίζονται στην συγκεκριμένη σκηνή, μία όχι ιδιαίτερα χρήσιμη πληροφορία.

Στην περίπτωση μας, το σημαντικότερο στοιχείο είναι εάν η σκηνή και η κλάση έχουν κοινά 1. Στο πολύ βασικό επίπεδο, εάν δεν έχουν κοινά 1 η σκηνή δεν εμπίπτει στην κλάση, ενώ εάν έχουν εμπίπτει. Για να εξασφαλιστεί αυτό, προτιμήθηκαν τα μέτρα που το a έχει μεγάλη βαρύτητα και "καθορίζει" τις ακραίες τιμές. Δηλαδή επιλέξαμε τα μέτρα που η μέγιστη τιμή 1 να μην μπορεί να επιτευχθεί χωρίς κοινά 1 μεταξύ των δύο διανυσμάτων και που η απουσία κοινών 1 να συνεπάγεται άμεσα την ελάχιστη τιμή.

Όπως αναφέραμε, για οποιοδήποτε μέτρο, απαραίτητη προϋπόθεση για να μεγιστοποιηθεί η τιμή του είναι ο μηδενισμός των διαφορετικών χαρακτηριστικών. Σε δύο περιπτώσεις, απαιτείται και μηδενισμός των κοινών μηδενικών. Αυτό όμως δεν διασφαλίζει ότι έχουμε κοινά 1 σε όλες τις περιπτώσεις. Για παράδειγμα, με μηδενισμένα *b* και το *c* το μέτρο ομοιότητας Sokal & Michener μας δίνεται από το πηλίκο , που ισούται με 1 σε κάθε περίπτωση που δεν μηδενίζονται και το a και το *d.* Μπορούμε να έχουμε την μέγιστη τιμή ομοιότητας με μηδενικό *α,* αρκεί το *d* να μην

μηδενίζεται, κάτι που συμβαίνει σχεδόν σε όλες τις περιπτώσεις. Στην αντίστοιχη περίπτωση, το μέτρο Dice & Sorenson μας δίνεται από το πηλίκο , που ισούται με 1 σε κάθε περίπτωση που δεν μηδενίζεται το a.

Όσον αφορά την ελάχιστη τιμή, το μέτρο Dice & Sorenson την παίρνει όταν μηδενίζεται το *α,* ενώ το Sokal & Michener την παίρνει όταν μηδενίζονται και το a και το *d.*

Κατ' αυτόν τον τρόπο, από τα μέτρα ομοιότητας που παραθέσαμε προηγουμένως, κρατήσαμε ως υποψήφια μόνο εκείνα που ο μηδενισμός του a επέφερε άμεσα την ελάχιστη τιμή (0) και η μέγιστη τιμή τους (1) εξασφάλιζε τον μη μηδενισμό του a. Τα μέτρα που πληρούν τις προϋποθέσεις αυτές είναι τα μέτρα ομοιότητας Jaccard, Russel & Rao, Dice & Sorenson, Sokal & Sneath I, Baroni-Urbani I, Sorgenfrei, Ochiai I, Simpson και Braun & Banquet.

6.1.2 Ανάλυση αποτελεσμάτων

Για να καταλήξουμε στο μέτρο ομοιότητας που θα χρησιμοποιήσουμε, όλα τα υποψήφια μέτρα εξετάστηκαν ώστε να δούμε με ποιο έχουμε καλύτερα αποτελέσματα. Έγινε κατηγοριοποίηση των σκηνών και των τριών ταινιών (Bangkok Dangerous, Death Defying Acts, The Walker) για κάθε υποψήφιο μέτρο και μεταβάλλοντας το κατώφλι, μεγαλύτερο του οποίου θα έπρεπε να είναι το μέτρο ομοιότητας μεταξύ μιας σκηνής και μιας κλάσης για να κατηγοριοποιηθεί σε αυτήν την κλάση η σκηνή. Οι σκηνές που δεν έχουν αρκετά ισχυρή ομοιότητα με τις κλάσεις στις οποίες κατηγοριοποιούνται θα έπρεπε να εμφανίζουν μικρά μέτρα ομοιότητας, οπότε αυξάνοντας το κατώφλι οι λάθος σκηνές δεν θα κατηγοριοποιούνται.

Σε κάθε τέτοια κατηγοριοποίηση μετρήθηκαν οι συνολικές σκηνές που κατηγοριοποιήθηκαν, πόσες από αυτές είναι σωστές και πόσες λάθος. Επίσης, μετρήθηκαν πόσες από τις σκηνές που είναι σκηνές ενδιαφέροντος δεν εντοπίστηκαν

καθόλου, καθώς και οι σκηνές που αν και δεν είναι επιλήψιμες, πρακτικά δεν είναι και λάθος. Για παράδειγμα, μια σκηνή που περιέχει ένα φιλί στο μάγουλο ή φωτιά πολύ μικρού μεγέθους.

Αρχικά το κατώφλι για όλα τα μέτρα ήταν ίσο με 0. Καθώς όλα τα υποψήφια μέτρα είχαν επιλεχτεί ώστε να μηδενίζονται για τις ίδιες συνθήκες, με κατώφλι το 0 οποιαδήποτε σκηνή είχε μη-μηδενικά κοινά χαρακτηριστικά με κάποια κλάση κατηγοριοποιούνται σε αυτή και τα αποτελέσματα για το κατώφλι αυτό είναι ίδια σε όλες τις περιπτώσεις. Για τις τρεις εξεταζόμενες ταινίες και για μηδενικό κατώφλι, ανεξάρτητα του μέτρου, τα αποτελέσματα της κατηγοριοποίησης είναι τα εξής :

	Death Defying Acts	**Bangkok Dangerous**	**The Walker**
Συνολικός αριθμός σκηνών	53	161	90
Σωστές	29	119	63
Δεκτές	3	9	5
Μη εντοπισμένες	2	13	7
Λάθος	21	33	22

Πίνακας 6.1: Αρχικά αποτελέσματα κατηγοριοποιήσεων

Αυτό που διαφέρει είναι οι τιμές που έχουν πάρει τα μέτρα και που θα καθορίσουν την περαιτέρω συμπεριφορά τους.

Παρακάτω φαίνονται τα διαφορετικά εύρη τιμών για κάθε μέτρο :

	Death Defying Acts	**Bangkok Dangerous**	**The Walker**
Jaccard	0.05 - 0.5	0.04 - 1.0	0.05 - 1.0
Russel & Rao	0.01 - 0.02	0.01 - 0.04	0.01 - 0.02
Dice & Sorenson	0.09 - 0.67	0.07 - 1.0	0.08 - 1.0
Sokal & Sneath I	0.02 - 0.33	0.02 - 1.0	0.02 - 1.0
Baroni &Urbani I	0.31 - 0.93	0.31 - 1.0	0.29 - 1.0
Sorgenfrei	0.02 - 0.5	0.01 - 1.0	0.01 - 1.0
Ochiai I	0.14 - 0.7	0.07 - 1.0	0.11 - 1.0
Simpson	0.33 - 1.0	0.08 - 1.0	0.2 - 1.0
Braun & Banquet	0.05 - 0.5	0.05 - 1.0	0.05 - 1.0

Πίνακας 6.2: Εύρη τιμών των μέτρων ομοιότητας

Αν και όλες οι τιμές κυμαίνονται μεταξύ του 0 και του 1, παρατηρούμε ότι υπάρχουν μεγάλες διαφορές μεταξύ τους. Οπότε, δεν θα είχαμε τα καλύτερα δυνατά αποτελέσματα εάν μεταβάλαμε από κοινού τα κατώφλια. Πρέπει να εξετάσουμε τα αποτελέσματα λαμβάνοντας υπόψιν τις διαφορές των τιμών τους, διατηρώντας όμως μια αναλογία μεταξύ των διάφορων μέτρων ώστε να έχει νόημα η σύγκριση των αποτελεσμάτων. Για επιτευχθεί αυτό, για κάθε μέτρο εξετάστηκαν 3 μεταβολές του κατωφλίου με την πρώτη τιμή να είναι ίση με το υψηλότερο κάτω άκρο των τριών ευρών και με το βήμα μεταβολής να είναι ίσο με το 10% του μικρότερου εύρους.

	1ο κατώφλι	2ο κατώφλι	3ο κατώφλι
Jaccard	0.05	0.1	0.15
Russel & Rao	0.01	0.011	0.012
Dice & Sorenson	0.09	0.15	0.21
Sokal & Sneath I	0.02	0.05	0.08
Baroni &Urbani I	0.31	0.37	0.43
Sorgenfrei	0.02	0.07	0.12
Ochiai I	0.14	0.2	0.26
Simpson	0.33	0.4	0.47
Braun & Banquet	0.05	0.1	0.15

Πίνακας 6.3: Κατώφλια μέτρων ομοιότητας

Όλες οι τιμές έχουν στρογγυλοποιηθεί στο 2ο δεκαδικό ψηφίο με μοναδική εξαίρεση τις τιμές του μέτρου Russel & Rao, λόγω του εξαιρετικά μικρού εύρους τιμών του.

Τα αποτελέσματα της διαδικασίας αυτής παρουσιάζονται αναλυτικά παρακάτω :

Jaccard

-0.05

	Death Defying Acts	**Bangkok Dangerous**	**The Walker**
Συν.αριθμός σκηνών	52	158	86
Σωστές	28	117	61
Δεκτές	3	9	5
Μη εντοπισμένες	3	15	9
Λάθος	21	32	20

Πίνακας 6.4: Αποτελέσματα για μέτρο ομοιότητας Jaccard και κατώφλι 0.05

-0.1

	Death Defying Acts	**Bangkok Dangerous**	**The Walker**
Συν.αριθμός σκηνών	40	99	59
Σωστές	25	78	44
Δεκτές	3	9	5
Μη εντοπισμένες	6	54	26
Λάθος	12	12	10

Πίνακας 6.5: Αποτελέσματα για μέτρο ομοιότητας Jaccard και κατώφλι 0.1

-0.15

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	31	73	44
Σωστές	21	60	33
Δεκτές	3	8	3
Μη εντοπισμένες	10	72	47
Λάθος	7	5	8

Πίνακας 6.6: Αποτελέσματα για μέτρο ομοιότητας Jaccard και κατώφλι 0.15

Russel & Rao

-0.01

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	18	38	44
Σωστές	18	36	40
Δεκτές	0	1	0
Μη εντοπισμένες	13	95	30
Λάθος	0	1	30

Πίνακας 6.7: Αποτελέσματα για μέτρο ομοιότητας Russel&Rao και κατώφλι 0.01

-0.011

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	4	21	2
Σωστές	4	19	2
Δεκτές	0	1	0
Μη εντοπισμένες	27	113	68
Λάθος	0	1	0

Πίνακας 6.8: Αποτελέσματα για μέτρο ομοιότητας Russel&Rao και κατώφλι 0.011

-0.012

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	4	21	2
Σωστές	4	19	2
Δεκτές	0	1	0
Μη εντοπισμένες	27	113	68
Λάθος	0	1	0

Πίνακας 6.9: Αποτελέσματα για μέτρο ομοιότητας Russel&Rao και κατώφλι 0.012

Dice & Sorenson

-0.09

	Death Defying Acts	**Bangkok Dangerous**	**The Walker**
Συν.αριθμός σκηνών	52	158	87
Σωστές	28	117	62
Δεκτές	3	9	5
Μη εντοπισμένες	3	15	8
Λάθος	21	32	20

Πίνακας 6.10: Αποτελέσματα για μέτρο ομοιότητας Dice&Sorenson και κατώφλι 0.09

-0.15

	Death Defying Acts	**Bangkok Dangerous**	**The Walker**
Συν.αριθμός σκηνών	45	110	74
Σωστές	27	83	53
Δεκτές	3	9	5
Μη εντοπισμένες	4	49	17
Λάθος	15	18	16

Πίνακας 6.11: Αποτελέσματα για μέτρο ομοιότητας Dice&Sorenson και κατώφλι 0.15

-0.21

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	38	92	54
Σωστές	25	70	40
Δεκτές	3	9	5
Μη εντοπισμένες	6	62	30
Λάθος	10	13	9

Πίνακας 6.12: Αποτελέσματα για μέτρο ομοιότητας Dice&Sorenson και κατώφλι 0.21

Sokal & Sneath I

-0.02

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	53	160	90
Σωστές	29	119	63
Δεκτές	3	9	5
Μη εντοπισμένες	2	13	7
Λάθος	21	32	22

Πίνακας 6.13: Αποτελέσματα για μέτρο ομοιότητας Sokal&Sneath I και κατώφλι 0.02

-0.05

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	38	103	67
Σωστές	25	79	50
Δεκτές	3	9	5
Μη εντοπισμένες	6	53	20
Λάθος	10	15	12

Πίνακας 6.14: Αποτελέσματα για μέτρο ομοιότητας Sokal&Sneath I και κατώφλι 0.05

-0.08

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	31	74	44
Σωστές	21	58	32
Δεκτές	3	8	3
Μη εντοπισμένες	10	74	38
Λάθος	7	8	9

