ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Εισηγητικό Σύστημα για Ταινίες με βάση το Περιεχόμενο και Έμφαση στους Υποτίτλους

Διπλωματικη Εργασια

του

ΜΠΟΥΓΙΑΤΙΩΤΗ Χ. ΚΩΝΣΤΑΝΤΙΝΟΥ

Επιβλέπων: Κόλλιας Στέφανος
Καθηγητής, Ε.Μ.Π.

Αθήνα, Οκτώβριος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Εισηγητικό Σύστημα για Ταινίες με βάση το Περιεχόμενο και Έμφαση στους Υποτίτλους

Διπλωματικη Εργασια

του

ΜΠΟΥΓΙΑΤΙΩΤΗ Χ. ΚΩΝΣΤΑΝΤΙΝΟΥ

Επιβλέπων: Κόλλιας Στέφανος
Καθηγητής, Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την η Οκτώβρη 2015

.....	.....	.....
Κόλλιας Στέφανος	Στάμου Γεώργιος	Σταφυλοπάτης Ανδρέας Γεώργιος
Καθηγητής, Ε.Μ.Π.	Καθηγητής, Ε.Μ.Π.	Καθηγητής, Ε.Μ.Π.

Αθήνα, Οκτώβριος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

.....

Κωνσταντίνος Χ. Μπουγιατιώτης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Ηλεκτρονικών Υπολογιστών

Περίληψη

Στόχος της παρούσας εργασίας είναι η μελέτη μεθόδων ανάλυσης θεματικού περιεχομένου των υποτίτλων καθώς και των ηχητικών χαρακτηριστικών των ταινιών, για την υλοποίηση ενός εισηγητικού συστήματος για ταινίες. Εφαρμόζοντας μεθόδους επεξεργασίας φυσικής γλώσσας και topic modeling αλγορίθμων στους υπότιτλους, εξαγάγουμε τη θεματική δομή των ταινιών και αποφαινόμαστε για την ομοιότητα μεταξύ τους, με βάση την αναπαράσταση αυτών στο θεματικό χώρο. Παράλληλα, χρησιμοποιούμε και διάφορες μεθόδους ανάλυσης των ηχητικών σημάτων των ταινιών που μας δίνουν πληροφορίες για το είδος της μουσικής και τα ηχητικά συμβάντα σε αυτές. Με τις νέες αυτές αναπαραστάσεις για τις ταινίες, επεκτείνουμε το σύστημα μας χρησιμοποιώντας διάφορες μεθόδους σύντηξης των διαφορετικών μοντέλων, για βελτιωμένα αποτελέσματα.

Προς επιβεβαίωση των μεθόδων αυτών, υλοποιείται ένα πλήρες σύστημα εισηγήσεων με τις παραπάνω τεχνικές και παρατηρούνται οι επιδόσεις του. Για το σκοπό αυτό, κατασκευάσαμε ένα dataset από 160 γνωστές ταινίες και εργαστήκαμε με αυτό. Διερευνήθηκαν οι διάφοροι τρόποι αξιολόγησης της ποιότητας των εισηγήσεων και δημιουργήθηκαν κατάλληλα μέτρα απόδοσης για το σύστημά μας. Τα πειραματικά αποτελέσματα αποδεικνύουν ότι η αναπαράσταση των ταινιών σε θέματα οδηγεί σε ποιοτικές εισηγήσεις και ότι η σύντηξη των διαφορετικών πηγών πληροφορίας, ωθεί το σύστημα σε βελτιωμένα αποτελέσματα. Συνοψίζοντας, το σύστημα εισηγήσεων που δημιουργήσαμε όχι μόνο παρέχει ικανοποιητικές συστάσεις και αναπαραστάσεις των ταινιών σε διάφορους χώρους πληροφορίας, αλλά προωθεί και τη καινοτόμα ιδέα της πλοήγησης στο χώρο των θεμάτων σε συσχετισμό με τις ταινίες.

Λέξεις Κλειδιά

εισηγητικό σύστημα, επεξεργασία υποτίτλων, εξαγωγή θεμάτων, Latent Dirichlet Allocation, ανάλυση ήχου ταινίας, πολυτροπική σύντηξη δεδομένων, ανάκτηση πληροφορίας

Abstract

This diploma thesis aims to create a content-based recommender system for movies, with emphasis on the topic representations of the subtitles, as well as cues stemming from the audio channel of the movie. Using natural language processing and topic modeling algorithms on the subtitles, we extract the latent topic structure of the movies and assert their similarities exploiting their topic representation distances. Moreover, using audio analysis techniques on the movies' audio clips, we represent the movies according to their distributions over the different kind of music and audio events. Finally, using different schemes of multimodal fusion between the different representation models, we manage to improve our results.

We implemented a complete recommendation system based on these techniques, and assessed its' performance. To this end we compiled a dataset of 160 well-known movies to work with. We analyze the different ways of evaluating the recommendations and we propose a new evaluation metric with emphasis on the recommendations' rankings. Experimental results show that topic models provide quality recommendations and that multimodal fusion achieves performance boosting related to the individual modalities. In summary, the proposed content based system did not only give high-grade recommendations and alternative movie representations, but also nurtures the novel idea of movie browsing and similarity discovery through the topic space.

Keywords

recommender system, subtitles processing, topic modeling, Latent Dirichlet Allocation, text mining, movie audio analysis, multimodal fusion, information retrieval

Ευχαριστίες

Αρχικά, να τονιστεί ότι η παρούσα εργασία προέκυψε από τη συνεργασία μεταξύ του Εργαστηρίου Ψηφιακής Επεξεργασίας Εικόνας, Βίντεο και Πολυμέσων, του τομέα Τεχνολογίας Πληροφορικής & Υπολογιστών της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Ε.Μ.Π και του Εργαστηρίου Υπολογιστικής Ευφυίας, του Ινστιτούτου Πληροφορικής και Τηλεπικοινωνιών, του Ε.Κ.Ε.Φ.Ε "Δημόκριτος".

Ολοκληρώνοντας αυτή την προσπάθεια λοιπόν, θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κ. Κόλλια Στέφανο για την εμπιστοσύνη που μου έδειξε αναλαμβάνοντας και επιβλέποντας αυτή τη διπλωματική εργασία. Ιδιαίτερα, θα ήθελα να ευχαριστήσω τον Δρ. Γιαννακόπουλο Θοδωρή για την ευκαιρία που μου έδωσε να εκπονήσω την εργασία αυτή στο Εργαστήριο Υπολογιστικής Ευφυίας, την συνεχή καθοδήγησή του και την πολύτιμη συνεργασία του.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου και την αδερφή μου (Χρήστος, Αρετή, Πωλίνα) και τους φίλους μου, για την συνεχή συμπαράσταση και διαρκή υποστήριξή τους.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικά

Η σύγχρονη εποχή ονομάζεται εποχή της πληροφορίας και όχι άδικα, καθώς έχουμε τη δυνατότητα πρόσβασης και ανταλλαγής πληροφορίας σε τέτοιο βαθμό, που ήταν αδύνατο να διανοηθεί κανείς στο παρελθόν. Το έναυσμα για αυτή την εποχή αποτέλεσε η ψηφιακή επανάσταση και εδραιώθηκε με την έλευση του διαδικτύου και του παγκόσμιου ιστού, που μας παρέχουν πρόσβαση σε ένα τεράστιο αριθμό δεδομένων. Σε αυτή την εποχή λοιπόν με την πληθώρα πληροφοριών και πολυμεσικών δεδομένων, έκαναν την εμφάνισή τους τα *εισηγητικά συστήματα* ή *συστήματα συστάσεων*(recommender ή recommendation systems). Τα εισηγητικά συστήματα αποτελούν τεχνικές που προτείνουν στο χρήστη να αλληλεπιδράσει με κάποιο αντικείμενο, θεωρώντας ότι μπορεί να τον ενδιαφέρει. Για παράδειγμα, όταν ψάχνουμε για κάποιο βιβλίο σε κάποια πλατφόρμα αγοράς αγαθών, όπως το *ebay*[1], μας προτείνονται σχετικά βιβλία με το συγκεκριμένο.

Τα εισηγητικά συστήματα δηλαδή απευθύνονται κυρίως σε άτομα που μπορεί να μην γνωρίζουν πολλά πράγματα ή να μην έχουν χρόνο και υπομονή, ώστε να αξιολογήσουν την πιθανώς ατελείωτη λίστα με σχετικά αντικείμενα που μπορεί να αφορούν κάτι που ψάχνουν. Έτσι, προσφέροντας στοχευμένες και εξατομικεύμενες, σε πολλές περιπτώσεις, συστάσεις σε κάποιο χρήστη, του επιλύουν το παραπάνω πρόβλημα και βελτιώνουν την αλληλεπίδρασή του με το σύστημα με το οποίο ασχολείται. Τα ίδια ισχύουν και πιο ειδικά όταν ο τομέας ασχολίας του ατόμου είναι η επιλογή ταινιών. Λόγω της μεγάλης γκάμας των ταινιών που υπάρχουν και το εύρος των διαφορετικών ειδών, είναι πολύ δύσκολο να επιλέξει κανείς ταινία, αν του παρου-

[1] http://www.ebay.com/

σιαστούν όλες μαζεμένες. Προς το σκοπό αυτό έχουν δημιουργηθεί συστήματα συστάσεων ειδικά για ταινίες, ώστε να ταιριάζουν τις απαιτήσεις του εκάστοτε καταναλωτή χωρίς αυτός να χρειαστεί να ψάξει ένα τεράστιο πλήθος ταινιών.

Πάρα πολλές ιστοσελίδες, όπως η RottenTomatoes[2], ή υπηρεσίες που προσφέρουν streaming ταινιών, όπως το Netflix[3], έχουν υλοποιήσει συστήματα εισηγήσεων για ταινίες. Όμως για να μπορέσει ένα σύστημα να εισηγηθεί αυτοματοποιημένα ταινίες αναλόγως με το πόσο ταιριάζουν με μία συγκεκριμένη ταινία, πρέπει πρώτα να έχει μία κατανόηση των ταινιών των ίδιων. Αυτό, σχεδόν σε όλες τις περιπτώσεις, το πετυχαίνουν τα υπάρχοντα συστήματα είτε μέσω collaborative filtering, δηλαδή σύμφωνα με τις προτιμήσεις χρηστών που έχουν δει την συγκεκριμένη ταινία και έχουν εκφράσει την αρέσκειά τους για κάποιες άλλες ταινίες, είτε μέσω content based μεθόδων, όπου κάθε ταινία εκφράζεται ως ένας συνδυασμός μεταδεδομένων που αφορούν την ταινία, όπως ποιος είναι ο σκηνοθέτης, το είδος της ταινίας, οι ηθοποιοί που εμφανίζονται κ.ο.κ, είτε τέλος μέσω υβριδικών μεθόδων των δύο προηγουμένων κατηγοριών[4].

Όμως οι παραπάνω μέθοδοι βασίζονται στη γνώμη των χρηστών είτε για να εξάγουν σχέσεις συνάφειας μεταξύ των ταινιών, είτε για το προσδιορισμό αυτών των μεταδεδομένων όπως για παράδειγμα το είδος της ταινίας. Δεν ασχολούνται δηλαδή με αυτό καθ'αυτό το περιεχόμενο της ταινίας, δηλαδή την εικόνα, τον ήχο και το κείμενο, αλλά ετεροπροσδιορίζουν κάθε ταινία σύμφωνα με την ανθρώπινη γνώμη για αυτές. Ωστόσο, θα είχε μεγάλο ενδιαφέρον ένα σύστημα εισήγησης ταινιών που δε θα βασιζόταν σε πληροφορίες έξωθεν των ταινιών, αλλά θα εξήγαγε συμπεράσματα μονάχα από τα στοιχεία που την απαρτίζουν, δηλαδή την εικόνα(βίντεο), τον ήχο και το κείμενο. Αυτή ακριβώς η διαφορετική προσέγγιση εξετάζεται στη παρούσα διπλωματική εργασία.

1.2 Συνεισφορά της Εργασίας

Σκοπός της εργασίας λοιπόν, όπως σκιαγραφήθηκε προηγουμένως, είναι η δημιουργία ενός εισηγητικού συστήματος για ταινίες, βασισμένο στο ουσιαστικό περιεχόμενο της ταινίας. Πιο συγκεκριμένα, στο πλαίσιο της εργασίας μας ασχολούμαστε κατά βάση με το κειμενικό μέρος

[2]http://www.rottentomatoes.com/
[3]https://www.netflix.com/
[4]Περισσότερα για αυτά στη συνέχεια (Κεφάλαιο **2**)

των ταινιών και δευτερευόντως με το ηχητικό σκέλος. Η προσφορά της εργασίας μας λοιπόν προς αυτό το σκοπό είναι πολύπλευρη.

Αρχικά, σχεδιάσαμε και υλοποιήσαμε μία αλυσίδα μεθόδων επεξεργασίας υποτίτλων, έτσι ώστε να καταλήξουν σε μορφή κατάλληλη είτε για εξαγωγή θεμάτων, είτε για εφαρμογή μετασχηματισμών, όπως ο tf-idf, ή για οποιαδήποτε άλλη μεθοδολογία που βασίζεται στην bag-of-words αναπαράσταση των κειμένων. Ακολούθως, διερευνήθηκε η διαδικασία topic modeling από αυτά τα δεδομένα, χρησιμοποιώντας τη μέθοδο *Latent Dirichlet Allocation*, όπου μελετήθηκαν οι διάφορες πτυχές και ιδιαιτερότητές της και τις οποίες παρουσιάζουμε αναλυτικά.

Επίσης, αναλύουμε συνοπτικά ένα πλαίσιο εξαγωγής ηχητικών χαρακτηριστικών και ειδών μουσικής για κάθε ταινία, που θα μας προσφέρει πληροφορίες για την ομοιότητα των ταινιών οι οποίες δεν είναι διαθέσιμες μέσω των υποτίτλων. Στο πλαίσιο χρησιμοποίησης αυτών των χαρακτηριστικών, παρουσιάζονται και δύο μέθοδοι σύντηξης πληροφορίας που προέρχεται από τις θεματικές αναπαραστάσεις των ταινιών και από αυτά τα ηχητικά χαρακτηριστικά και παρουσιάζεται η μεγάλη χρησιμότητα που προσφέρει αυτή η σύντηξη.

Επιπροσθέτως, μελετάται και παρουσιάζεται μία διαδικασία ανάκτησης πληροφορίας από τις αναπαραστάσεις των ταινιών στους διαφορετικούς χώρους πληροφορίας. Αντιπαραβάλλονται διάφορα μετρικά και ιδέες που τα προωθούν και παρουσιάζεται και ένα νέο μετρικό, που μπορεί να χρησιμοποιηθεί ως μέσω αξιολόγησης ενός εισηγητικού συστήματος το οποίο δίνει έμφαση στη σειρά κατάταξης των εισηγήσεων.

Προς ολοκλήρωση των παραπάνω έγινε πειραματική υλοποίηση των περιγραφόμενων μεθόδων σε μία μικρή βάση ταινιών, ώστε να αποδειχθεί η ορθότητα των ισχυρισμών. Πέρα από τη παρουσίαση των πειραματικών αποτελεσμάτων και των συμπερασμάτων που προκύπτουν από αυτά, έγινε και η καταγραφή μελλοντικών επεκτάσεων του συστήματος και ιδεών που θα οδηγούσαν σε δυνητικά καλύτερα αποτελέσματα.

Συνοψίζοντας οι δύο πιο σημαντικές συνεισφορές αυτής της εργασίας είναι οι εξής:

- Αφ'ενός, η δημιουργία ενός νέου είδους εισηγητικού συστήματος για ταινίες, το οποίο βασίζεται κυρίως στο θεματικό περιεχόμενο των ταινιών και δευτερευόντως σε πληροφορίες από τα ηχητικά τους χαρακτηριστικά για να κάνει κατάλληλες εισηγήσεις.

- Αφ'ετέρου, η διαφοροποίηση στο τρόπο θέασης των ταινιών, καθώς μας δίνεται η δυνατότητα με αυτή την υλοποίηση να πλοηγηθούμε στο θεματικό χώρο που ανήκουν οι ταινίες και να δούμε τις συσχετίσεις μεταξύ αυτών και των θεμάτων που τις αποτελούν. Ανοίγει έτσι ο δρόμος για μία ποιοτική αναπαράσταση των ταινιών ως οντότητες με πολλές διαφορετικές πληροφορίες, εδώ αναλύσαμε μόνο στοιχεία περιεχομένου(ήχος και κείμενο) αλλά θα μπορούσε να επεκταθεί και με μεταδεδομένα ή προτιμήσεις χρηστών σχετικά με αυτές, που θα μας οδηγήσουν σε μία νέα, πιο ολιστική θεώρηση των ταινιών και των συσχετίσεων μεταξύ αυτών.

1.3 Διάρθρωση της Εργασίας

Η εργασία οργανώνεται σε κεφάλαια με την ακόλουθη δομή:

⋆ **Κεφάλαιο 2**: Σε αυτό το κεφάλαιο γίνεται μία περιεκτική περιγραφή της υπάρχουσας βιβλιογραφίας σχετικά με το αντικείμενο που πραγματεύεται η εργασία.

⋆ **Κεφάλαιο 3**: Το κεφάλαιο αυτό περιέχει όλη την απαραίτητη θεωρητική περιγραφή για την ανάλυση του κειμένου και του ήχου, καθώς και τη σύντηξη αυτών των δύο πηγών πληροφορίας. Επίσης, αναφερόμαστε στον τρόπο παραγωγής των πινάκων ομοιότητας μεταξύ των ταινιών.

⋆ **Κεφάλαιο 4**: Περιγράφεται πλήρως η διαδικασία υλοποίησης και όλες οι τεχνικές λεπτομέρειες αυτής. Ακολούθως, αναλύουμε λεπτομερώς τη διαδικασία ανάκτησης πληροφορίας και τα μέτρα απόδοσης που χρησιμοποιούμε για το σύστημά μας. Τέλος, παρουσιάζουμε τα πειραματικά αποτελέσματα και σχολιάζουμε τα συμπεράσματα που προκύπτουν από αυτά.

⋆ **Κεφάλαιο 5**: Συνοψίζονται τα συμπεράσματα και η συνεισφορά της εργασίας μας και περιγράφονται μελλοντικές επεκτάσεις της παρουσιαζόμενης μεθόδου.

Κεφάλαιο 2

Σχετική Βιβλιογραφία

2.1 Εισηγητικά Συστήματα

Τα *εισηγητικά συστήματα ή συστήματα συστάσεων*(recommender ή recommendation systems) είναι μοντέλα και τεχνικές επεξεργασίας πληροφορίας που σκοπό έχουν να προτείνουν κατάλληλα "άντικείμενα" σε χρήστες, προβλέποντας τις προτιμήσεις τους σχετικά με αυτά τα "άντικείμενα"[85, 82]. Ως "άντικείμενο" νοείται οτιδήποτε ο χρήστης μπορεί να επιλέξει από μία συλλογή ειδών, όπως για παράδειγμα ταινίες, βιβλία ή ιστοσελίδες. Μπορούν να θεωρηθούν εφαρμογές της αναγνώρισης χαρακτηριστικών, αλλά στο πλαίσιο της δημιουργίας εξατομικευμένων προτάσεων για κάθε χρήστη. Αποτελούν απαραίτητη τεχνολογία στη σημερινή εποχή με την πληθώρα διαφορετικών επιλογών που προσφέρονται σε μία μεγάλη γκάμα προϊόντων.

Στην πιο απλή τους και συνηθισμένη μορφή οι εισηγήσεις αυτές παρουσιάζονται στο χρήστη ως μία ταξινομημένη λίστα "άντικειμένων", με τη σειρά κατάταξης τους να δείχνει ποιά από αυτά τα "άντικείμενα" ταιριάζουν πιο πολύ στον εκάστοτε χρήστη. Η αρχή των εισηγητικών συστημάτων έγινε με βάση την πολύ απλή παρατήρηση ότι ως άτομα οι καθημερινές μας αποφάσεις ορίζονται από τις εισηγήσεις κοντινών μας, ή και όχι, ατόμων[61]. Για παράδειγμα όπως όταν επιλέγουμε βιβλίο για να διαβάσουμε ενδιαφερόμαστε για τις προτιμήσεις των ομότιμων μας ή όπως οι εργοδότες ενδιαφέρονται για συστατικές επιστολές των πιθανών εργαζομένων τους. Τα πρώτα λοιπόν συστήματα συστάσεων προσπαθούσαν να μιμηθούν αυτό το τρόπο σκέψης, δημιουργώντας συσχετίσεις της μορφής: Αν στον Α αρέσει το x και y αντικείμενο και στον Β αρέσουν τα x, y, z, τότε πιθανώς στον Α να αρέσει και το z αντικείμενο.