Πίνακας 6.15: Αποτελέσματα για μέτρο ομοιότητας Sokal&Sneath I και κατώφλι 0.08

Baroni & Urbani I

-0.31

	Death Defying Acts	**Bangkok Dangerous**	**The Walker**
Συν.αριθμός σκηνών	53	161	88
Σωστές	29	119	62
Δεκτές	3	9	5
Μη εντοπισμένες	2	13	8
Λάθος	21	33	21

Πίνακας 6.16: Αποτελέσματα για μέτρο ομοιότητας Baroni&Urbani I και κατώφλι 0.31

-0.37

	Death Defying Acts	**Bangkok Dangerous**	**The Walker**
Συν.αριθμός σκηνών	52	158	87
Σωστές	28	117	62
Δεκτές	3	9	5
Μη εντοπισμένες	3	15	8
Λάθος	21	32	20

Πίνακας 6.17: Αποτελέσματα για μέτρο ομοιότητας Baroni&Urbani I και κατώφλι 0.37

-0.43

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	51	153	84
Σωστές	28	115	60
Δεκτές	3	9	5
Μη εντοπισμένες	3	17	10
Λάθος	20	29	19

Πίνακας 6.18: Αποτελέσματα για μέτρο ομοιότητας Baroni&Urbani I και κατώφλι 0.43

Sorgenfrei

-0.02

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	53	149	82
Σωστές	29	114	58
Δεκτές	3	9	5
Μη εντοπισμένες	2	18	12
Λάθος	21	26	19

Πίνακας 6.19: Αποτελέσματα για μέτρο ομοιότητας Sorgenfrei και κατώφλι 0.02

-0.07

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	36	87	58
Σωστές	23	67	42
Δεκτές	3	9	3
Μη εντοπισμένες	8	65	28
Λάθος	10	11	13

Πίνακας 6.20: Αποτελέσματα για μέτρο ομοιότητας Sorgenfrei και κατώφλι 0.07

-0.12

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	30	59	43
Σωστές	20	45	33
Δεκτές	3	9	3
Μη εντοπισμένες	11	87	37
Λάθος	7	5	7

Πίνακας 6.21: Αποτελέσματα για μέτρο ομοιότητας Sorgenfrei και κατώφλι 0.12

Ochiai I

-0.14

	Death Defying Acts	**Bangkok Dangerous**	**The Walker**
Συν.αριθμός σκηνών	53	150	84
Σωστές	29	114	59
Δεκτές	3	9	5
Μη εντοπισμένες	2	18	11
Λάθος	21	27	20

Πίνακας 6.22: Αποτελέσματα για μέτρο ομοιότητας Ochiai I και κατώφλι 0.14

-0.2

	Death Defying Acts	**Bangkok Dangerous**	**The Walker**
Συν.αριθμός σκηνών	47	128	73
Σωστές	27	100	54
Δεκτές	3	9	5
Μη εντοπισμένες	4	32	16
Λάθος	17	19	14

Πίνακας 6.23: Αποτελέσματα για μέτρο ομοιότητας Ochiai I και κατώφλι 0.2

-0.26

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	36	87	58
Σωστές	23	66	43
Δεκτές	3	9	3
Μη εντοπισμένες	8	66	27
Λάθος	10	12	12

Πίνακας 6.24: Αποτελέσματα για μέτρο ομοιότητας Ochiai I και κατώφλι 0.26

Simpson

-0.33

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	53	137	68
Σωστές	29	105	50
Δεκτές	3	9	3
Μη εντοπισμένες	2	27	20
Λάθος	21	23	15

Πίνακας 6.25: Αποτελέσματα για μέτρο ομοιότητας Simpson και κατώφλι 0.33

-0.4

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	47	111	65
Σωστές	25	86	47
Δεκτές	3	9	3
Μη εντοπισμένες	6	46	23
Λάθος	19	16	15

Πίνακας 6.26: Αποτελέσματα για μέτρο ομοιότητας Simpson και κατώφλι 0.4

-0.47

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	47	110	65
Σωστές	25	85	47
Δεκτές	3	9	3
Μη εντοπισμένες	6	47	23
Λάθος	19	16	15

Πίνακας 6.27: Αποτελέσματα για μέτρο ομοιότητας Simpson και κατώφλι 0.47

Braun & Banquet

-0.05

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	52	160	87
Σωστές	28	118	62
Δεκτές	3	9	5
Μη εντοπισμένες	3	14	8
Λάθος	21	33	20

Πίνακας 6.28: Αποτελέσματα για μέτρο ομοιότητας Braun&Banquet και κατώφλι 0.05

-0.1

	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	38	105	65
Σωστές	25	81	48
Δεκτές	3	9	5
Μη εντοπισμένες	6	51	22
Λάθος	10	15	12

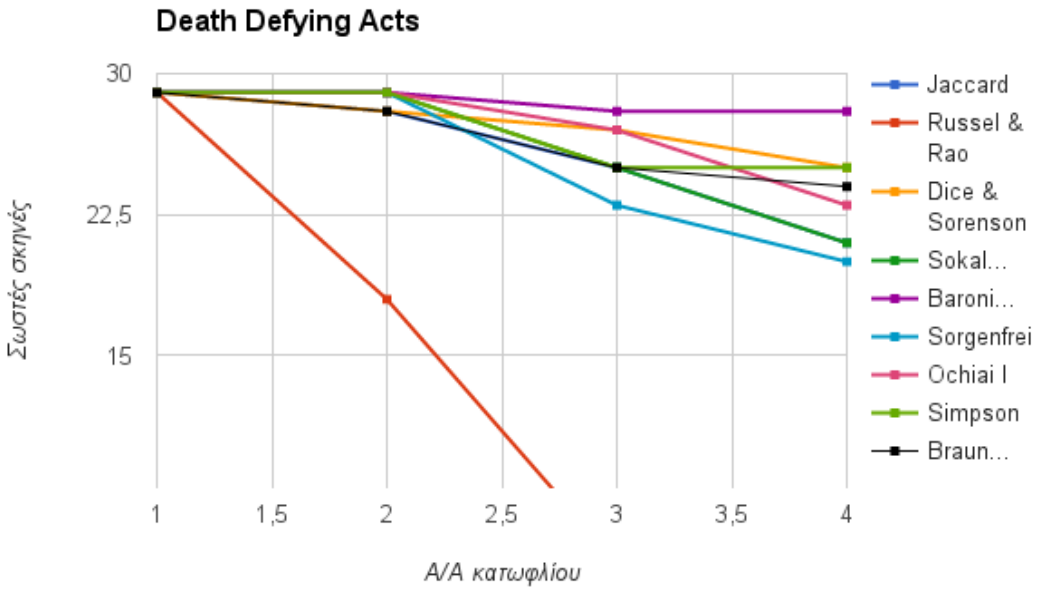
Πίνακας 6.29: Αποτελέσματα για μέτρο ομοιότητας Braun&Banquet και κατώφλι 0.1

-0.15

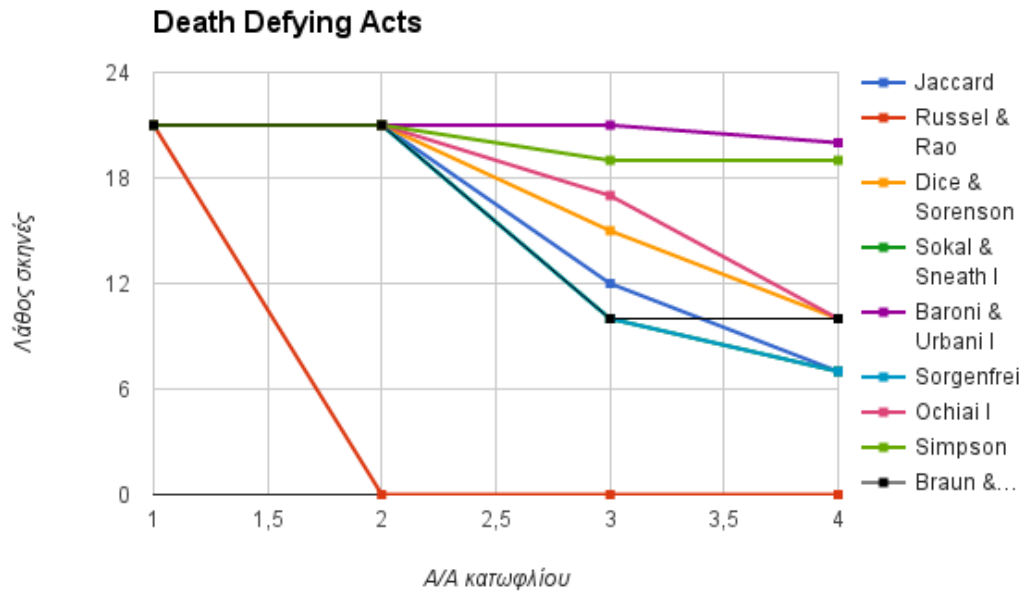
	Death Defying Acts	Bangkok Dangerous	The Walker
Συν.αριθμός σκηνών	37	89	61
Σωστές	24	70	44
Δεκτές	3	8	5
Μη εντοπισμένες	7	62	26
Λάθος	10	11	12

Πίνακας 6.30: Αποτελέσματα για μέτρο ομοιότητας Braun&Banquet και κατώφλι 0.15

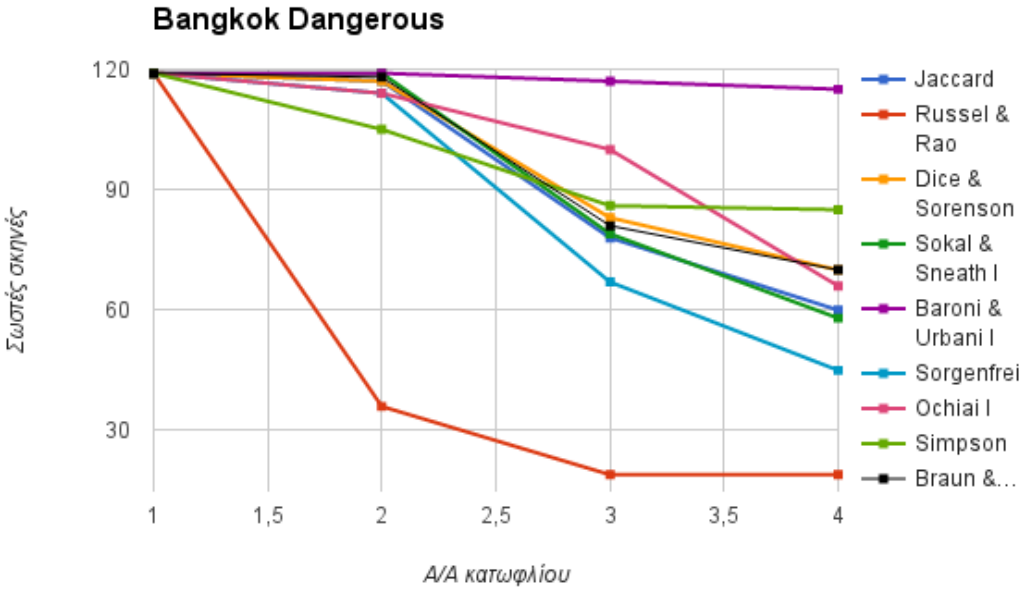
Στη συνέχεια, παραθέτουμε συγκριτικά διαγράμματα των διαφόρων μέτρων ομοιότητας και των αντίστοιχων αποτελεσμάτων τους ανά ταινία που εξετάστηκε :



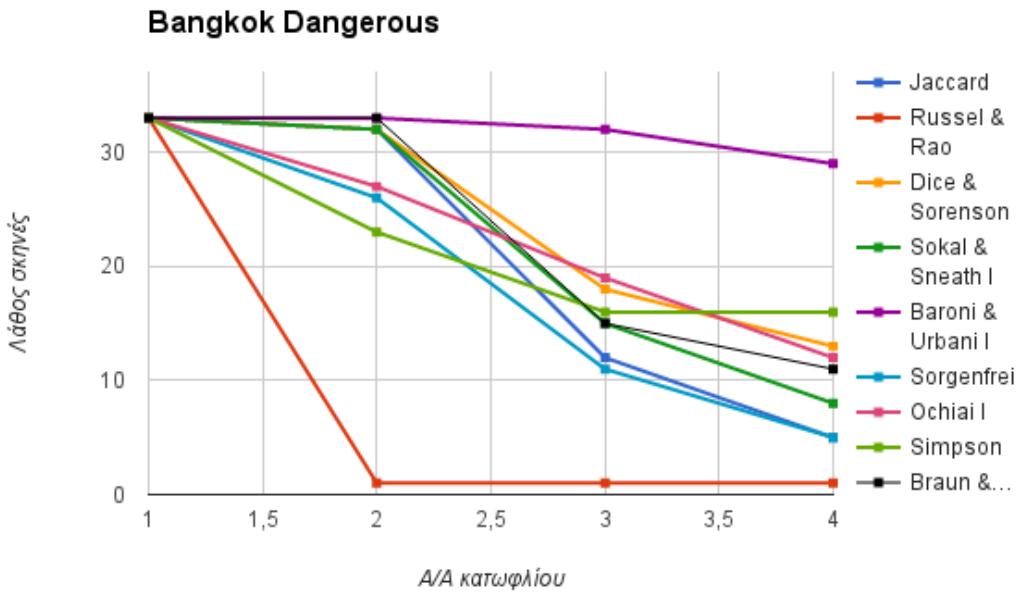
Σχήμα 6.1 : Διάγραμμα σωστών σκηνών για την ταινία Death Defying Acts