Αυτό μας φέρνει και στις 3 βασικές κατηγορίες εισηγητικών συστημάτων όπως είναι ευρέως αναγνωρισμένες[2]:

- Η πρώτη, και μεγαλύτερη από άποψη πλήθος εργασιών, κατηγορία εισηγητικών συστημάτων ονομάζεται *collaborative filtering*. Αυτή η κατηγορία εκφράζει το σκεπτικό που αναφέραμε παραπάνω, δηλαδή οι εισηγήσεις μας προς κάποιο χρήστη βασίζονται στις προτιμήσεις άλλων χρηστών που έχουν τα ίδια ενδιαφέροντα με τον συγκεκριμένο χρήστη. Ονομάζεται και people to people correlation[87] και στη πρώτη εφαρμογή της μεθόδου[35] απαιτούνταν από κάθε χρήστη να ορίσει τις προτιμήσεις του ο ίδιος και με βάση αυτές θα έβρισκε το σύστημα τους κατάλληλους συσχετισμούς του με άλλους χρήστες. Στη συνέχεια δημιουργήθηκε μία πληθώρα αυτοματοποιημένων τέτοιων εφαρμογών με πιο γνωστές το σύστημα RINGO[89] για τη μουσική και τα συστήματα GroupLens[50] και VideoRecommender[42] για τις ταινίες.

- Η δεύτερη βασική κατηγορία είναι τα *content based* συστήματα. Τέτοιου είδους συστήματα προτείνουν στο χρήστη "αντικείμενα" που έχουν κάποια ομοιότητα με τις προηγούμενες επιλογές του χρήστη[76]. Για να γίνει αυτό, το σύστημα πρέπει να βρει κάποιες ομοιότητες μεταξύ χαρακτηριστικών των "αντικειμένων" τα οποία επεξεργάζεται, πράγμα που σημαίνει ότι κάθε "αντικείμενο" αναπαρίσταται ως ένα *διάνυσμα χαρακτηριστικών*(feature vector). Για παράδειγμα, αν ένας χρήστης είχε εκφράσει θετική κριτική για μία ταινία που ήταν κωμωδία, τότε ένα τέτοιο σύστημα θα του σύστηνε κωμωδίες ως επόμενες ταινίες προς παρακολούθηση. Μια γνωστή τέτοια εφαρμογή είναι το σύστημα Fab[8], το οποίο πρότεινε ιστοσελίδες στους χρήστες, αναπαριστώντας τες ως μία συλλογή από τις 100 πιο συνηθισμένες λέξεις σε αυτές και υπολογίζοντας την ομοιότητα μεταξύ τους.

- Η τρίτη κατηγορία αποτελείται από τα *hybrid systems* και όπως συνεπάγεται από την ονομασία τους, αποτελεί ένα μείγμα των δύο προηγουμένων κατηγοριών. Ο σκοπός αυτής της προσέγγισης είναι να εκμεταλλευτούμε τα θετικά και των δύο κατηγοριών, αντιμετωπίζοντας στη πορεία τα μειονεκτήματα της καθεμίας. Υπάρχουν διάφοροι τρόποι για να γίνει αυτό, όπως για παράδειγμα η δημιουργία δύο ξεχωριστών συστημάτων, ενός collaborative filtering και ενός content based και ο συνδυασμός των εισηγήσεων τους σε μία τελική λίστα από εισηγήσεις. Ένα γνωστό παράδειγμα τέτοιου είδους, είναι το σύστημα Foxtrot[66], το οποίο εισηγείται επιστημονικά άρθρα με βάση τόσο τις προτιμήσεις των συνεργατών του χρήστη, όσο και την αναπαράσταση των άρθρων σε

θεματικές οντολογίες.

Πέρα από τα παραπάνω όμως υπάρχουν και συνήθεις αναφορές σε διάφορα άλλα είδη με τα βασικότερα από αυτά να είναι[17]:

⋆ Demographic συστήματα, όπου μας ενδιαφέρουν δημογραφικά και προσωπικά χαρακτηριστικά του χρήστη, όπως το φύλο ή η ηλικία. Ένα πολύ συνηθισμένο παράδειγμα είναι με τις μηχανές αναζήτησης στο διαδίκτυο, οι οποίες επιστρέφουν αποτελέσματα με προτεραιότητα τις ιστοσελίδες που εδρεύουν στην ίδια χώρα με το χρήστη.

⋆ Knowledge based συστήματα, όπου το σύστημα "γνωρίζει" τι χρειάζεται ο χρήστης με βάση πληροφορίες για αυτόν και τις πρότερες προτιμήσεις του. Παράδειγμα τέτοιου συστήματος είναι, ένα εισηγητικό σύστημα για πτήσεις[1] το οποίο λαμβάνει υπ'όψιν για τις εισηγήσεις τη διάρκεια του ταξιδιού, το κόστος, τον αριθμό των αλλαγών κ.α.

⋆ Utility based, όπου ορίζεται είτε από το χρήστη είτε από το σύστημα μία συνάρτηση "χρησιμότητας", με βάση την οποία μετριέται η χρησιμότητα των "άντικειμένων" για το χρήστη και γίνονται οι πιο κατάλληλες εισηγήσεις[24].

2.2 Συστήματα για Ταινίες

Ειδικά για συστάσεις ταινιών, τα πιο πολλά και γνωστά εισηγητικά συστήματα βασίζονται σε collaborative filtering μεθόδους, όπως το MovieLens[1][41] ή το Netflix[2]. Νέες προσεγγίσεις σε αυτό το τομέα περιλαμβάνουν, εκμετάλλευση των κοινωνικών δικτύων για ενσωμάτωση γνώσης για το χρήστη που θα οδηγήσει σε καλύτερες εισηγήσεις[34], Υπάρχουν μερικές προσεγγίσεις που είναι βασισμένες στο περιεχόμενο των ταινιών, δηλαδή content based, οι οποίες όμως στην πραγματικότητα βασίζονται σε μεταδεδομένα(όπως το είδος ή ο σκηνοθέτης) για τις ταινίες και όχι σε αυτό καθ'αυτό το περιεχόμενό τους. Ένα από τα πιο πετυχημένα συστήματα από αυτά είναι το jinni[3], το οποίο προσφέρει τη δυνατότητα στους χρήστες να χαρακτηρίσουν τις ταινίες με *ετικέτες*(tags) με σημασιολογικό περιεχόμενο. Έτσι οι ταινίες μπορούν να ομαδοποιηθούν κατά είδος, πλοκή ή καστ, επιτρέποντας έναν πιο υψηλού γνωσιακού επιπέδου συσχετισμό των ταινιών. Υπάρχουν και άλλες ιδέες, όπως να συμπεριληφθούν

[1]https://movielens.org/
[2]https://www.netflix.com/
[3]http://www.jinni.com/

οι δημογραφικές ιδιότητες του χρήστη σε μία υβριδική μορφή των content based με τα demographic συστήματα[27], ή το MORE[4] ένας συνδυασμός των μεταδεδομένων αυτών και των προτιμήσεων του χρήστη και των γνωστών του, όπως αυτές εκφράζονται μέσω των κοινωνικών δικτύων[69]. Μία ακόμη σύγχρονη εργασία[57] μοντελοποιεί τις σχέσεις των ταινιών με τις ετικέτες(movie-director, movie-plot κ.α.) ως *πράκτορες*(agents) και το τελικό σύστημα μας δίνει τις βέλτιστες εισηγήσεις με βάση τις προηγούμενες επιλογές του χρήστη.

Όμως όπως προείπαμε, τα παραπάνω contet based συστήματα δεν λαμβάνουν υπ'όψιν τους το πραγματικό περιεχόμενο των ταινιών , αλλά στηρίζονται σε ετικέτες που παρήγαγαν οι χρήστες και αφορούν δικές τους εντυπώσεις. Εμείς ενδιαφερόμαστε για εισηγήσεις που βασίζονται στη πολυμεσική πληροφορία της κάθε ταινίας. Σε αυτό τον τομέα, οι εργασίες που αναλύουν τις ταινίες κατά αυτό τον τρόπο έχουν ως στόχο κάποια άλλη πιο ειδική εφαρμογή και όχι τις εισηγήσεις παρόμοιων ταινιών. Δύο πολύ συνηθισμένες τέτοιες εφαρμογές είναι ο εντοπισμός βίαιων σκηνών, η κατηγοριοποίηση των οποίων ως τέτοιες προκύπτει κυρίως από την εικόνα(βίντεο)[74] ή τον ήχο[31] ή και τα δύο μαζί[32, 70] και η περίληψη ταινιών[26, 51], η οποία βασίζεται κυρίως σε οπτικά και ακουστικά στοιχεία και δευτερευόντως στο κείμενο των υποτίτλων(που συνήθως προκύπτει από *αυτόματη αναγνώριση φωνής*(automatic speech recognition-asr».

Ειδικά τώρα για ανακάλυψη ομοιότητας μεταξύ των ταινιών βασισμένη σε multimodal ανάλυση του περιεχομένου τους υπάρχουν λίγες σχετικές εργασίες. Η πιο ενδιαφέρουσα συνδυάζει χαρακτηριστικά και από τα 3 κανάλια πληροφορίας, κείμενο, ήχος και εικόνα, με σκοπό την εύρεση συσχετίσεων μεταξύ αυτών των low-level χαρακτηριστικών από αυτά τα κανάλια και της ομοιότητας μεταξύ των ταινιών[53]. Άλλη εφαρμογή σε αυτό το τομέα είναι το σύστημα συστάσεων Video-Search[65] για online υπηρεσίες, το οποίο βασίζεται σε χαρακτηριστικά του ήχου, της εικόνας και κειμενικών μεταδεδομένων της ταινίας για να προτείνει τις κατάλληλες ταινίες στον χρήστη. Τέλος, μία διαφορετική κάπως ιδέα παρουσιάζεται από τον Rabinovich[78], καθώς δημιουργεί ένα σύστημα στο οποίο και τα οπτικά και τα ακουστικά και τα κειμενικά στοιχεία της ταινίας, αναπαριστώνται σε θεματική μορφή, πράγμα που επιτρέπει την περιήγησή μας στο χώρο των ταινιών από τελείως διαφορετική οπτική γωνία.

Ολοκληρώνοντας, να τονίσουμε ότι στην προσέγγισή μας θέλουμε να δώσουμε έμφαση στην ανάλυση κειμένου με την αποκάλυψη της λανθάνουσας θεματικής δομής των κειμένων των υ-

[4]http://apps.facebook.com/movie-recommendation/

ποτίτλων σε κάθε ταινία, μέσω της Latent Dirichlet Allocation-LDA μεθόδου. Αν και η LDA έχει χρησιμοποιηθεί στο παρελθόν γενικώς για εισηγητικά συστήματα, όπως για παράδειγμα για την εύρεση θεματικής ομοιότητας ανάμεσα σε ιστοσελίδες και εισήγηση αυτών[49], εντούτοις είναι πολύ λίγες οι εφαρμογές που χρησιμοποιείται για εισήγηση ταινιών. Συγκεκριμένα σε μία εφαρμογή που ξεχωρίζει, δεν έχουμε εφαρμογή της LDA στους υπότιτλους της ταινίας αλλά σε reviews σχετικές με τη ταινία και η θεματική ομοιότητα ανάμεσα στα reviews αυτά μας δίνει την ομοιότητα μεταξύ των ταινιών[10]. Τέλος, σε μία αξιοσημείωτη εργασία που έχει στόχο την παραγωγή της περίληψης της ταινίας σε βίντεο, με δεδομένα μόνο από κείμενο, γίνεται ένας συνδυασμός των υποτίτλων της ταινίας και της περιγραφής της πλοκής της, όπως αυτή δίνεται στη Wikipedia[5], με αποτέλεσμα να παραχθούν τα θέματα από τα οποία αποτελείται αυτή και να γίνει ένας συσχετισμός χρονικός των θεμάτων αυτών με τις σκηνές στη ταινία, έτσι ώστε η τελική περίληψη της ταινίας να περιέχει τις σκηνές εκείνες στις οποίες είναι έντονη η παρουσία των συγκεκριμένων θεμάτων[81].

[5]https://en.wikipedia.org/wiki

Κεφάλαιο 3

Προτεινόμενη Μέθοδος

Σε αυτό το κεφάλαιο θα αναφερθούμε στις διάφορες πτυχές του εισηγητικού συστήματος που σχεδιάσαμε, κυρίως αναφερόμενοι στις θεωρητικές βάσεις των μεθόδων που χρησιμοποιήσαμε. Αρχικά, θα δώσουμε μία συνοπτική περιγραφή του συστήματος(**3.1**) και στη συνέχεια θα αναφερθούμε στη μεθοδολογία που ακολουθήσαμε για την επεξεργασία των υποτίτλων(**3.2**) και για την ανάλυση του ήχου(**3.3**). Τελικά, θα περιγράψουμε τη κατασκευή των πινάκων ομοιότητας μεταξύ των ταινιών(**3.4**) και θα αναλύσουμε και τις μεθόδους σύντηξης πληροφορίας(**3.5**) που θα χρησιμοποιηθούν στη συνέχεια.

3.1 Γενικό Διάγραμμα Μεθόδου

Ξεκινώντας θα κάνουμε μία ανάλυση της μεθόδου που προτείνουμε για τη δημιουργία ενός συστήματος εισηγήσεων για ταινίες βασισμένο στο περιεχόμενο. Η κεντρική ιδέα είναι ένα σύστημα το οποίο θα προσφέρει συστάσεις για διάφορες ταινίες, κατανοώντας την ουσία κάθε ταινίας. Δηλαδή, οι ταινίες θα αναπαριστώνται σε ένα σύστημα γνώσης με βάση το περιεχόμενο που εμπεριέχουν τα διάφορα κανάλια πληροφορίας(κείμενο, ήχος και εικόνα) και με βάση αυτή την αναπαράσταση θα δίνονται οι κατάλληλες εισηγήσεις ανά ταινία, καθώς και η δυνατότητα στο χρήστη να πλοηγηθεί κατάλληλα σε αυτό το σύστημα αναπαράστασης γνώσης. Στο πλαίσιο της εργασίας αυτής, η έμφαση δόθηκε στην ανάλυση του κειμένου αρχικά και δευτερευόντως, και κυρίως επικουρικά για το σύστημά μας όπως θα φανεί παρακάτω, στην ανάλυση του ήχου. Πιο συγκεκριμένα :

- Όσον αφορά στο κομμάτι της ανάλυσης του κειμένου, αναφερόμαστε στους υπότιτλους της ταινίας, οι οποίοι παρέχονται αυτούσιοι για κάθε ταινία και υπόκεινται σε μία σειρά

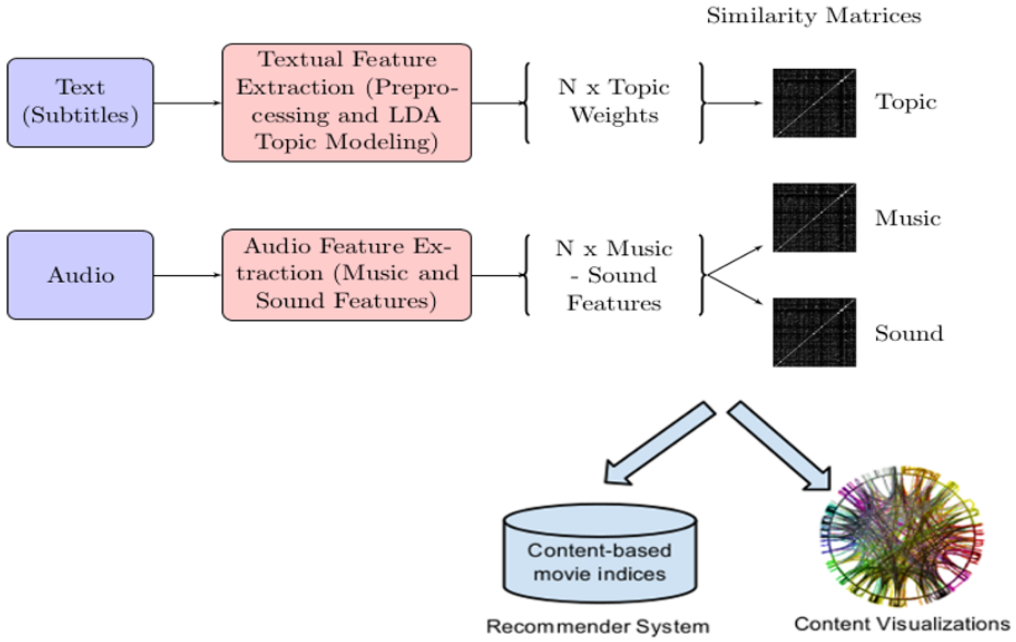
από διεργασίες ώστε να έρθουν σε κατάλληλη μορφή για εξαγωγή θεμάτων από αυτές. Ο τελικός δηλαδή στόχος της ανάλυσης κειμένου είναι η αναπαράσταση της κάθε ταινίας ως ένα μείγμα από διαφορετικά *θέματα*(topics), τα οποία προκύπτουν από τη συλλογή των υποτίτλων που έχουμε στη διάθεση μας. Τα θέματα αυτά αποτελούνται από διάφορες λέξεις που υπάρχουν στη συλλογή των υποτίτλων μας, με διαφορετική βαρύτητα συμμετοχής κάθε λέξης σε κάθε θέμα. Για το σκοπό αυτό εφαρμόζουμε τον αλγόριθμο **Latent Dirichlet Allocation(LDA)**, ο οποίος θα μας δώσει μία συλλογή από θέματα και την αναπαράσταση των ταινιών αυτών σε αυτό το *θεματικό χώρο*(topic space). Τα παραπάνω θα αναλυθούν με λεπτομέρεια στο (**3.2**).

- Όσον αφορά στην ανάλυση ήχου, το κίνητρο μας είναι πάλι το ίδιο, δηλαδή να αναπαραστήσουμε τις ταινίες σε μορφή διανυσμάτων και να βρούμε συσχέτιση μεταξύ αυτών. Στη συγκεκριμένη περίπτωση από τον ήχο θα εξάγουμε δύο διανυσματικές αναπαραστάσεις, μία που αφορά στο είδος της μουσικής στη ταινία(κλασσική, τζαζ κ.α.) και μία που αφορά στο είδος των ήχων που ακούγονται σε αυτή(μουσική, ομιλία, πυροβολισμοί, κ.α.). Όπως αναφέραμε και προηγουμένως, η συμμετοχή του ήχου ως κανάλι πληροφορίας είναι δευτερεύουσα στη παρούσα εργασία και αποτελεί κυρίως βοήθημα στο *θεματικό μοντέλο*(topic model) που σχολιάσαμε πριν. Το κομμάτι αυτό θα αναλυθεί συνοπτικά στο (**3.3**).

- Επίσης, έχοντας τις αναπαραστάσεις των ταινιών σε διάφορους χώρους πληροφορίας θα πρέπει να βρούμε ένα τρόπο μέτρησης της συνάφειας αυτών. Σε αυτή την ενότητα(**3.4**) λοιπόν θα αναφερθούμε στο μετασχηματισμό ομοιότητας *cosine similarity* και στους *πίνακες ομοιότητας*.

- Τέλος, θα αναφερθούμε και με συντομία στις δυνατότητες ένωσης αυτών των δύο μεθόδων στη τελευταία ενότητα αυτού του κεφαλαίου(**3.5**). Εκεί θα αναλύσουμε στους λόγους που δοκιμάζουμε τη *σύντηξη*(fusion) των καναλιών πληροφορίας και τους δύο τρόπους που έγινε αυτό.

Το προτεινόμενο σύστημα που περιγράψαμε προηγουμένως φαίνεται και σχηματικά στο διάγραμμα ροής της μεθόδου στο **Σχήμα 3.1**.



Σχήμα 3.1: Γενικό σχεδιάγραμμα προτεινόμενης μεθόδου.

3.2 Ανάλυση Κειμένου

3.2.1 Γενικά

Σε αυτό το κεφάλαιο θα κάνουμε τη πλήρη θεωρητική περιγραφή των θεμάτων που αφορούν στην ανάλυση κειμένου στο σύνολο της. Θα αναφερθούμε τόσο στο κομμάτι της προεπεξεργασίας(**3.2.2**), όσο και στο κομμάτι *εξαγωγής θεμάτων*(topic extraction) μέσω της *LDA* μεθόδου(**3.2.3**). Θα εξηγήσουμε και διάφορες άλλες μεθόδους, όπως η *tf-idf*, που δεν αποτελούν βασικό κομμάτι της επεξεργασίας κειμένου για τη μέθοδο μας, αλλά είτε τις χρησιμοποιούμε δευτερευόντως σε διάφορα πλαίσια στην εργασία μας είτε τις αγγίζουμε ως θεωρητικά σχετικές με το αντικείμενο.