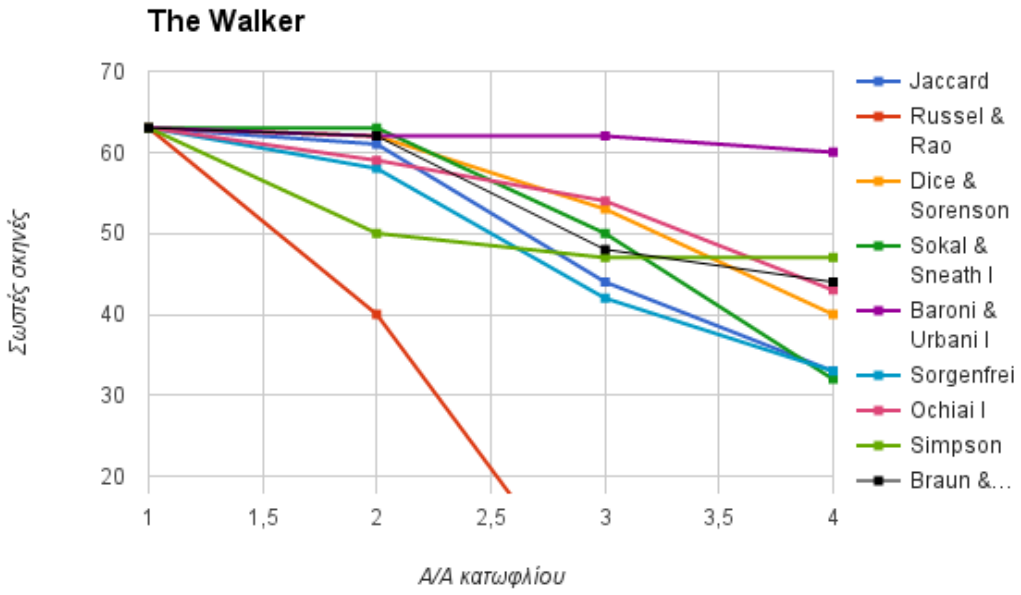
Σχήμα 6.2 : Διάγραμμα λάθος σκηνών για την ταινία Death Defying Acts



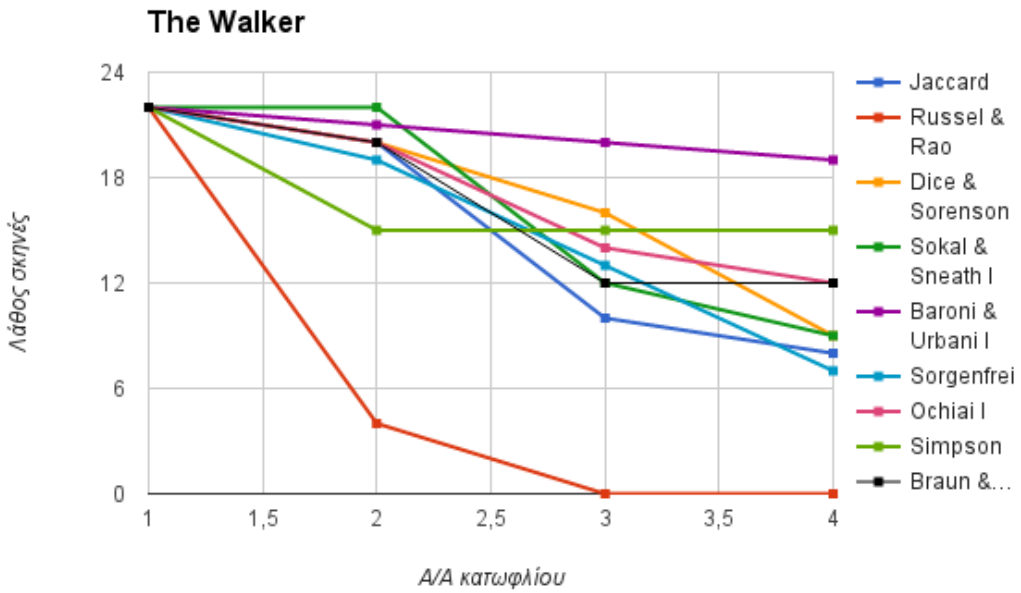
Σχήμα 6.3 : Διάγραμμα σωστών σκηνών για την ταινία Bangkok Dangerous



Σχήμα 6.4 : Διάγραμμα λάθος σκηνών για την ταινία Bangkok Dangerous



Σχήμα 6.5 : Διάγραμμα σωστών σκηνών για την ταινία The Walker



Σχήμα 6.6 : Διάγραμμα λάθος σκηνών για την ταινία The Walker

Καθώς τα κατώφλια αυξάνονται, είναι αναμενόμενο ότι μειώνεται ο αριθμός των σκηνών που κατηγοριοποιούνται, είτε σωστά είτε λάθος, καθώς τα μέτρα ομοιότητας ξεπερνούν δυσκολότερα το κατώφλι. Ιδανικά θα θέλαμε οι σωστές κατηγοριοποιήσεις να μειώνονταν ελάχιστα ή και καθόλου, ενώ οι λάθος κατηγοριοποιήσεις να εξαλείφονταν. Καθώς αυτό είναι πρακτικά αδύνατο, στόχος μας είναι να βρούμε ποιο μέτρο ομοιότητας και για ποιο αντίστοιχο κατώφλι, μας δίνουν τον καλύτερο συνδυασμό αποτελεσμάτων, ώστε οι σωστές κατηγοριοποιήσεις να μένουν σε υψηλά επίπεδα ενώ οι λάθος κατηγοριοποιήσεις να μειώνονται επαρκώς.

Καθώς το πλήθος των συνολικών σκηνών διαφέρει μεταξύ των ταινιών, δεν θα δουλέψουμε με τους απόλυτους αριθμούς των σωστών και των λάθος σκηνών, αλλά με ποσοστά επί του συνόλου των σωστών κατηγοριοποιήσεων που υπάρχουν σε κάθε ταινία. Για τις εξεταζόμενες ταινίες, υπάρχουν 31 στην ταινία *Death Defying Acts*,

132 στην ταινία *Bangkok dangerous* και 70 στην ταινία *The Walker*.

Για τις σωστές σκηνές, θέσαμε ως κατώτατο όριο το 65% και για τις λάθος, θέσαμε ως ανώτερο όριο το 25%. Παρακάτω θα δούμε τους συνδυασμούς που εξετάστηκαν και ποια μέτρα ομοιότητας, αν υπάρχουν, ικανοποίησαν τις συνθήκες αυτές. Ο αριθμός που παρατίθεται δίπλα στο μέτρο ομοιότητας μας υποδεικνύει το αντίστοιχο κατώφλι του μέτρου για το οποίο επιτυγχάνεται ο συνδυασμός.

-The Walker

		Λάθος σκηνές		
		<15%	**<20%**	**<25%**
	>80%	**---**	**---**	**---**
Σωστές σκηνές	**>75%**	**---**	**---**	Ochiai 3
	>70%	**---**	Sokal & Sneath 3	Dice & Sorenson 3 Sokal & Sneath 3 Ochiai 3 Simpson 2
	>65%	**---**	Sokal & Sneath 3 Ochiai 3 Braun & Banquet 3	Dice & Sorenson 3 Sokal & Sneath 3 Ochiai 3 Simpson 2/3/4 Braun & Banquet 3

Πίνακας 6.31: Συγκεντρωτικά αποτελέσματα για την ταινία The Walker

-Bangkok Dangerous

		Λάθος σκηνές		
		<15%	**<20%**	**<25%**
	>80%	**---**	**---**	Jaccard 2 Dice & Sorenson 2 Sokal & Sneath 2 Baroni & Urbani 3/4 Sorgenfrei 2 Ochiai 2/3
Σωστές σκηνές	**>75%**	Ochiai 3	Ochiai 3 Simpson 2	Jaccard 2 Dice & Sorenson 2 Sokal & Sneath 2 Baroni & Urbani 3/4 Sorgenfrei 2 Ochiai 2/3 Simpson 2
	>70%	Ochiai 3	Ochiai 3 Simpson 2	Jaccard 2 Dice & Sorenson 2 Sokal & Sneath 2 Baroni & Urbani 3/4 Sorgenfrei 2 Ochiai 2/3 Simpson 2
	>65%	Ochiai 3	Dice & Sorenson 3 Sorgenfrei 2 Ochiai 3 Simpson 2	Jaccard 2 Dice & Sorenson 2 Sokal & Sneath 2 Baroni & Urbani 3/4 Sorgenfrei 2 Ochiai 2/3 Simpson 2

Πίνακας 6.32: Συγκεντρωτικά αποτελέσματα για την ταινία Bangkok Dangerous

-Death Defying Acts

		Λάθος σκηνές		
		<15%	**<20%**	**<25%**
	>80%	**---**	**---**	**---**
Σωστές σκηνές	**>75%**	**---**	**---**	**---**
	>70%	**---**	**---**	**---**
	>65%	**---**	**---**	Jaccard 4, Sokal & Sneath 4

Πίνακας 6.33: Συγκεντρωτικά αποτελέσματα για την ταινία Death Defying Acts

Είναι εμφανές ότι το μέτρο ομοιότητας που μας δίνει τα καλύτερα αποτελέσματα είναι το μέτρο Ochiai για κατώφλι ίσο με 0.2. Εξαίρεση αποτελεί η ταινία *Death Defying Acts*, η οποία έχοντας πολύ λίγες επιλήψιμες σκηνές στο σύνολο της, πληρεί δύσκολα τα κριτήρια που θέσαμε προηγουμένως. Αν και οι σωστές σκηνές της για το μέτρο αυτό κρατούνται σε πολύ υψηλά επίπεδα (87%), οι λάθος κατηγοριοποιήσεις δεν μειώνονται επαρκώς και μένουν στο 55%. Για την ταινία *Bangkok dangerous*, τα αντίστοιχα ποσοστά βρίσκονται στο 76% και 14% και για την ταινία *The Walker,* στο 77% και 20%.

6.2 Αποτελέσματα αντιστοίχησης

Στην διαδικασία της αντιστοίχησης, η ομοιότητα που θα χρησιμοποιήσουμε είναι η ομοιότητα συνημιτόνου. Το ζητούμενο μας είναι το κατώφλι που θα ορίσουμε για την ομοιότητα αυτή.

Σε αντίθεση με την περίπτωση της κατηγοριοποίησης, το κατώφλι εδώ δεν

χρησιμοποιήθηκε ως απόλυτο όριο.

Σε εκείνη την περίπτωση, θεωρούσαμε ότι εάν η ομοιότητα μιας σκηνής με μία κατηγορία δεν υπερέβαινε το κατώφλι η σκηνή δεν ανήκε στην κατηγορία αυτή και εάν αυτό συνέβαινε για όλες τις κατηγορίες, η σκηνή δεν ανήκε σε καμία κατηγορία και δεν ήταν επιλήψιμη.

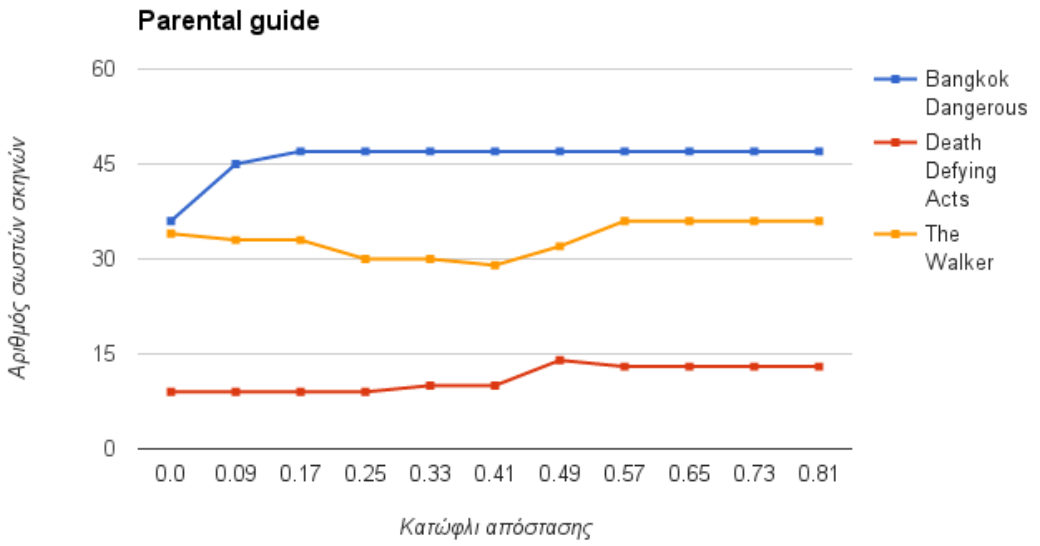
Εδώ, εφόσον οι προτάσεις περιγράφουν επιλήψιμες σκηνές, είναι αδύνατο να μην αντιστοιχούν σε κάποιες σκηνές. Στην περίπτωση που καμία από τις εξεταζόμενες σκηνές δεν έχει ομοιότητα που να υπερβαίνει το όποιο κατώφλι, θεωρούμε ότι αυτό συμβαίνει επειδή η πρόταση περιγράφει κάτι με πολύ γενικό ή περιληπτικό τρόπο και δεν μας δίνει επαρκή περαιτέρω στοιχεία ώστε να βρούμε μια πιο ειδική αντιστοίχηση με κάποια συγκεκριμένη σκηνή. Οπότε, στην πρόταση αυτή αντιστοιχούμε όλες τις σκηνές με τις οποίες έχει ίδιο τύπο.