3.2.2 Προεπεξεργασία Υποτίτλων

Ξεκινώντας λοιπόν θα αναφερθούμε στη προεπεξεργασία που χρειάζονται οι υπότιτλοι για να είναι σε κατάλληλη μορφή για την εξαγωγή θεμάτων από αυτούς. Αυτή η διαδικασία είναι απαραίτητη γιατί υπάρχει άχρηστη πληροφορία στους υπότιτλους στην αρχική τους μορφή. Γενικώς, σε εφαρμογές *ανάκτησης πληροφορίας*(information retrieval) και *εξόρυξης κειμένου*(text mining) είναι συνηθισμένη η θεώρηση ότι σε ένα κείμενο οι λέξεις είναι ανεξάρτητες η μία από την άλλη και η σειρά εμφάνισης τους δεν παίζει σπουδαίο ρόλο στο

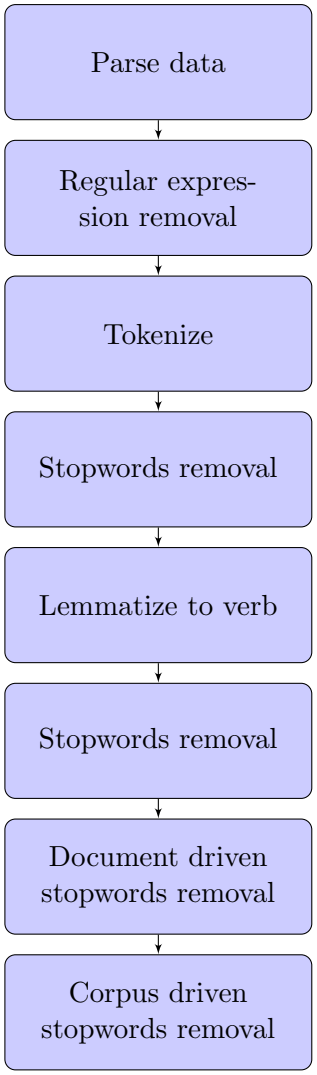
περιεχόμενο αυτών. Αυτή το μοντέλο ονομάζεται bag of words και είναι προφανώς μία απλούστευση, καθώς για τις περισσότερες εφαρμογές δεν ισχύει, όπως και για τη δική μας. Από την άλλη, απλοποιεί πάρα πολύ την επεξεργασία του κειμένου καθώς κάθε κείμενο αντιπροσωπεύεται ως ένα διάνυσμα συχνοτήτων των λέξεων που εμφανίζονται σε αυτό. Για να εξηγήσουμε τα παραπάνω, πρώτα απ'όλα να ξεκαθαρίσουμε ότι με τη λέξη κείμενο εδώ, εννοείται το σύνολο των υποτίτλων κάθε ταινίας και ως συλλογή κειμένων, ή *σώμα*(corpus), εννοούμε το σύνολο των υποτίτλων για τις ταινίες, με τις οποίες ασχολούμαστε. Για να γίνει η αναπαράσταση των κειμένων σύμφωνα με το παραπάνω μοντέλο πρέπει να ορίσουμε πρώτα ένα *λεξιλόγιο*(vocabulary) το οποίο θα περιέχει όλες τις διαφορετικές λέξεις που υπάρχουν στα κείμενα των υποτίτλων. Το διάνυσμα που αναφέραμε προηγουμένως περιέχει τόσες διαστάσεις όσες και οι λέξεις του λεξιλογίου μας και υπάρχει μία "1-1" αντιστοιχία ανάμεσα στις λέξεις αυτές και στις διαστάσεις. Έτσι, κάθε κείμενο εκφράζεται ως ένα διάνυσμα σε αυτό το πολυδιάστατο χώρο, η τιμή του οποίου σε κάθε διάσταση είναι οι φορές εμφάνισης της συγκεκριμένης λέξης στο κείμενο.

Όμως για να γίνει αυτή η αναπαράσταση πρέπει να αφαιρεθεί από το κάθε κείμενο αρκετή άχρηστη πληροφορία όπως σημεία στίξης, λέξεις που δεν προσφέρουν κάτι στο νόημα του κειμένου(όπως το και) κ.α. Αυτή η διαδικασία είναι στο σύνολό της η προεπεξεργασία που πρέπει να γίνει. Το διάγραμμα ροής στο **Σχήμα 3.2.2** εμφανίζει συνολικά τη διαδικασία που θα ακολουθήσουμε.

Εν πρώτοις, είναι το διάβασμα των υποτίτλων από τα .srt αρχεία. Δεν μας ενδιαφέρει σε αυτό το σημείο για το πως αποκτήσαμε τους υπότιτλους σε αυτή τη μορφή, παρά μόνο ότι οι υπότιτλοι δίνονται σε μορφή .srt αρχείων για κάθε ταινία(για τη διαδικασία επιλογής των ταινιών και εύρεσης των υποτίτλων αναφέρονται περισσότερα στο **4.1**). Τα περιεχόμενα των .srt αρχείων είναι ως ακολούθως:

```
261
00:26:38,141 --> 00:26:39,311
Look again.

262
00:26:39,647 --> 00:26:40,922
He's right...
```



Σχήμα 3.2: Διάγραμμα ροής για τη διαδικασία προεπεξεργασίας των υποτίτλων

```
263
00:26:41,239 --> 00:26:42,293
...there.

265
00:26:54,646 --> 00:26:56,851
-Maybe it's rabid.
-Oh, my Christ.
```

Όπως βλέπει κανείς τα δεδομένα δεν είναι σε μορφή κατάλληλη για να φτάσουμε στην διανυσματική απεικόνιση που είπαμε προηγουμένως. Πρώτο στάδιο όπως φαίνεται και στο διάγραμμα είναι το διάβασμα του αρχείου και το φιλτράρισμα ώστε να κρατήσουμε μόνο τις φράσεις του αρχείου. Στη συνέχεια, κάνουμε αφαίρεση συγκεκριμένων χαρακτήρων για να απομακρύνουμε διάφορα σημεία στίξης(μαζί με παρενθέσεις κ.α.), καθώς και συγκεκριμένους *markup* κώδικες για italic και άλλα στοιχεία που προκύπτουν σε αυτά τα αρχεία. Για να το πετύχουμε αυτό χρησιμοποιούμε *κανονικές εκφράσεις*(regular expressions), οι οποίες βρίσκουν τα συγκεκριμένα στοιχεία στο κείμενο και τα απαλείφουν. Ακολούθως, κάνουμε *tokenization* μία καθιερωμένη διαδικασία σε τέτοιες εφαρμογές, η οποία συνίσταται στο χωρισμό των προτάσεων σε λέξεις και η μετατροπή όλων των χαρακτήρων από κεφαλαία σε πεζά(για να αποφύγουμε να υπάρχουν λέξεις όπως "tomorrow" και "Tomorrow" ως διακριτές λέξεις στο λεξιλόγιο μας).

Συνεχίζοντας, προχωράμε σε αφαίρεση συγκεκριμένων λέξεων κλειδιών(stopwords removal) οι οποίες δεν προσφέρουν κάτι το διακριτικό σε κάθε κείμενο. Αυτές είναι πολύ συνηθισμένες λέξεις όπως I, it, and, very, yes κ.α., λίστες των οποίων υπάρχουν διαθέσιμες στο διαδίκτυο. Εμείς, εκτός από τις συγκεκριμένες που βρήκαμε στο διαδίκτυο, κατασκευάσαμε και μία λίστα με τέτοιες λέξεις οι οποίες δεν υπήρχαν στις προαναφερθείσες λίστες και προέκυψαν έπειτα από παρατήρηση των υποτίτλων των ταινιών, οι οποίοι πολλές φορές περιέχουν και αδόκιμες εκφράσεις και λάθη όπως aint, ill, theres ή εκφράσεις συνηθισμένων λέξεων σε μορφή αργκό όπως yeap, yeah τις οποίες είναι δύσκολο να προβλέψει κανείς. Στο σύνολο είναι 609 stopwords, οι οποίες αποτελούν φίλτρο και αφαιρούν τις λέξεις αυτές από το κείμενο των υποτίτλων, όπου τις πετυχαίνουν.

Το επόμενο βήμα είναι η *λημματοποίηση*(lemmatization) σε ρήματα, δηλαδή η μετατροπή όλων των ομόρριζων λέξεων στο κοινό τους λήμμα(σε ρηματική μορφή). Για παράδειγμα, οι διαφορετικές εκφάνσεις της λέξης "καλώ", όπως "κλήση", "καλέσαμε", μετατρέπονται όλες στη ρηματική μορφή "καλώ". Αυτό το κάνουμε γιατί όλες αυτές οι λέξεις εκφράζουν θεματικά και νοηματικά την ίδια έννοια, οπότε θα θέλαμε η ομοιότητα μεταξύ κειμένων που περιέχουν αυτές οι διαφορετικές εκφάνσεις να γίνεται αντιληπτή από το σύστημά μας. Επιπροσθέτως, μειώνουμε έτσι το κατακερματισμό του διανυσματικού χώρου αναπαράστασης των κειμένων, μειώνοντας το πλήθος του λεξιλογίου, με αποτέλεσμα να είναι πιο ευκρινής η σαφήνεια μεταξύ των κειμένων. Δεν είναι υποχρεωτική σαν διαδικασία και σε διάφορες εφαρμογές που θέλουμε

να διατηρήσουμε τη διαφοροποίηση αυτών των εκφάνσεων η χρήση της επιδεινώνει τα τελικά αποτελέσματα. Να σημειωθεί εδώ ότι η υλοποίηση που χρησιμοποιήσαμε είναι αυτή της βιβλιοθήκης *Natural Language ToolKit-NLTK*[58], που είναι βασισμένη στη βάση δεδομένων *WordNet*[29].

Μετά και τη λημματοποίηση που περιγράψαμε προηγουμένως, ξαναφιλτράρουμε τα κείμενα των υποτίτλων από τα παραπάνω stopwords. Η επανάληψη αυτού του βήματος είναι για να φροντίσουμε να αφαιρεθούν λέξεις που ήταν σε διαφορετική μορφή, απ'ότι αυτές στη λίστα με τα stopwords, αλλά έχουν κοινό λήμμα. Σε αυτό το σημείο που φτάσαμε και με τη προεργασία που κάναμε έχουμε ένα μέσο πλήθος λέξεων ανά κείμενο(υπότιτλο): $\approx 2945 \frac{words}{doc}$. Πέρα από το γεγονός ότι θα θέλαμε να μειώσουμε το πλήθος των λέξεων ανά κείμενο περαιτέρω, δεν έχουμε ασχοληθεί με το θέμα των συχνοτήτων των λέξεων στη συλλογή των κειμένων. Για το σκοπό αυτό, οι επόμενες δύο διεργασίες αποτελούν αφαίρεση λέξεων, αλλά η μεν πρώτη είναι αφοσιωμένη στο να βρίσκει λέξεις μικρής σημασίας για τη συλλογή μας βασισμένη σε κάθε ένα κείμενο ξεχωριστά και να τις αφαιρεί, η δε δεύτερη εντοπίζει αυτές τις λέξεις ήσσονος σημασίας λαμβάνοντας υπ'όψιν την κατανομή αυτών σε ολόκληρη τη συλλογή υποτίτλων.