Το βήμα με το οποίο μεταβάλλεται το κατώφλι της απόστασης καθορίστηκε από το μικρότερος εύρος τιμών, που στην συγκεκριμένη περίπτωση ήταν της ταινίας *Death Defying Acts*, με τις τιμές των αποστάσεων να κυμαίνονται από 0.25-1. Το βήμα ορίστηκε στο 10% του εύρους αυτού, 0.08.

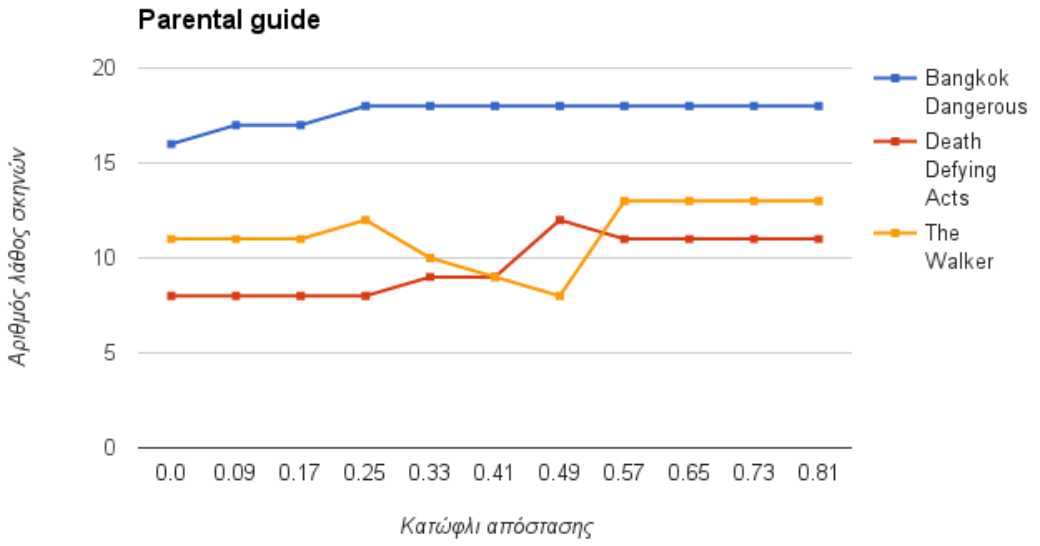
Αρχικά, η πρώτη τιμή κατωφλίου είχε οριστεί ως το ψηλότερο άκρο των ευρών των αποστάσεων, αλλά κρίνοντας ότι η διαφορά από το 0.03 που ήταν το χαμηλότερο άκρο στο 0.25 που ήταν το υψηλότερο ήταν ιδιαιτέρως μεγάλη, προστέθηκαν άλλες δύο τιμές μεταξύ του αρχικού 0.0 και του 0.25.

Στην καταμέτρηση των σκηνών που αντιστοιχούνται σωστά ή λάθος με τις προτάσεις του οδηγού παρατηρήθηκε ότι υπήρχαν προτάσεις που το σύστημα δεν θα μπορούσε ποτέ να αντιστοιχήσει. Αυτό συμβαίνει είτε επειδή περιγράφουν κάτι το οποίο δεν φαίνεται να συμβαίνει στην ταινία αλλά υπονοείται (*Rape is implied, Sex is implied*) είτε επειδή το συγκεκριμένο γεγονός, αν και συμβαίνει στην ταινία, δεν έχει

περιγραφεί στο σενάριο (*A woman's bare back can be seen, as well as a partial side shot of her breast, A few shots of other topless women beside a pool.*). Για τις δύο περιπτώσεις που το γεγονός υπονοείται, δεν υπάρχει καμία αντιστοίχηση καθώς καμία σκηνή της ταινίας εμπίπτει στην ίδια κατηγορία με την εκάστοτε πρόταση και για τις άλλες δύο, εμφανίζονται κάποιες λίγες, προφανώς λάθος, σκηνές που εμπίπτουν στην ίδια κατηγορία με την πρόταση. Αν και τα αποτελέσματα που παρουσιάζονται για τις σκηνές αυτές είναι συγκριτικά με τις συνθήκες καλά, οι προτάσεις αυτές αγνοήθηκαν στην καταμέτρηση των αποτελεσμάτων που θα παρουσιαστούν στην συνέχεια, καθώς δεν επηρεάζονται ιδιαιτέρως από το σύστημα και δεν θα μας προσφέρουν χρήσιμες πληροφορίες για την επίδοση του.



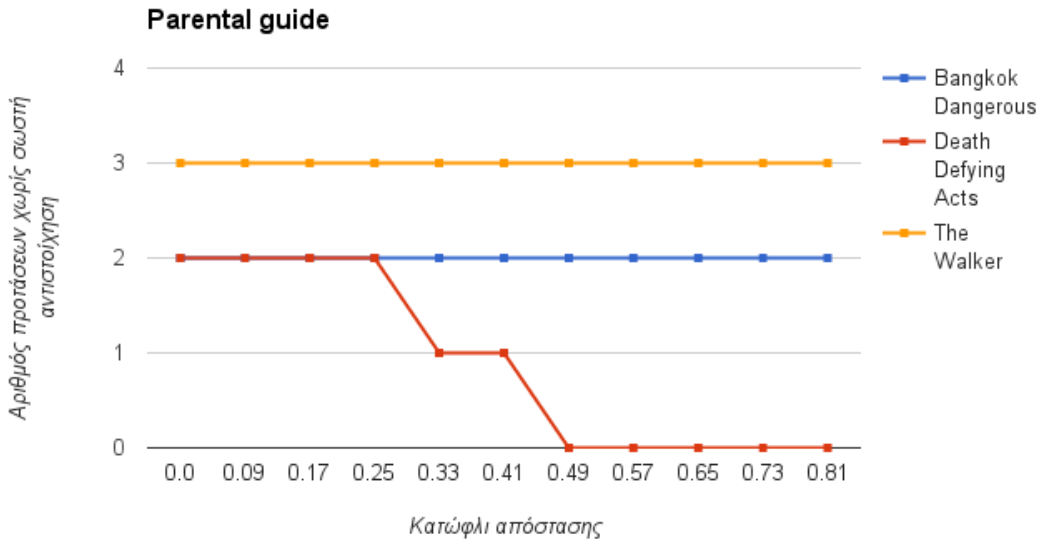
Σχήμα 6.7 : Διάγραμμα σωστών σκηνών για τις περιγραφές του οδηγού



Σχήμα 6.8 : Διάγραμμα λάθος σκηνών για τις περιγραφές του οδηγού

Παρατηρούμε ότι στην πλειοψηφία των περιπτώσεων όταν το κατώφλι έχει αυξηθεί πολύ, ο αριθμός των σκηνών σταθεροποιείται. Στις περιπτώσεις αυτές, ελάχιστες αποστάσεις μεταξύ των σκηνών και της πρότασης καταφέρνουν να ξεπεράσουν το κατώφλι και έτσι η αντιστοίχηση πέφτει στο επίπεδο της κοινής κατηγορίας. Το γεγονός αυτό ευνοεί τις προτάσεις με γενικές περιγραφές, αλλά αυξάνει και τα λάθη στις πιο ειδικές περιγραφές.

Μία άλλη παράμετρος που εξετάστηκε, εκτός από το σύνολο των λάθος και σωστών περιγραφών, ήταν ο αριθμός των προτάσεων που ανάμεσα στις σκηνές που τους αντιστοιχήθηκαν δεν περιλαμβανόταν ούτε ένα σωστό. Θεωρήθηκε ιδιαίτερα σημαντικό να υπάρχει τουλάχιστον μία σωστή αντιστοίχηση για κάθε πρόταση.



Σχήμα 6.9 : Διάγραμμα μηδενικών αντιστοιχίσεων για τις περιγραφές του οδηγού

Για τις δύο από τις τρεις ταινίες, ο αριθμός των προτάσεων παραμένει σταθερός. Για την ταινία *Death Defying Acts* όμως, η μεταβολή του κατωφλίου επιτυγχάνει την μείωση των προτάσεων. Εξίσου σημαντική, αν και ως αρνητικό αποτέλεσμα, θα θεωρούσαμε αν είχαμε κάποια μεταβολή προς τα πάνω.

Τα αποτελέσματα σε αυτήν την περίπτωση δεν ελαττώνονται σταθερά, όπως συνέβαινε στα αποτελέσματα της κατηγοριοποίησης, καθώς αυξάνεται το κατώφλι. Κατ' επέκταση, λόγω των σχετικά αυτόνομων αυξομειώσεων, δεν μπορεί να βρεθεί αρκετά εύκολα κάποιο κατώφλι που να ικανοποιεί όλες τις περιπτώσεις.

Εφόσον για κατώφλι μεγαλύτερο του 0.49 μηδενίζονται οι προτάσεις χωρίς αντιστοίχηση για την ταινία *Death Defying Acts*, θα προτιμήσουμε κάποιο κατώφλι σε αυτές τις τιμές. Είναι ιδιαίτερα σημαντικό όχι μόνο για να καταφέρουμε να έχουμε έστω και μία αντιστοίχηση σε περισσότερες προτάσεις, αλλά ιδιαίτερα για την ταινία αυτή τα αποτελέσματα μεταξύ των λάθος και των σωστών σκηνών είναι άμεσα συγκρίσιμα και το γεγονός αυτό ανοίγει εμμέσως την ψαλίδα. Καθώς για κατώφλι μεγαλύτερο του 0.57, όλες οι τιμές σταθεροποιούνται, οι δύο εναλλακτικές είναι

κατώφλι 0.49 και 0.57.

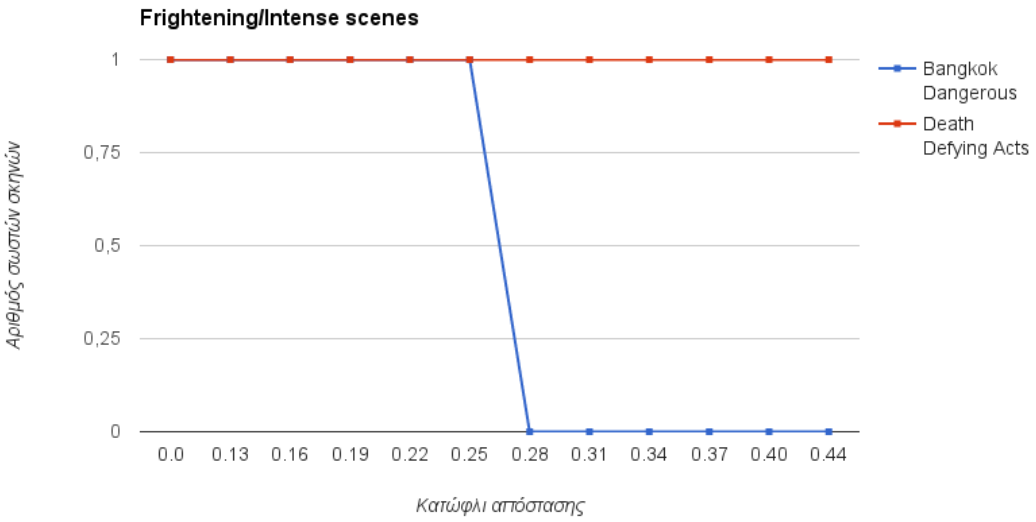
Για κατώφλι 0.57 έχουμε είτε την υψηλότερη τιμή είτε την δεύτερη υψηλότερη, με ελάχιστη διαφορά από την πρώτη, όσον αφορά τις σωστές σκηνές. Το ίδιο συμβαίνει και για τις λάθος σκηνές, αλλά δεν έχουμε κάποια ιδιαίτερα καλύτερη επιλογή. Για την περίπτωση αυτή έχουμε επί του συνόλου των σκηνών που εμφανίζονται ποσοστό σωστών σκηνών 72% για την ταινία *Bangkok Dangerous* , 54% για την ταινία *Death Defying Acts* και 73% για την ταινία *The Walker*.

6.3 Αποτελέσματα αντιστοιχίσεων κατηγορίας Frightening/Intense Scenes

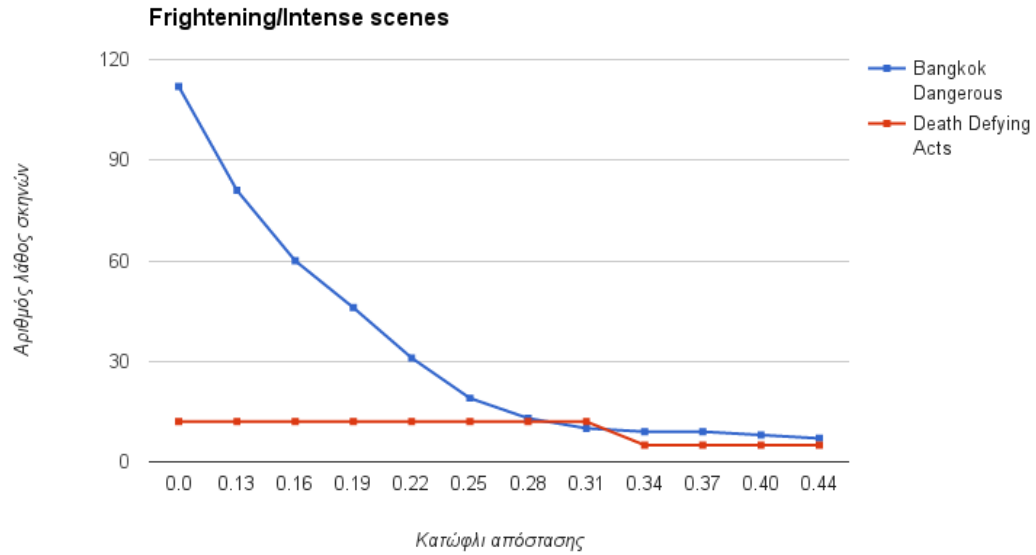
Για την κατηγορία αυτή ακολουθήθηκε κατευθείαν η διαδικασία της αντιστοίχησης, χωρίς να προηγηθούν οι κατηγοριοποιήσεις των σκηνών. Γι' αυτόν τον λόγο, εξετάστηκε ξεχωριστά, καθώς δεν θα είχε συγκρίσιμα αποτελέσματα με τις άλλες κατηγορίες και πιθανόν θα επωφελούνταν από χαμηλότερα κατώφλια για να έχει καλύτερα αποτελέσματα.

Από τις 3 εξεταζόμενες ταινίες, εδώ δουλέψαμε με τις 2. Η ταινία *The Walker* περιλάμβανε μόνο μία πρόταση για τον τομέα αυτό, "*See Violence & Gore*", που δεν δίνει επί της ουσίας πληροφορίες στον χρήστη για τις σκηνές αλλά τον κατευθύνει σε μία άλλη κατηγορία. Ακόμα και με το μικρότερο δυνατό κατώφλι (0.0), καμία σκηνή δεν μπόρεσε να το ξεπεράσει. Οπότε, για την ταινία αυτή δεν έχουμε καμία αντιστοίχηση, γεγονός που παραμένει σταθερό με την αύξηση του κατωφλίου.

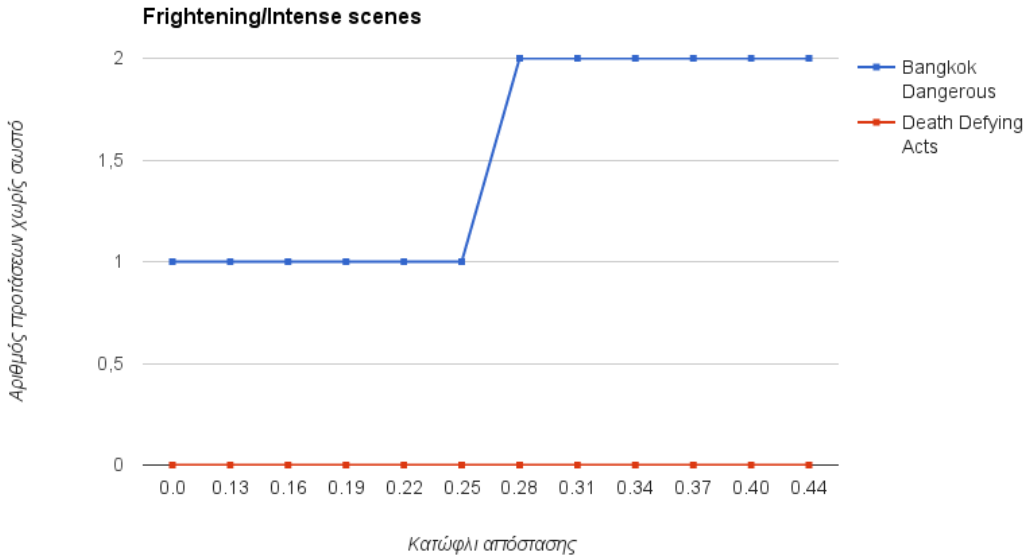
Για τις άλλες δύο ταινίες, το βήμα του κατωφλίου επιλέχτηκε πάλι ως το 10% του μικρότερου εύρους τιμών (0.33-0.57). Όπως όμως και προηγουμένως, θεωρήθηκε ότι καθώς τα κατώτερα όρια των ευρών διέφεραν πολύ (0.08, 0.33), θα ήταν προτιμότερο να μην ξεκινήσουμε από το υψηλότερο όριο ως πρώτο κατώφλι.



Σχήμα 6.10 : Διάγραμμα σωστών σκηνών για την κατηγορία Frightening/Intense scenes



Σχήμα 6.11 : Διάγραμμα λάθος σκηνών για την κατηγορία Frightening/Intense scenes



Σχήμα 6.12 : Διάγραμμα μηδενικών αντιστοιχίσεων για την κατηγορία Frightening/Intense scenes

Όπως έχει ήδη αναφερθεί, η συγκεκριμένη κατηγορία παρουσιάζει ιδιαίτερες δυσκολίες λόγω των σκηνών που περιλαμβάνονται και τον τρόπο περιγραφής τους, που προσφέρουν ελάχιστες ή δύσκολα εντοπίσιμες πληροφορίες. Σε συνδυασμό με το γεγονός ότι εδώ εξετάζουμε το σύνολο των σκηνών, τα αποτελέσματα είναι αναμενόμενα χειρότερα, ακόμα και για μικρότερα κατώφλια από τα αντίστοιχα των υπόλοιπων κατηγοριών.

Είναι αρκετά εμφανές από το γεγονός ότι για κατώφλι μεγαλύτερο του 0.25, οι σωστές προτάσεις της ταινίας *Bangkok Dangerous* μηδενίζονται ότι δεν μπορούμε να χρησιμοποιήσουμε κάποια από τις τιμές αυτές. Για τις μικρότερες τιμές, η ταινία *Death Defying Acts* παρουσιάζει σταθερές τιμές, ενώ η ταινία *Bangkok Dangerous* έχει σταθερό αριθμό σωστών σκηνών και οι λάθος σκηνές μειώνονται όσο αυξάνεται το κατώφλι. Επομένως, η καλύτερη τιμή για το κατώφλι είναι η 0.25.

7

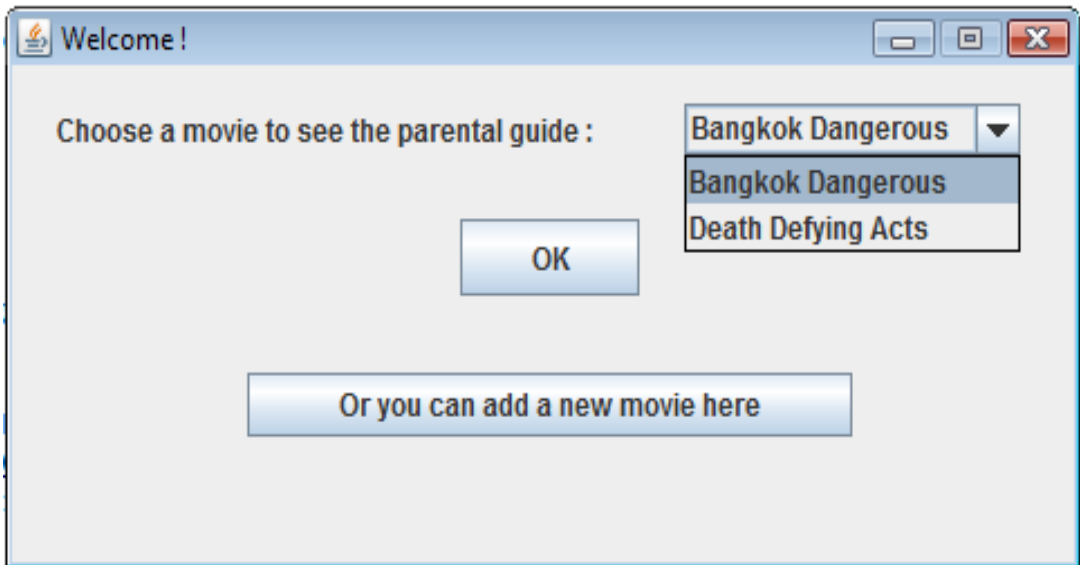
Περιγραφή συστήματος

Στο κεφάλαιο αυτό θα παρουσιάσουμε τις λειτουργίες που προσφέρει το σύστημα, τις κλάσεις που το αποτελούν, καθώς και τα εργαλεία και τις βιβλιοθήκες που χρησιμοποιήθηκαν για την υλοποίηση του.

7.1 Λειτουργίες του συστήματος

Το σύστημα που δημιουργήσαμε προσφέρει δύο βασικές λειτουργίες στον χρήστη, την δυνατότητα να προσθέσει μία νέα ταινία στο σύστημα και την δυνατότητα να δει τις "προβληματικές" σκηνές ταινίες που ήδη υπάρχουν στο σύστημα.

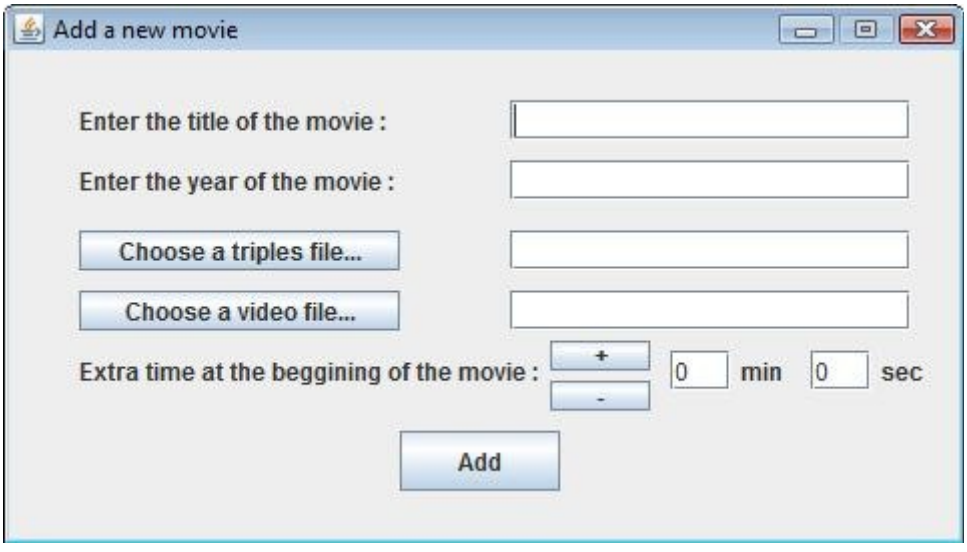
Στο σχήμα 7.1, μπορούμε να δούμε την αρχική οθόνη του συστήματος, που μας δείχνει τις δύο αυτές επιλογές. Στο drop-down menu, εμφανίζονται οι ταινίες που έχουν ήδη προστεθεί.



Σχήμα 7.1 : Αρχική οθόνη του συστήματος

7.1.1 Προσθήκη νέας ταινίας

Επιλέγοντας ο χρήστης την προσθήκη μίας νέας ταινίας, το σύστημα του παρουσιάζει την παρακάτω οθόνη, όπου του ζητά να συμπληρώσει τα απαραίτητα στοιχεία.