Για να εξηγήσουμε τα παραπάνω και οι δύο μέθοδοι έχουν αρχικά ως σκοπό, τη κατασκευή μία λίστας από stopwords τις οποίες στις συνέχεια θα αφαιρέσουμε από τα κείμενα των υποτίτλων. Η διαφορά τους είναι η σκοπιά από την οποία αντιμετωπίζουν το πρόβλημα, καθώς η πρώτη κοιτάει τις ιδιότητες των λέξεων σε κάθε κείμενο ξεχωριστά ενώ η δεύτερη τις ιδιότητες των λέξεων στη συλλογή ως σύνολο. Ας αναλύσουμε την πρώτη, η οποία όπως προείπαμε είναι document driven και ασχολείται με κάθε κείμενο ξεχωριστά. Πιο συγκεκριμένα, μία λέξη προστίθεται στη λίστα με τις ανεπιθύμητες αν εμφανίστηκε το πολύ δύο φορές σε ένα κείμενο ή αν έχει το πολύ δύο γράμματα. Τέτοιες λέξεις είναι στη πρώτη περίπτωση πολύ σπάνια χρησιμοποιούμενες, στη δε δεύτερη άρθρα ή λέξεις που δεν περιέχουν θεματικό ενδιαφέρον. Διατρέχουμε λοιπόν όλη τη συλλογή των υποτίτλων και κατασκευάζουμε τη λίστα των ανεπιθύμητων λέξεων με αυτές που πληρούν κάποια από τις δύο παραπάνω ιδιότητες. Στη συνέχεια, αφαιρούμε αυτές τις λέξεις από κάθε κείμενο κατά τα γνωστά. Έπειτα από αυτή τη διαδικασία, το μέσο πλήθος λέξεων ανά κείμενο γίνεται: $\approx 1564 \frac{words}{doc}$.

Η δεύτερη μέθοδος κατασκευάζει τη λίστα με τις ανεπιθύμητες λέξεις λαμβάνοντας υπ'όψιν την συχνότητα εμφάνισής τους σε ολόκληρη τη συλλογή των υποτίτλων, γι'αυτό την ονομάζουμε και corpus driven. Συγκεκριμενοποιώντας τα προηγούμενα, εννοούμε ότι μία λέξη

προστίθεται στη λίστα με τις ανεπιθύμητες αν εμφανίζεται στο 70% των κειμένων ή αν εμφανίζεται σε λιγότερα από 3 κείμενα στο σύνολο. Αυτές είναι λέξεις που είναι πολύ συχνές, αφ'ενός, οι οποίες δεν προσφέρουν ιδιαίτερη πληροφορία στο κάθε κείμενο καθώς είναι αρκετά συνηθισμένες και αφ'ετέρου σπάνιες λέξεις που χρησιμοποιούνται σε πολύ λίγες ταινίες, οπότε η εμφάνιση τους δεν προσφέρει καμία δυνατότητα συσχετισμού με άλλες ταινίες. Προφανώς, με αυτούς τους τρόπους χάνεται διαθέσιμη πληροφορία, αλλά για την εφαρμογή μας είναι μία διαδικασία που βοηθάει στα τελικά αποτελέσματα καθώς τονίζει τις συσχετίσεις μεταξύ των ταινιών. Οι τιμές 70% και 3 προέκυψαν πειραματικά κρατώντας τις τιμές που μας έδιναν καλύτερη ποιότητα κειμένου και αποτελεσμάτων συνάφειας μεταξύ των ταινιών(περισσότερα για τα μέτρα απόδοσης στο **4.4**). Έχοντας διατρέξει λοιπόν όλη τη συλλογή κειμένων και δημιουργώντας τη κατάλληλη λίστα με τα stopwords, αφαιρούμε αυτές τις λέξεις από όλα τα κείμενα στη διάθεση μας. Τελειώνοντας και με τη παραπάνω αφαίρεση ο τελικός μέσος αριθμός λέξεων ανά κείμενο είναι : $\approx 989 \frac{words}{doc}$.

Με τη διαδικασία που περιγράψαμε καταφέραμε να δημιουργήσουμε μία συλλογή κειμένων που περιέχουν έναν αριθμό λέξεων. Συνεχίζοντας το προηγούμενο υπόδειγμα υποτίτλων, αν αυτό το κομμάτι αποτελούσε ένα κείμενο μόνο του, το αποτέλεσμα έπειτα από αυτή τη διαδικασία θα ήταν το εξής: $[look, rabid, christ]$. Δηλαδή, κάθε κείμενο υποτίτλων εκφράζεται τώρα ως μία λίστα λέξεων όπως η προηγούμενη, απλά με μέσο μήκος 989 λέξεις. Βέβαια, αυτές οι λέξεις δεν είναι μοναδικές αλλά σε κάθε κείμενο κάποιες επαναλαμβάνονται. Επομένως, για να φέρουμε τα δεδομένα μας στη τελική μορφή ως διανύσματα που τα κελιά τους αντιστοιχούν σε συχνότητες εμφάνισης των λέξεων και οι δείκτες αντιπροσωπεύουν τις λέξεις αυτές, χρειαζόμαστε και μία τελευταία μετατροπή. Αρχικά, ξεχωρίζουμε τις διαφορετικές λέξεις που υπάρχουν στη συλλογή των υποτίτλων στη τελική τους μορφή και αυτό είναι το λεξιλόγιο μας. Είναι : 1498 διαφορετικές λέξεις. Αυτή είναι και η διάσταση του διανυσματικού χώρου αναπαράστασης των κειμένων. Έπειτα, για κάθε κείμενο στη συλλογή μας μετράμε πόσες φορές εμφανίζεται κάθε μία από αυτές τις 1498 λέξεις και αποθηκεύουμε την τιμή του πλήθους στο αντίστοιχο κελί στο διάνυσμα αναπαράστασης. Κατά αυτό το τρόπο έχουμε την αναπαράσταση των κειμένων ως *διανύσματα χαρακτηριστικών*(feature vectors), όπου τα χαρακτηριστικά είναι οι λέξεις του λεξιλογίου. Με αυτή την αναπαράσταση, το μέσο πλήθος διαφορετικών λέξεων ανά ταινία προκύπτει: $\approx 159 \frac{uniquewords}{doc}$. Δηλαδή , για να κλείσουμε και με το παράδειγμα που σχολιάσαμε και προηγουμένως, η διανυσματική του μορφή θα ήταν ένα διάνυσμα με μηδενικά όλα τα στοιχεία εκτός από τα 3 κελιά στα οποία αντιστοιχούν οι λέξεις

look, rabid, christ, στα οποία θα είχε μονάδες.

Επεξεργάζοντας λοιπόν τους υπότιτλους με το παραπάνω τρόπο, τους μετατρέπουμε σε μορφή κατάλληλη για εφαρμογή της μεθόδου **Latent Dirichlet Allocation-LDA** που θα μας οδηγήσει στην εξαγωγή των θεμάτων από το κείμενο. Πριν προχωρήσουμε όμως στο επόμενο κεφάλαιο θα κάνουμε και μία σύντομη περιγραφή της μεθόδου *term frequency-inverse document frequency*(tf-idf), η οποία μοιάζει αρκετά με τη διαδικασία που ακολουθήσαμε στο τέλος της επεξεργασίας των υποτίτλων και εμείς και θα χρησιμοποιηθεί στη συνέχεια (**4.1**).

Η μέθοδος **term frequency-inverse document frequency(tf-idf)** είναι ένα *σχήμα βαρών*(weighting scheme) που αντιστοιχεί σε κάθε λέξη καθενός κειμένου στη συλλογή μας ένα βάρος, ανάλογο με την συχνότητα εμφάνισης της λέξης αυτής στο κείμενο αυτό και αντιστρόφως ανάλογο με τη συχνότητα εμφάνισης της λέξης σε ολόκληρη τη συλλογή κειμένων[84, 63, 86]. Αποτελείται από δύο μέλη, με πρώτο τον όρο να είναι η **term frequency** $tf_{i,d}$ που εκφράζει τη σημαντικότητα του όρου i για το κείμενο d[59]. Ο πιο απλός τρόπος υπολογισμού αυτού του όρου είναι, απλά να του αποδώσουμε τον αριθμό εμφανίσεων της λέξης i στο συγκεκριμένο κείμενο d, δηλαδή αν η συχνότητα εμφάνισης στο κέιμενο είναι $f_{i,d}$ τότε $tf_{i,d} = f_{i,d}$, αν και υπάρχουν και αρκετοί άλλοι μέθοδοι. Ο δεύτερος όρος είναι η **inverse document frequency** idf_i που εκφράζει τη πληροφορία που εμπεριέχει ο όρος i, λαμβάνοντας υπ'όψιν τη συχνότητα εμφάνισης του όρου αυτού σε ολόκληρη τη συλλογή των κειμένων μας[92]. Συνήθως υπολογίζεται ως : $idf_i = \log_2 \frac{N}{n_i}$, όπου N είναι το πλήθος των κειμένων και n_i ο αριθμός των κειμένων που περιέχουν το συγκεκριμένο όρο. Δηλαδή, είναι ο λογάριθμος του αντίστροφου της συχνότητας εμφάνισης της λέξης στη συλλογή μας. Τελικά, το γινόμενο αυτών των δύο μας δίνει το τελικό βάρος για τη λέξη i στο κείμενο d ως : $tfidf_{i,d} = tf_{i,d} \times idf_i$. Αυτή η μετατροπή έχει σαν αποτέλεσμα η συλλογή κειμένων μας, να αναπαρίσταται ως ένας πίνακας βαρών που οι γραμμές αντιπροσωπεύουν δείκτες λέξεων και οι στήλες τα διαφορετικά κείμενα στη συλλογή μας, με τη τιμή του κελιού το βάρος tf-idf. Αυτή η διαδικασία γίνεται για να πάρουν μεγαλύτερο βάρος οι λέξεις που εμφανίζονται πιο πολύ σε κάθε κείμενο, αλλά δεν εμφανίζονται και σε όλα τα διαφορετικά κείμενα της συλλογής μας, καθιστώντας τες φορείς πληροφορίας για το συγκεκριμένο κείμενο.