Σχήμα 7.2 : Οθόνη προσθήκης νέας ταινίας

Τα στοιχεία που ζητούνται από τον χρήστη είναι τα εξής :

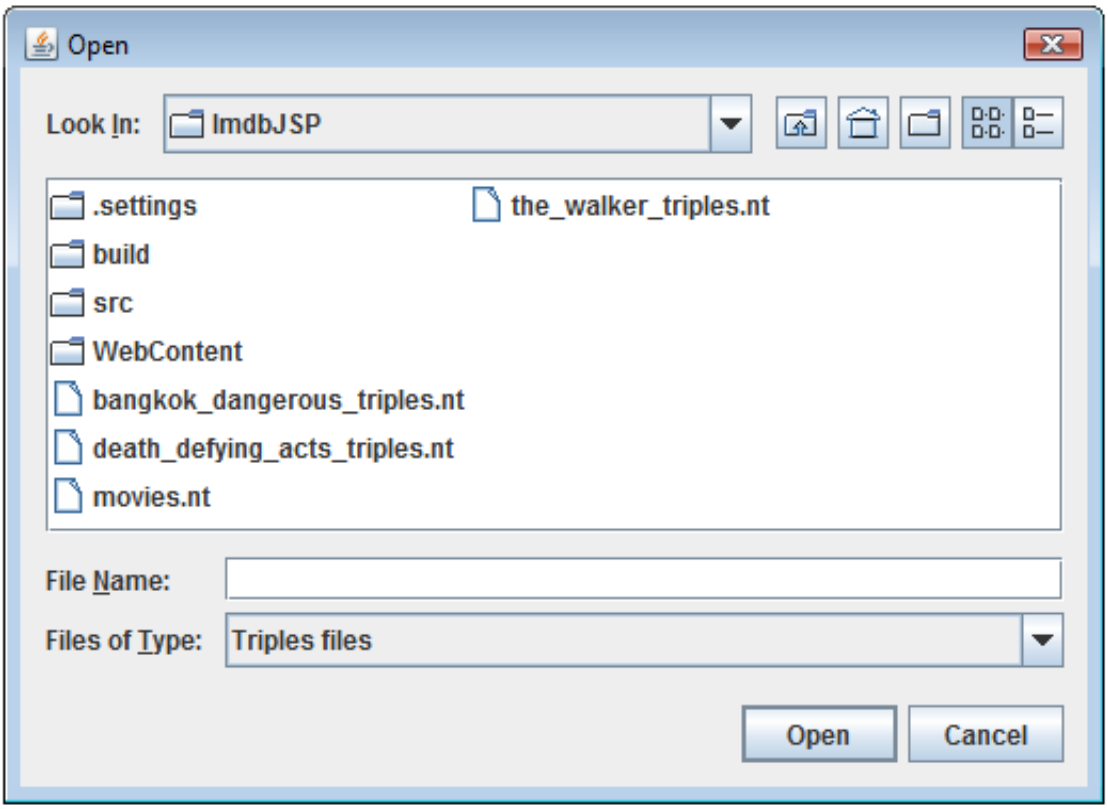
- ο τίτλος της ταινίας
- το έτος κυκλοφορίας της ταινίας
- ένα αρχείο .nt με το σενάριο της ταινίας σε τρίπλες σύμφωνα με την scriptontology
- ένα αρχείο .mp4 με το βίντεο της ταινίας
- η χρονική διαφορά μεταξύ του βίντεο και του σεναρίου

Ο τίτλος της ταινίας παρέχεται για να εντοπιστεί η σωστή σελίδα του imdb που περιέχει τον οδηγό γονέων για την συγκεκριμένη ταινία και για να γίνει η εγγραφή της νέας αυτής ταινίας στο σύστημα.

Πολλές φορές διαφορετικές ταινίες έχουν τον ίδιο τίτλο, όπως συμβαίνει και για μία από τις ταινίες που εξετάσαμε, *The Walker,* που είναι μία από τις τρεις ταινίες και σειρές με το ίδιο όνομα. Γι' αυτόν τον λόγο προστέθηκε ως απαιτούμενο από τον χρήστη και το έτος κυκλοφορίας της ταινίας καθώς είναι μια πληροφορία που εμφανίζεται στα αποτελέσματα μαζί με τους τίτλους και μπορεί να ταυτοποιήσει την ζητούμενη ταινία.

Καθώς το σύστημα πρέπει να εντοπίσει την σελίδα του οδηγού γονέων και να την ανακτήσει, απαραίτητη προϋπόθεση για την λειτουργία αυτή είναι ο υπολογιστής στον οποίον θα τρέξει το σύστημα να είναι συνδεδεμένος στο ίντερνετ.

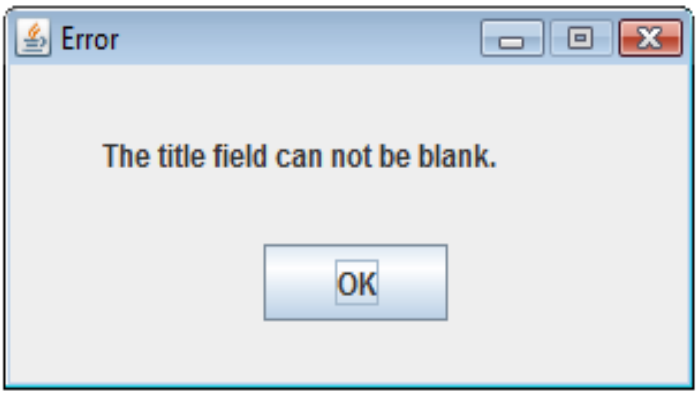
Για ευκολία του χρήστη, τα κουμπιά που είναι δίπλα στην επιλογή των αρχείων .nt και .mp4 δημιουργούν μια οθόνη επιλογής αρχείων, ώστε ο χρήστης να μπορεί να επιλέξει τα επιθυμητά αρχεία.



Σχήμα 7.3 : Οθόνη επιλογής αρχείων

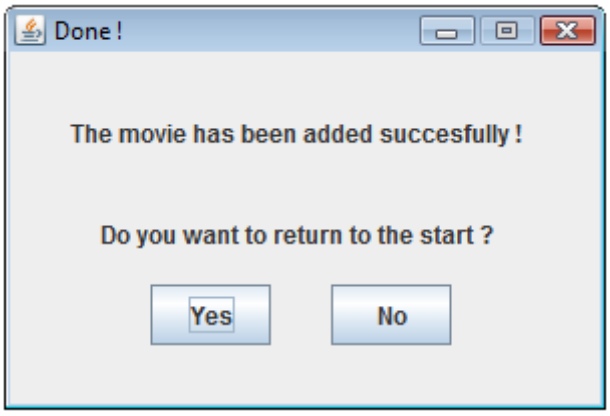
Στο σενάριο, ως αρχή της ταινίας θεωρείται η χρονική στιγμή που είναι σημειωμένη ως η αρχή της πρώτης σκηνής. Τις περισσότερες φορές, η χρονική στιγμή αυτή δεν συμπίπτει με την αντίστοιχη του βίντεο, καθώς στο σενάριο δεν περιλαμβάνονται διαφημίσεις των εταιριών παραγωγής, κενός χρόνος που στο βίντεο έχει μαύρη οθόνη ή ενημερωτικά μηνύματα σχετικά με την παράνομη αντιγραφή και τα πνευματικά δικαιώματα. Επομένως, για να συμβαδίζουν οι χρόνοι έναρξης και λήξης των σκηνών που θα πάρουμε από τις τρίπλες με το βίντεο ώστε να κοπούν σωστά τα αποσπάσματα, χρειαζόμαστε την διαφορά μεταξύ της χρονικής στιγμής που υποτίθεται πως αρχίζει η ταινία σύμφωνα με το σενάριο και της χρονικής στιγμής που όντως αρχίζει στο βίντεο.

Σε περίπτωση που ο χρήστης δεν συμπληρώσει κάποιο από τα απαραίτητα πεδία ή συμπληρώσει κάποιο λάθος, εάν για παράδειγμα το έτος και ο τίτλος της ταινίας δεν συμβαδίζουν, εμφανίζονται οθόνες λάθους με μηνύματα που περιγράφουν τι έχει συμβεί.



Σχήμα 7.4 : Οθόνη λάθους λόγω κενού πεδίου τίτλου

Αφού συμπληρωθούν σωστά τα ζητούμενα πεδία, το σύστημα εμφανίζει μια οθόνη αναμονής, ενώ τρέχει παράλληλα τις διαδικασίες που απαιτούνται για να ολοκληρώσει την προσθήκη της νέας ταινίας. Μόλις ολοκληρωθούν οι διαδικασίες αυτές, το σύστημα ενημερώνει τον χρήστη ότι η ταινία έχει προστεθεί επιτυχώς και του δίνει την δυνατότητα να επιστρέψει στην αρχική σελίδα. Από εκεί μπορεί αν θέλει να συνεχίσει για να δει τις "προβληματικές" σκηνές της ταινίας που μόλις προσέθεσε.



Σχήμα 7.5 : Οθόνη ολοκλήρωσης προσθήκης νέας ταινίας

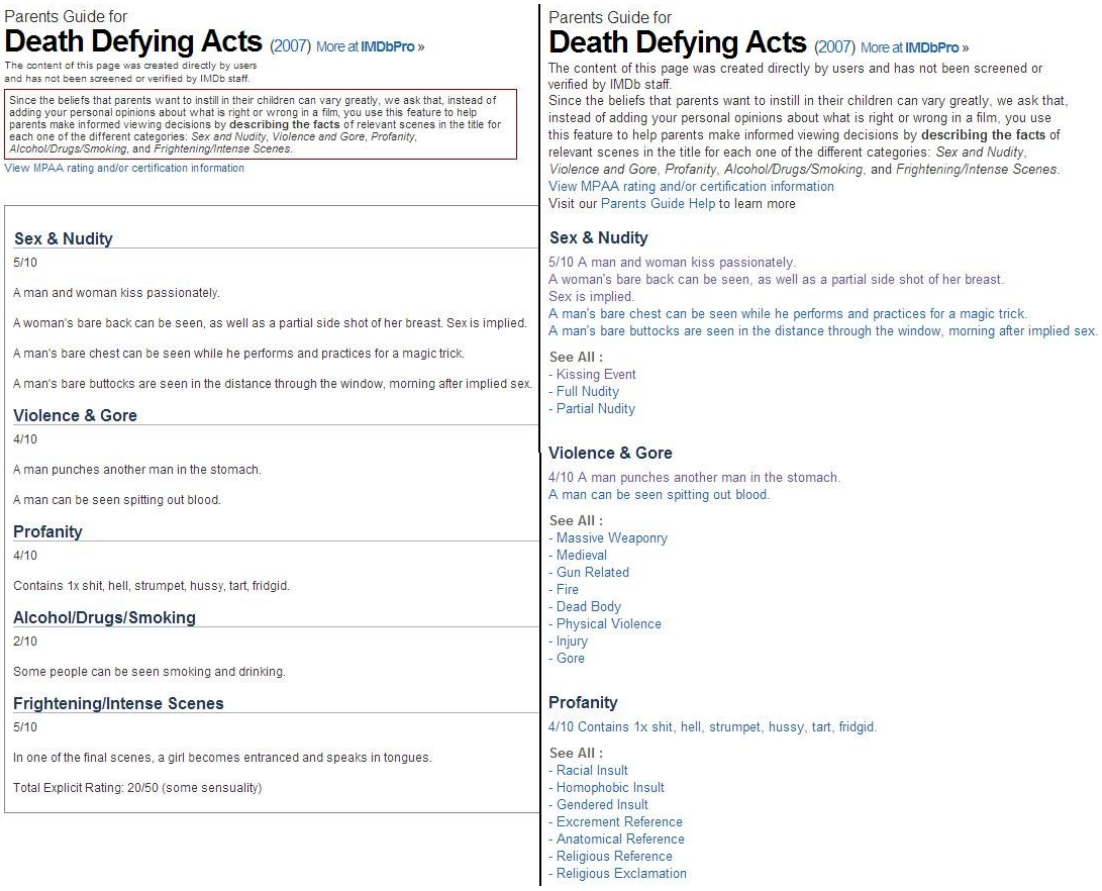
7.1.2 Επιλογή υπάρχουσας ταινίας

Εάν ο χρήστης επιλέξει κάποια από τις ταινίες του drop-down menu της αρχικής οθόνης, το σύστημα ανοίγει ένα παράθυρο στον προεπιλεγμένο πρόγραμμα περιήγησης του υπολογιστή ώστε να παρουσιάσει τις προβληματικές σκηνές της επιλεγμένης ταινίας.

7.1.2.1 Παρουσίαση προβληματικών σκηνών ταινίας

Ο τρόπος που επιλέχθηκε για να παρουσιαστούν τα αποτελέσματα στον χρήστη ήταν μέσω της σελίδας του οδηγού γονέων. Κατά την διαδικασία της προσθήκης της νέας ταινίας, δημιουργείται είναι ένα τροποποιημένο html του οδηγού γονέων με link στις θέσεις των προτάσεων. Τα links αυτά οδηγούν σε html, τα οποία περιέχουν video των σκηνών που το σύστημα έχει αντιστοιχήσει με τις αντίστοιχες προτάσεις.

Ως επιπλέον λειτουργία, στο html αυτό προσθέτονται links για τις κλάσεις της οντολογίας, προσφέροντας στον χρήστη την δυνατότητα να δει το σύνολο των σκηνών που εμπίπτουν σε αυτές, ανεξάρτητα με το αν έχουν περιγραφεί στον οδηγό ή όχι.



Σχήμα 7.6: Αρχικό html parental guide(αριστερά)-τροποποιημένο html (δεξιά)

7.2 Περιγραφή λειτουργίας κλάσεων του συστήματος

Παρακάτω θα δούμε αναλυτικά τις κλάσεις του συστήματος και την λειτουργία τους.

7.2.1 OptionScreen

Η κλάση αυτή δημιουργεί την αρχική οθόνη του συστήματος. Επίσης, δημιουργεί το SailRepository, στο οποίο θα αποθηκεύονται οι τρίπλες RDF του συστήματος και εγκαθιστά την σύνδεση μαζί του.

Για την δημιουργία του drop-down menu, χρησιμοποιεί ένα αρχείο .nt, στο οποίο

είναι αποθηκευμένες οι ταινίες με τον τίτλο και τον κωδικό τους. Οι τρίπλες αυτές προστίθενται στο SailRepository και στη συνέχεια γίνεται SPARQL query για να πάρουμε τους τίτλους ώστε να μπουν στο μενού.

Ανάλογα με την επιλογή του χρήστη, είτε καλεί την *CreateScreen* είτε ανοίγει το πρόγραμμα περιήγησης με το βασικό html της επιλεγμένης ταινίας.

7.2.2 CreateScreen

Η κλάση αυτή δημιουργεί την οθόνη προσθήκης νέας ταινίας. Επίσης, δημιουργεί τις συμπληρωματικές οθόνες επιλογής αρχείων όπου χρειάζονται.

Όταν ο χρήστης πατήσει το κουμπί "OK", ελέγχει εάν όλα τα στοιχεία έχουν συμπληρωθεί σωστά. Εάν έχουν συμπληρωθεί σωστά, βρίσκει την κατάλληλη σελίδα του imdb, συγκεντρώνει τα δεδομένα που έχει συμπληρώσει ο χρήστης και τα προωθεί στην κλάση *WaitingScreen,* την οποία καλεί.

Εάν τα στοιχεία δεν έχουν συμπληρωθεί σωστά, δημιουργεί το κατάλληλο μήνυμα λάθους και καλεί την *ErrorScreen*. Τα λάθη για τα οποίο ελέγχει είναι τα εξής :

- αν κάποιο από τα πεδία τίτλου ή χρονιάς είναι κενό
- αν για τον συνδυασμό τίτλου και χρονιάς δεν υπάρχει κάποια εγγραφή στο imdb
- αν δεν έχει επιλεχτεί .nt ή .mp4 αρχείο
- αν η ταινία που προσπαθούμε να προσθέσουμε υπάρχει ήδη στο σύστημα

Εάν ο χρήστης δεν έχει προσδιορίσει κάποιο χρονικό διάστημα, θεωρούμε ότι δεν υπάρχει κάποια καθυστέρηση ή προήγηση.

7.2.3 ErrorScreen

Η κλάση αυτή δημιουργεί την οθόνη με το μήνυμα λάθους που έχει προκύψει από την *CreateScreen.*

7.2.4 WaitingScreen

Η κλάση αυτή δημιουργεί την οθόνη αναμονής. Καλεί την κλάση *MovieScenesCateg* και όταν αυτή ολοκληρωθεί την *GuideVector*. Όταν και αυτή ολοκληρωθεί, τερματίζει την οθόνη αναμονής και καλεί την κλάση *OKScreen* .

7.2.4 OKScreen

Η κλάση αυτή δημιουργεί την οθόνη ολοκλήρωσης προσθήκης νέας ταινίας. Προσφέρει στον χρήστη την επιλογή να επιστρέψει στην αρχή ή να κλείσει το παράθυρο. Εάν ο χρήστης επιλέξει επιστροφή, καλεί την κλάση *OptionScreen.* Διαφορετικά, τερματίζει την οθόνη.

7.2.5 MovieScenesCateg

Η κλάση αυτή κατηγοριοποιεί όλες τις σκηνές μίας ταινίας, όπως εμφανίζονται στο σενάριο.

Αρχικά, καλεί την *LexiconCreation* δύο φορές, μία για το τμήμα των περιγραφών και μία για το τμήμα του διαλόγου. Έτσι, δημιουργούνται δυο HashMap<String, Integer> που θα περιέχουν τις λέξεις-κλειδιά του εξεταζόμενου τμήματος, στην θέση του String, και την θέση που καταλαμβάνει η λέξη στο αντίστοιχο διάνυσμα χαρακτηριστικών, στην θέση του Integer.

Στη συνέχεια, δημιουργούνται τα διανύσματα χαρακτηριστικών των κλάσεων, καλώντας την κλάση *CategoriesVector*, επίσης δύο φορές. Τα διανύσματα αυτά αποθηκεύονται σε δύο πίνακες, έναν για τον διάλογο και έναν για την περιγραφή.

Τέλος, η κλάση αυτή κάνει SPARQL query στο SailRepository, στο οποίο έχει προσθέσει τις τρίπλες της ταινίας αυτής, ζητώντας τις σκηνές της ταινίας. Για κάθε σκηνή, καλεί την κλάση *SimilarityDis* δύο φορές, προωθώντας το IRI της σκηνής μαζί με τον πίνακα των κλάσεων και το HashMap, του διαλόγου την μία φορά και της περιγραφής την άλλη.

7.2.6 LexiconCreation

Η κλάση αυτή δημιουργεί ένα HashMap με τις λέξεις-κλειδιά του εκάστοτε εξεταζόμενου τμήματος και την θέση που καταλαμβάνει η λέξη στο αντίστοιχο διάνυσμα χαρακτηριστικών.

Παίρνει ως παράμετρο ένα αρχείο .xls, το οποίο προτιμήθηκε για την ευκολία που προσφέρει στην δημιουργία πινάκων. Τα αρχείο αυτό, που χρησιμοποιείται και από την κλάση *CategoriesVector*, περιέχει μία στήλη με τις λέξεις του διανύσματος χαρακτηριστικών και οι υπόλοιπες στήλες συμπληρώνονται από τις κλάσεις που αντιστοιχούν στο διάνυσμα αυτό. Οι στήλες των κλάσεων είναι κενές, εκτός από τις γραμμές που περιέχουν κάποια λέξη-κλειδί της κλάσης που συμπληρώνονται με 1.

Η κλάση αυτή ασχολείται μόνο με την πρώτη στήλη. Για κάθε κελί, παίρνει τις λέξεις που το αποτελούν, τις περνάει από ένα PorterStemmer ώστε να είναι στην ίδια μορφή με το κείμενο, και τοποθετεί τις λέξεις στο HashMap με το αντίστοιχο νούμερο της γραμμής στην οποία βρίσκονται.

7.2.7 CategoriesVector

Η κλάση αυτή δημιουργεί έναν πίνακα με τα διανύσματα χαρακτηριστικών των διάφορων κλάσεων.

Παίρνει ως παράμετρο ένα αρχείο .xls, όπως αυτό που περιγράψαμε προηγουμένως.

Για κάθε μία από τις κλάσεις, ελέγχει την στήλη που τις αντιστοιχεί και συμπληρώνει το διάνυσμα της με 0, εάν το κελί είναι κενό, και με 1, εάν όχι.

7.2.8 SimilarityDis

Η κλάση αυτή κατηγοριοποιεί την εξεταζόμενη σκηνή στις κατηγορίες στις οποίες εμπίπτει.

Αρχικά, η κλάση αυτή καλεί την κλάση *SceneVector,* η οποία δημιουργεί το διάνυσμα χαρακτηριστικών της δεδομένης σκηνής.

Στη συνέχεια, για κάθε ένα από τα διανύσματα των κλάσεων, που βρίσκονται στον πίνακα που έχει περάσει ως παράμετρος, και το διάνυσμα της σκηνής, υπολογίζει τα κοινά 1, κοινά 0 και τους συνδυασμούς 01 και 10 που υπάρχουν ανάμεσα στα δύο αυτά διανύσματα. Με βάση αυτά, υπολογίζει την απόσταση ομοιότητας μεταξύ της σκηνής και κάθε κλάσης.

Τέλος, αν η απόσταση που έχει υπολογιστεί ξεπερνάει το κατώφλι που έχουμε θέσει, το αρχείο .nt τροποποιείται και προστίθεται μια τρίπλα που λέει ότι η δεδομένη σκηνή έχει τύπο την όποια κλάση.

7.2.9 SceneVector

Η κλάση αυτή δημιουργεί το διάνυσμα χαρακτηριστικών διαλόγου ή περιγραφής μίας σκηνής.

Δημιουργείται ένας μονοδιάστατος πίνακας, συμπληρωμένος μόνο με 0, που θα αποτελέσει το διάνυσμα χαρακτηριστικών

Ανάλογα με την περίπτωση, η κλάση αυτή κάνει SPARQL query στο SailRepository,

για να πάρει όλα τα κομμάτια διαλόγου ή περιγραφής που εμπίπτουν στην σκηνή αυτή. Το κάθε ένα από αυτά τα κομμάτια υποβάλλεται σε tokenisation και POS-tagging. Οι λέξεις που έχουν κατάλληλη ετικέτα υποβάλλονται και σε stemming.

Κάθε μία από αυτές τις λέξεις ελέγχεται αν υπάρχει στο HashMap. Αν υπάρχει, ο πίνακας που αναπαριστά το διάνυσμα συμπληρώνεται με 1 στην κατάλληλη θέση.

7.2.10 GuideVector

Η κλάση αυτή αντιστοιχίζει τις σκηνές του σεναρίου με τις περιγραφές και δημιουργεί τα τελικά .html αρχεία που θα χρησιμοποιηθούν για την παρουσίαση των αποτελεσμάτων.

Καλώντας την κλάση *LexiconCreation*, δημιουργεί ένα HashMap<String,Integer> που περιέχει την λίστα με τις λέξεις εξαίρεσης (stop word list). Θα μπορούσε να έχει δημιουργηθεί κάποια άλλη κλάση για την λειτουργία αυτή, αφού δεν μας ενδιαφέρει η θέση της λέξης σε αυτήν την περίπτωση και ο Integer του HashMap είναι ουσιαστικά περιττός. Καθώς όμως το HashMap προσφέρει πολύ εύκολη πρόσβαση στα δεδομένα του και η επιπλέον κόπωση στο σύστημα δεν είναι υπερβολική, δεν θεωρήθηκε απαραίτητο.

Στη συνέχεια, για κάθε μία από τις κατηγορίες του οδηγού καλείται η μέθοδος *sector,* η οποία τροποποιεί το κείμενο της αντίστοιχης κατηγορίας καταλλήλως. Το τροποποιημένο .html της σελίδας του οδηγού γονέων αποθηκεύεται σε ένα αρχείο.

7.2.10.1 Μέθοδος sector

Εάν η μέθοδος αυτή έχει κληθεί για την κατηγορία *Frightening/Intense scenes,* για κάθε πρόταση του τομέα, η μέθοδος καλεί την κλάση *FITfIdf*, που επιστρέφει μία λίστα με τις σκηνές που έχει αντιστοιχήσει με την πρόταση. Στη συνέχεια, καλεί τις

μεθόδους *videoparts,* που αποκόπτει τα κατάλληλα αποσπάσματα βίντεο που αντιστοιχούν στις ζητούμενες σκηνές και την μέθοδο *htmlcreation,* η οποία δημιουργεί το .html στο οποίο θα μεταβαίνει ο χρήστης από την πρόταση αυτή και θα περιέχει τα παραπάνω βίντεο. Τέλος, μετατρέπει το απλό κείμενο της πρότασης σε σύνδεσμο για το αρχείο .html το οποίο δημιουργήθηκε.

Εάν η μέθοδος αυτή έχει κληθεί για οποιαδήποτε άλλη κατηγορία, καλεί την *LexiconCreation* και την *CategoriesVector ,* για να δημιουργηθεί το HashMap και ο πίνακας των διανυσμάτων των κλάσεων αυτής της κατηγορίας αντίστοιχα.

Στη συνέχεια, για κάθε πρόταση του τομέα καλείται η μέθοδος *sim,* που αντιστοιχεί την πρόταση με τις σκηνές του σεναρίου, δημιουργώντας παράλληλα τα κατάλληλα .html, και το κείμενο της πρότασης μετατρέπεται σε σύνδεσμο για το αρχείο .html το οποίο δημιουργήθηκε.

Τέλος, προσθέτονται σύνδεσμοι στο βασικό .html, μετά το κείμενο των περιγραφών, για όλες τις κλάσεις που ανήκουν σε αυτήν την κατηγορία. Για κάθε μία από τις κλάσεις αυτές, γίνεται SPARQL query, ζητώντας τις σκηνές που ανήκουν σε αυτήν, Εάν υπάρχουν τέτοιας σκηνές, προστίθεται ένας σύνδεσμος και δημιουργείται ένα .html με βίντεο των σκηνών αυτών.

7.2.10.2 Μέθοδος sim

Η μέθοδος αυτή κατηγοριοποιεί την δεδομένη πρόταση στις κλάσεις στις οποίες ανήκει.