3.2.3 Εξαγωγή Θεμάτων

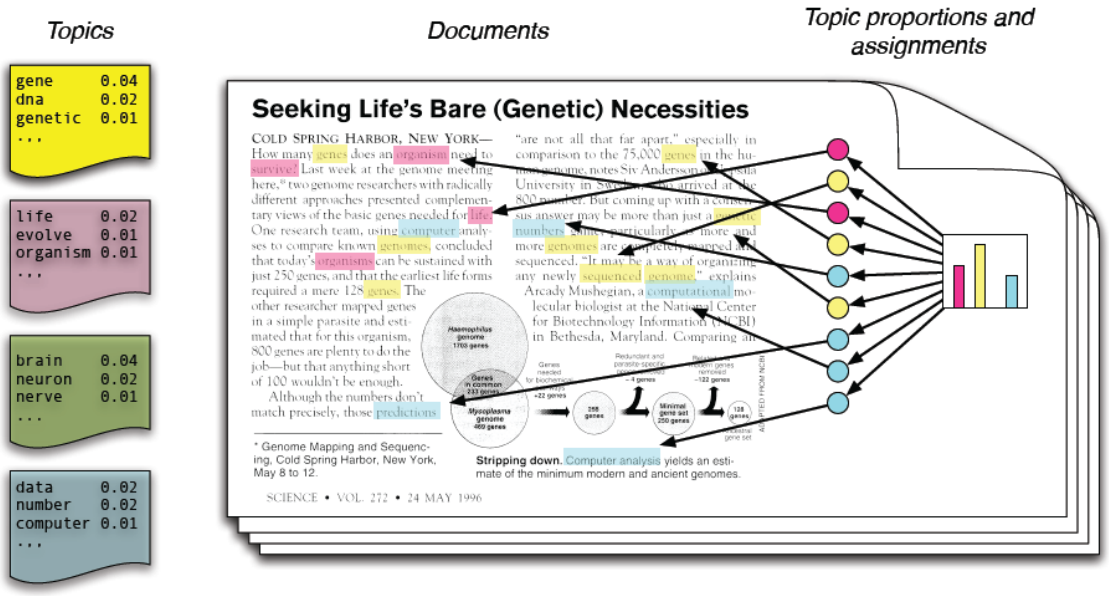
Εισαγωγή

Σε αυτή την ενότητα θα ασχοληθούμε με την *εξαγωγή θεμάτων*(topic extraction) από τη συλλογή των κειμένων μας. Η μέθοδος που θα χρησιμοποιήσουμε είναι η **Latent Dirichlet Allocation-LDA**[15]. Με την συνεχώς αυξανόμενη δημιουργία και ψηφιοποίηση πληροφοριών, βρισκόμαστε σε μία εποχή αφθονίας δεδομένων που γίνεται όλο και πιο δύσκολο να βρει κανείς αυτό ακριβώς το οποίο ψάχνει, να το καταλάβει και να το οργανώσει. Τα εργαλεία που έχουμε στη διάθεσή μας είναι κατά βάση η αναζήτηση με *λέξεις κλειδιά*(keywords) και οι σχετικοί σύνδεσμοι για κάθε δεδομένο. Για κείμενα λόγου χάρη, ψάχνουμε κάποιες συγκεκριμένες λέξεις κλειδιά και βρίσκουμε μία ομάδα κειμένων που σχετίζονται με αυτά και συνήθως κάποιους συνδέσμους με άλλα κείμενα που σχετίζονται με το αρχικό αυτό κείμενο. Αν και αυτό μας καλύπτει τις ανάγκες τις περισσότερες φορές, ένας πιο δυναμικός τρόπος προσπέλασης της πληροφορίας θα ήταν να μπορούσαμε να διερευνήσουμε τα διάφορα δεδομένα στη διάθεσή μας με βάση τα θέματα με τα οποία ασχολούνται. Θα μπορούσαμε αντί να ανασύρουμε κείμενα με αναζητήσεις βασισμένες σε λέξεις κλειδιά, να βρίσκουμε πρώτα το θέμα που μας ενδιαφέρει και στη συνέχεια τα κείμενα που σχετίζονται με αυτό το θέμα. Είτε να παρακολουθήσουμε τη χρονική εξέλιξη των κειμένων που σχετίζονται με κάποιο θέμα και τις διαφοροποιήσεις τους στο χρόνο.

Αυτή την πιο διαισθητική προσέγγιση προσφέρει η συγκεκριμένη μέθοδος. Είναι ένα σύνολο αλγορίθμων *πιθανοτικής μοντελοποίησης θεμάτων*(probabilistic topic modeling) με σκοπό την εύρεση της κρυμμένης θεματικής δομής των διαφόρων κειμένων. Δεν απαιτείται κάποια παραπάνω πληροφορία πάνω στο είδος των κειμένων για αυτή τη μέθοδο, καθώς τα διάφορα θέματα και η σχέση τους με τα κείμενα, προκύπτουν από τη στατιστική ανάλυση των κειμένων. Να τονιστεί εδώ, ότι η συγκεκριμένη μέθοδος μπορεί να λειτουργήσει αυτούσια και σε άλλα είδη δεδομένων, όπως εικόνες[90], ήχους[46] κ.α.[45], απλώς το ενδιαφέρον μας στο πλαίσιο της εργασίας εστιάζεται σε κείμενο.

Η βασική ιδέα πίσω από αυτή τη μέθοδο είναι η πεποίθηση ότι κάθε κείμενο σε μία συγκεκριμένη συλλογή, εκφράζεται ως ένας συνδυασμός θεμάτων, τα οποία προκύπτουν από το σύνολο των κειμένων αυτών. Ο τρόπος που υλοποιείται αυτό με τη συγκεκριμένη μέθοδο είναι μέσω ενός *γενεσιουργού μοντέλου*(generative process), το οποίο δημιουργεί όλα τα κείμενα της συλλογής μας. Για να γίνει πιο κατανοητό το παραπάνω θα θέσουμε ένα σύντομο παράδειγμα[13].

Στο **Σχήμα 3.3** φαίνεται ένα άρθρο με τίτλο Seeking Life's Bare (Genetic) Necessities το οποίο είναι σχετικό με τη χρήση ανάλυσης δεδομένων για την εύρεση του πλήθους των γονιδίων που χρειάζεται ένας οργανισμός για να εξελιχθεί. Με το χέρι, έχουμε τονίσει συγκεκριμένες λέξεις που αφορούν σε κάποια θέματα. Με μπλε είναι χρωματισμένες λέξεις σχετικές με την *ανάλυση δεδομένων*, όπως computer, numbers, με ροζ λέξεις σχετικές με τη *βιολογία* όπως life, organism και με κίτρινο λέξεις σχετικές με τη *γενετική* όπως gene, genome. Συνεχίζοντας αυτή τη διαδικασία για όλο το άρθρο, εξαιρώντας λέξεις μηδενικής θεματικής σημασίας (όπως and, it, etc), θα βλέπαμε ότι το άρθρο αυτό αποτελείται από συνδυασμό, με διαφορετική συμμετοχή, των θεμάτων ανάλυσης δεδομένων, βιολογίας και γενετικής. Αυτή την πληροφορία προσπαθούμε να αποκαλύψουμε με την LDA μέθοδο. Η generative process που είπαμε προηγουμένως, είναι η τυχαία διαδικασία μέσω της οποίας, δοθέντων των παραπάνω θεμάτων μπορούμε να δημιουργήσουμε αυτό το άρθρο. Ως θέμα ορίζουμε μία κατανομή πάνω σε ένα πεπερασμένο λεξιλόγιο. Στο παράδειγμά μας, το θέμα γενετική έχει λέξεις όπως gene, genome με μεγάλη πιθανότητα ενώ τις λέξεις computer, numbers με μικρή.



Σχήμα 3.3: Αναπαράσταση των βασικών ιδεών για την LDA μέθοδο. Στα αριστερά φαίνονται διάφορα θέματα, τα οποία είναι κατανομές πάνω στις λέξεις του λεξιλογίου μας. Στα δεξιά, το ιστόγραμμα που φαίνεται, είναι η κατανομή των θεμάτων για το συγκεκριμένο κείμενο. Για κάθε λέξη επιλέγεται μία ανάθεση σε θέμα(τα χρωματιστά νομίσματα δεξιά) και έπειτα επιλέγεται η λέξη αυτή από τη κατανομή του συγκεκριμένου θέματος στις λέξεις.

Θεωρώντας τώρα δεδομένα, για αρχή, τα θέματα αυτά, μπορούμε να δημιουργήσουμε κάθε κείμενο στη συλλογή μας σε δύο βήματα:

1. Επιλέγουμε μία κατανομή θεμάτων για το κείμενο μας.

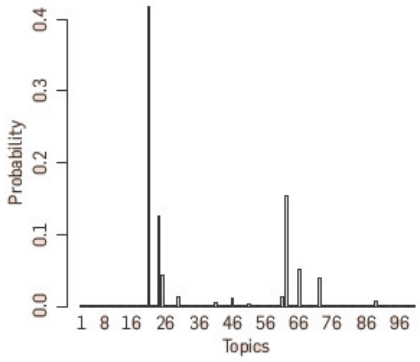
2. Για κάθε λέξη στο κείμενο που θέλουμε να βάλουμε:

   (α′) Τυχαία διαλέγουμε ένα θέμα στο οποίο ανήκει αυτή η λέξη, από τη κατανομή που έχουμε από το βήμα 1

   (β′) Τυχαία διαλέγουμε από τη κατανομή του συγκεκριμένου θέματος στις λέξεις του λεξιλογίου, τη λέξη που τοποθετείται στο κείμενο.

Με αυτό το στατιστικό μοντέλο κάθε κείμενο αποτελείται από ένα συνδυασμό θεμάτων(**1**), κάθε λέξη σε κάθε κείμενο λαμβάνεται από ένα από τα θέματα (**2β′**), το οποίο θέμα έχει προκύψει από τη κατανομή θεμάτων για το συγκεκριμένο κείμενο (**2α′**). Η δύναμη του μοντέλου αυτού έγκειται στη δυνατότητα λειτουργίας του και ανάποδα. Αφού μπορούμε να κατασκευάσουμε όσα άρθρα θέλουμε με αυτό το τρόπο δοθέντων κάποιων θεμάτων, αν αντιστρέψουμε τη διαδικασία και έχουμε έτοιμα τα άρθρα μπορούμε να συμπεράνουμε τα θέματα από τα οποία προέκυψαν αυτά. Δηλαδή, εμείς στη πραγματικότητα στη διάθεση μας έχουμε μόνο τη συλλογή των κειμένων, ενώ τα θέματα, η κατανομή των θεμάτων για κάθε κείμενο και η ανάθεση σε θέμα ανά λέξη για κάθε κείμενο, είναι *κρυφές μεταβλητές*(hidden variables). Οπότε πρέπει να αντιστρέψουμε τη διαδικασία και να βρούμε τη κρυφή δομή που δημιούργησε το κείμενο που έχουμε στα χέρια μας.