Η πρόταση υποβάλλεται σε tokenisation και POS-tagging. Για κάθε λέξη ελέγχεται, αν η ετικέτα που έχει πάρει είναι δεκτή. Εάν ναι, ακολουθεί stemming. Στη συνέχεια, ελέγχεται εάν η λέξη ανήκει στις λέξεις-κλειδιά και αν ανήκει, το διάνυσμα

χαρακτηριστικών αποκτά ένα 1 στην αντίστοιχη θέση. Ελέγχεται επίσης εάν ανήκει στην λίστα με τις λέξεις εξαίρεσης. Εάν δεν ανήκει, προστίθεται στην λίστα που θα χρησιμοποιηθεί από την κλάση *TfIdf*.

Υπολογίζεται το μέτρο ομοιότητας με κάθε υποψήφια κλάση και η πρόταση-σκηνή κατηγοριοποιείται αντίστοιχα. Για κάθε κλάση στην οποία ανήκει η πρόταση, καλείται η κλάση *TfIdf,* η οποία επιστρέφει μία λίστα με τις σκηνές που αντιστοιχίζει στην πρόταση.

Τέλος, καλούνται οι μέθοδοι *videoparts* και *htmlcreation*, για να παράξουν τα αποσπάσματα βίντεο των σκηνών με τις οποίες έχει αντιστοιχηθεί η πρόταση και το αντίστοιχο .html.

7.2.10.3 Μέθοδος videoparts

Η μέθοδος αυτή δημιουργεί αποσπάσματα βίντεο για μία λίστα σκηνών που της δίνεται ως παράμετρος.

Για κάθε σκηνή, χρησιμοποιώντας το IRI της κάνει SPARQL query ώστε να πάρει τις χρονικές στιγμές έναρξης και λήξης. Σε αυτές προσθέτει ή αφαιρεί τον χρόνο που έχει προσδιορίσει ο χρήστης και με τη βοήθεια του εργαλείου *ffmpeg,* αποκόπτει το αρχικό αρχείο .mp4 για να πάρει το κατάλληλο απόσπασμα.

7.2.10.4 Μέθοδος htmlcreation

Η μέθοδος αυτή δημιουργεί το .html αρχείο που θα περιέχει τα βίντεο των σκηνών.

7.2.11 FITfIdf

Η κλάση αυτή αντιστοιχεί μία πρόταση της κατηγορίας *Frightening/Intense scenes* με σκηνές του σεναρίου.

Η πρόταση υποβάλλεται σε tokenisation και POS-tagging. Για κάθε λέξη ελέγχεται, αν η ετικέτα που έχει πάρει είναι δεκτή. Εάν ναι, ακολουθεί stemming και γίνεται έλεγχος εάν ανήκει στην λίστα με τις λέξεις εξαίρεσης. Εάν δεν ανήκει, προστίθεται στην λίστα έρευνας με βάση την οποία θα βρεθούν τα βάρη tf-idf.

Δημιουργούνται τρεις πίνακες, ένας μονοδιάστατος για τα idf που θα υπολογιστούν, και δύο δυδιάστατοι για τα tf και τα βάρη tf-idf που θα προκύψουν. Γίνεται SPARQL query για να βρεθούν οι σκηνές της ταινίας. Για κάθε σκηνή, καλείται η μέθοδος *matchcount,* στην οποία θα υπολογιστούν τα αντίστοιχα tf και θα ενημερωθούν τα idf.

Στη συνέχεια, υπολογίζονται τα βάρη και στο τέλος, υπολογίζεται η απόσταση συνημιτόνου μεταξύ της πρότασης και κάθε υποψήφιας σκηνής. Εάν η απόσταση ξεπερνάει ένα όριο, η σκηνή αντιστοιχίζεται με την πρόταση και προστίθεται στην λίστα των σκηνών που θα επιστρέψει η κλάση.

7.2.11.1 Μέθοδος matchcount

Η μέθοδος αυτή ενημερώνει τους πίνακες tf και idf για κάθε σκηνή που εξετάζεται.

Γίνεται δύο SPARQL queries για την σκηνή ώστε να πάρουμε όλα τα κομμάτια διαλόγου και περιγραφών της σκηνής. Το κάθε κομμάτι υποβάλλεται στην γνωστή επεξεργασία φυσικής γλώσσας και αν κάποια από τις λέξεις της λίστας έρευνας εμφανίζεται στο κομμάτι αυτό, το αντίστοιχο tf και idf ενημερώνονται καταλλήλως.

7.2.12 TfIdfMatch

Η κλάση αυτή αντιστοιχεί μία πρότασης κάποιας κατηγορίας, εκτός από την *Frightening/Intense scenes,* με σκηνές του σεναρίου, που ανήκουν στην ίδια κλάση με αυτή.

Η δομή και η λειτουργία της κλάσης αυτής είναι εξαιρετικά παρόμοια με την προηγούμενη κλάση. Για να μην επαναλαμβάνουμε κείμενο ανούσια, θα αναφέρουμε τις διαφορές και παραπέμπουμε στο 7.2.11 για περισσότερες λεπτομέρειες.

Η μοναδική διαφορά που υπάρχει μεταξύ των δύο αυτών κλάσεων είναι ότι η μελέτη tf-idf δεν γίνεται στο σύνολο των σκηνών της ταινίας, αλλά περιορίζεται στις σκηνές που ανήκουν στην ίδια κλάση με την εξεταζόμενη πρόταση.

7.3 Βιβλιοθήκες και εργαλεία που χρησιμοποιήθηκαν

Το σύστημα υλοποιήθηκε σε γλώσσα Java, χρησιμοποιώντας το περιβάλλον Eclipse. Όλα τα γραφικά δημιουργήθηκαν με χρήση της ενσωματωμένης βιβλιοθήκης Swing.

Οι εξωτερικές βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι εξής :

- *apache.poi-3.11[14]* : χρησιμοποιήθηκε για την επικοινωνία με τα αρχεία .xls στα οποία ήταν αποθηκευμένα τα διάφορα διανύσματα χαρακτηριστικών καθώς και η λίστα με τις λέξεις εξαίρεσης.
- *jsoup-1.9.1[15]* : χρησιμοποιήθηκε για την εξαγωγή δεδομένων από αρχεία .html και την μεταποίηση στοιχείων των αρχείων αυτών.
- *opennlp.tools-1.5.3[16]*: χρησιμοποιήθηκε για την επεξεργασία φυσικής γλώσσας και συγκεκριμένα για τον διαχωρισμό λέξεων και για την ανάθεση ετικετών ως προς το μέρος του λόγου στις λέξεις.
- *lucene.snowball-3.0.2[17]* : χρησιμοποιήθηκε για την επεξεργασία φυσικής γλώσσας και συγκεκριμένα για την αποκοπή καταλήξεων.
- *openrdf-sesame-2.7.11[18]* : χρησιμοποιήθηκε για την αποθήκευση και την επικοινωνία, χρησιμοποιώντας SPARQL, με τις τρίπλες RDF.

Επιπλέον εργαλεία που χρησιμοποιήθηκαν :

- *en-pos-maxent.bin, en-token.bin* : για την λειτουργία της βιβλιοθήκης

opennlp.tools χρειάστηκε η υποστήριξη από αυτά τα ήδη εκπαιδευμένα μοντέλα για να εκτελεστούν οι λειτουργίες που ήδη αναφέραμε.

- *ffmpeg[19]* : χρησιμοποιήθηκε για την αποκοπή κομματιών από το βίντεο της ταινίας που αντιστοιχούν στις σκηνές που θέλουμε να παρουσιάσουμε.

8

Επίλογος

8.1 Σύνοψη και συμπεράσματα

Στόχος της διπλωματικής εργασίας ήταν η δημιουργία μιας εφαρμογής, που θα επιτύγχανε να εντοπίσει τις "προβληματικές" σκηνές μιας ταινίας και να τις αντιστοιχίσει με τις συγκεκριμένες περιγραφές του οδηγού γονέων, δίνοντας την δυνατότητα στον χρήστη να μπορεί να μεταβεί άμεσα από τις περιγραφές στις ίδιες τις "προβληματικές" σκηνές.

Κατά την διαδικασία ανάπτυξης της εφαρμογής αυτής, έγινε ανάπτυξη οντολογίας, εξαγωγή λέξεων-κλειδιά, έρευνα αποτελεσμάτων μέτρων ομοιότητας και εύρεση παρόμοιων όρου κειμένου.

Ο στόχος της διπλωματικής επιτεύχθηκε κατά το μεγαλύτερο μέρος. Τα ποσοστά επιτυχίας και για το μεσαίο στάδιο των κατηγοριοποιήσεων και για τις τελικές αντιστοιχήσεις είναι πάνω από 70%, με μοναδική εξαίρεση τα τελικά αποτελέσματα της ταινίας *Death Defying Acts* που πέφτουν στο 55%.

Αν και τα αποτελέσματα της κατηγορίας *Frightening/Intense scenes* είναι ιδιαιτέρως

χαμηλά, δεν το θεωρούμε αποτυχία, λόγω της υπερβολικής δυσκολίας των συγκεκριμένων δεδομένων.
Δόθηκε ιδιαίτερη προσοχή, ιδιαίτερα στην εξαγωγή της οντολογίας και στην εξαγωγή λέξεων-κλειδιών, ώστε η εφαρμογή να είναι όσο πιο περιεκτική γίνεται και να μην αναπτυχθεί γύρω από τα δεδομένα των ταινιών που είχαμε. Λόγω του μικρού αριθμού των ταινιών προς εξέταση όμως, είναι αρκετά πιθανό τα κατώφλια που επιλέχτηκαν να μην αποδειχθούν τα καλύτερα, εξετάζοντας ένα μεγαλύτερο δείγμα.

8.2 Μελλοντικές επεκτάσεις

Επεκτάσεις που μπορούν να γίνουν στην εφαρμογή :

- ανεξαρτητοποιώντας την εφαρμογή από το IMDB, μπορούμε, επιπρόσθετα στις πληροφορίες που παίρνουμε από αυτήν την σελίδα, να προσθέσουμε και άλλες πηγές. Μια πιθανή ιστοσελίδα είναι η *commonsensemedia.org*, που προσφέρει πληροφορίες για τις επιλήψιμες σκηνές σε παρόμοιο πλαίσιο με αυτό του οδηγού γονέων του IMDB, καθώς και κριτικές χρηστών. Επίσης, μπορούν να χρησιμοποιηθούν διάφορες σελίδες που διατηρούνται από χρήστες με ειδικό προσανατολισμό σε κάποιο θέμα όπως η *istheresuicideinit.tumblr.com*, για εμφανίσεις αυτοκτονίας στις ταινίες, *self-injury.net/media/movie*, για αυτοτραυματισμό και *isitconsensual.tumblr.com,* για βιασμό.
- περαιτέρω επέκταση της οντολογίας που να αφορά επιπλέον ανεξάρτητα στοιχεία του περιεχομένου (τόπος που η σκηνή λαβαίνει χώρα, αντικείμενα που εμφανίζονται, κ.λ.π.) και περίληψη τους στην διαδικασία της αντιστοίχησης σκηνών.

9

Βιβλιογραφία

[1] “IMDb”, http://www.imdb.com.

[2] “RDF 1.1 Concepts and Abstract Syntax ”, https://www.w3.org/TR/rdf11-concepts/

[3] “RDF Schema 1.1 ”, https://www.w3.org/TR/rdf-schema/

[4] “SPARQL Query Language for RDF”, https://www.w3.org/TR/rdf-sparql-query/

[5] “OWL Web Ontology Language Overview ”, https://www.w3.org/TR/owl-features/

[6] “Tokenizer (Apache OpenNLP Tools 1.5.2-incubating API) ”, https://opennlp.apache.org/documentation/1.5.2-incubating/apidocs/opennlp-tools/opennlp/tools/tokenize/Tokenizer.html

[7] Porter, M.f. "An Algorithm for Suffix Stripping." *Program: Electronic Library and Information Systems* 14.3 (1980): 130-37. Web.

[8] S. Choi, S. Cha, and C. C. Tappert, "A Survey of Binary Similarity and Distance Measures," Journal of Systemics, Cybernetics and Informatics, Vol 8, No 1, 2010, pp 43-48.

[9] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze.*Introduction to Information Retrieval*. New York: Cambridge UP, 2008. Print.

[10] “IMDb Top 250 - IMDb”, http://www.imdb.com/chart/top

[11] Fellbaum, Christiane. *WordNet: An Electronic Lexical Database*. Cambridge, Mass: MIT, 1998. Print.

[12] “Notes on WordNet Domains”, http://courses.washington.edu/englhtml/wndomains_notes.html

[13] “Penn Treebank P.O.S. Tags ”, https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

[14]“Apache POI, ”https://poi.apache.org

[15] “Jsoup”, https://jsoup.org/

[16] “Apache OpenNLP”, https://opennlp.apache.org/

[17] “Apache Lucene” , http://lucene.apache.org/

[18] “Sesame”, http://www.openrdf.org/

[19] “Ffmpeg”, https://ffmpeg.org/