Συνεχίζοντας στο παράδειγμα μας, για να γίνει αυτό, εκπαιδεύεται ένα topic model με εκατό θέματα χρησιμοποιώντας ως δεδομένα 17000 άρθρα του περιοδικού Science, έτσι ώστε να βρεθεί η λανθάνουσα θεματική μορφή. Το αποτέλεσμα για το άρθρο του παραδείγματος φαίνεται στο **Σχήμα 3.4**. Στα αριστερά φαίνεται η κατανομή των θεμάτων για το συγκεκριμένο άρθρο(είναι το αντίστοιχο του μινιμαλιστικού ιστογράμματος στο **Σχήμα 3.3**). Παρατηρεί κανείς ότι ενώ υπάρχουν 100 διαφορετικά θέματα, η κατανομή για αυτό το κείμενο έχει ενεργοποιήσει πολύ μικρό αριθμό από αυτά, μόνο ≈ 4 έχουν πιθανότητα εμφάνισης ≥ 0.05. Στα δεξιά του σχήματος φαίνονται τα 4 πιο πιθανά θέματα για το άρθρο, υπό τη μορφή των 15 πιο πιθανών λέξεων που περιέχονται στα θέματα αυτά. Κοιτάζοντας κανείς τις λέξεις αυτές παρατηρεί ότι πράγματι μερικά από αυτά είναι θέματα που αφορούν τη γενετική, τη βιολογική και την ανάλυση δεδομένων, όπως υποθέσαμε στην αρχή του παραδείγματος μας. Να τονίσουμε ότι οι τίτλοι πάνω από τα 4 αυτά θέματα στο αναφερόμενο σχήμα, είναι χειροποίητοι και μεταγενέστεροι. Ανθρώπινος παρατηρητής μετά τη δημιουργία των θεμάτων, τους απέδωσε



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Σχήμα 3.4: Τα αποτελέσματα σε αυτό το σχήμα είναι από ένα μοντέλο με εκατό θέματα που εκπαιδεύτηκε σε 17000 άρθρα του Science και αφορούν το άρθρο του σχήματος **3.3**. Στα αριστερά είναι η πραγματική κατανομή των 100 θεμάτων για το κείμενο. Στα δεξιά οι 15 πιο πιθανές λέξεις για τα 4 πιο πιθανά θέματα που εμφανίζονται σε αυτό το κείμενο

τίτλους ώστε να είναι και ποιοτικά εμφανής η έννοια της λέξης θέμα. Τέλος, έμφαση πρέπει να δοθεί ξανά στο γεγονός ότι το μοντέλο μας δε γνωρίζει τίποτα για το θέμα των άρθρων, δηλαδή δεν υπάρχει κάποια ετικέτα ανά άρθρο, ούτε για τα θέματα γνωρίζει ποιες λέξεις πρέπει να βάλει σε ποια θέματα εκ των προτέρων(για παράδειγμα ότι το θέμα νούμερο 21 τη γενετική, το θέμα νούμερο 61 την εξέλιξη κ.ο.κ).

Latent Dirichlet Allocation

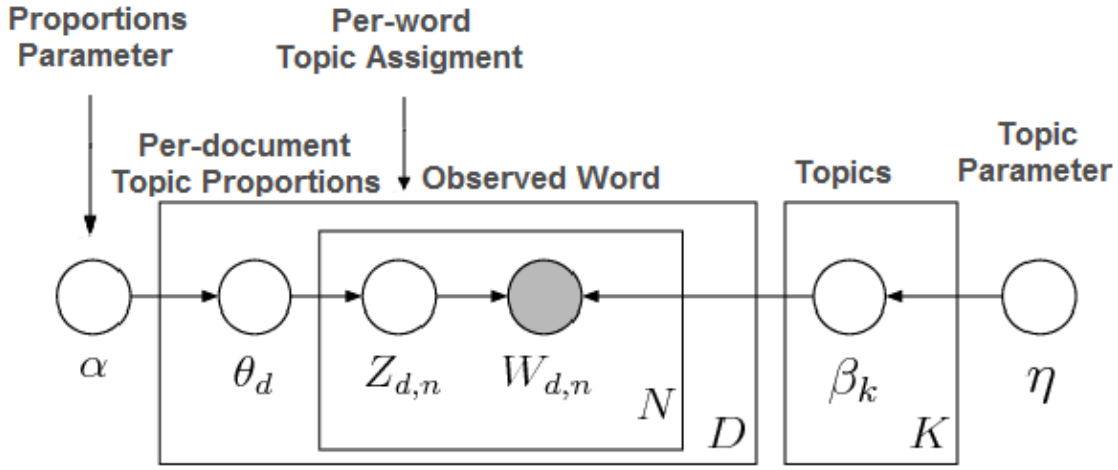
Μετά τα παραπάνω μπορούμε να εξηγήσουμε γιατί η μέθοδος ονομάζεται *Latent Dirichlet Allocation*. Η λέξη *Dirichlet* προκύπτει επειδή η κατανομή από την οποία επιλέγεται τυχαία η κατανομή των θεμάτων ανά κείμενο ονομάζεται dirichlet distribution, *Allocation* καθώς κατανέμουμε τις λέξεις των κειμένων σε θέματα και *Latent* διότι η δομή που περιγράψαμε παραπάνω και θέλουμε να βρούμε είναι λανθάνουσα(=latent). Όπως προείπαμε η LDA ανήκει σε ένα ευρύτερο επιστημονικό πεδίο, αυτό των *probabilstic modeling* μεθόδων, για τα οποία θεωρούμε ότι τα δεδομένα μας προέκυψαν από μία generative process η οποία περιέχει κρυφές μεταβλητές, όπως περιγράφηκε παραπάνω. Μέσω αυτής της διαδικασίας είναι δυνατόν να ορισθεί μία *από κοινού συνάρτηση πιθανότητας*(joint probability distribution) μεταξύ των παρατηρούμενων μεταβλητών(εδώ των λέξεων στα κείμενα) όσο και των κρυφών(στη περίπτωση μας, η κατανομή των λέξεων σε κάθε θέμα, η κατανομή των θεμάτων ανά κείμενο και η

ανάθεση λέξεων σε θέματα ανά κείμενο).

Αυτή η από κοινού κατανομή, μας δίνει την *δεσμευμένη κατανομή*(conditional probability) των κρυφών μεταβλητών με γνωστές τις παρατηρούμενες μεταβλητές, η οποία ονομάζεται και *posterior distribution*. Αυτό προκύπτει άμεσα ως: $P(Hid|Obs) = \frac{P(Hid,Obs)}{P(obs)}$, όπου Hid είναι οι κρυφές μεταβλητές και Obs οι παρατηρούμενες. Να αναφέρουμε εδώ, καθώς θα χρησιμοποιηθεί αργότερα, ότι η $P(Obs|Hid)$ ονομάζεται *πιθανοφάνεια*(likelihood) και εκφράζει τη πιθανότητα δοθέντων των κρυφών μεταβλητών, δηλαδή των μεταβλητών που απαρτίζουν το μοντέλο μας, να παραχθούν τα πραγματικά δεδομένα που έχουμε στα χέρια μας. Επομένως, το αντίστοιχο αλγοριθμικό πρόβλημα στη περιγραφή που κάναμε παραπάνω για την εύρεση της κρυφής θεματικής δομής των κειμένων είναι η εύρεση ακριβώς αυτής της δεσμευμένης κατανομής των άγνωστων μεταβλητών, δοθέντων των παρατηρούμενων λέξεων στα κείμενα. Θα περιγράψουμε τώρα πιο φορμαλιστικά τα προηγούμενα. Αρχικά, θεωρούμε για τη συλλογή μας ότι έχουμε D έγγραφα(documents), N λέξεις ανά κείμενο(χωρίς βλάβη της γενικότητας και χάριν απλότητας θα θεωρήσουμε σταθερό αριθμό λέξεων ανά κείμενο), K ο αριθμός των topics, τον οποίο ορίζουμε εμείς πριν ξεκινήσουμε, και $|V|$ το πλήθος των διαφορετικών λέξεων στο λεξιλόγιο μας. Τα διαφορετικά topics συμβολίζονται με $\beta_{1..K}$, όπου το β_k είναι η κατανομή των λέξεων για το k-οστό θέμα και $\beta_{k,i}$ είναι η συμμετοχή της i-οστής λέξης του λεξιλογίου στο θέμα k. Η μεταβλητή αυτή αντιστοιχεί στα μίνι θέματα που φαίνονται στο **Σχήμα 3.3** αριστερά. Με $\theta_{1..D}$ συμβολίζονται οι διαφορετικές κατανομές θεμάτων ανά κείμενο, αντίστοιχα το ιστόγραμμα στα αριστερά στο **Σχήμα 3.4**. Πιο συγκεκριμένα, με $\theta_{d,k}$ είναι η τιμή συμμετοχής του k-οστού θέματος στο d-οστό κείμενο, δηλαδή η τιμή του ιστογράμματος που αφορά το κείμενο d, για το θέμα με δείκτη k. Οι αναθέσεις σε θέματα των λέξεων των κειμένων είναι $z_{1..D,1..N}$ και είναι οι χρωματιστοί κύκλοι στα δεξιά του **Σχήματος 3.3**. Στη πραγματικότητα κάθε $z_{d,n}$ είναι ένας ακέραιος δείκτης που μας λέει σε ποιο θέμα ανήκει η n λέξη στο d κείμενο. Τέλος, με $w_{d,n}$ συμβολίζουμε τη n-οστή λέξη στο d-οστό κείμενο και αυτή είναι η μεταβλητή που παρατηρούμε.

Ένας τρόπος να αναπαραστήσουμε τις συσχετίσεις μεταξύ αυτών, πράγμα που θα μας βοηθήσει να κατανοήσουμε και τις εξαρτήσεις τους, οπότε θα μπορέσουμε να σχηματίσουμε την από κοινού κατανομή πιθανότητας, είναι μέσω των *γραφικών μοντέλων*(probabilistic graphical models). Το αντίστοιχο μοντέλο για την μέθοδο που περιγράψαμε είναι αυτό που φαίνεται στο **Σχήμα 3.5**. Εξηγούμε και στο υπόμνημα της εικόνας την αναπαράσταση επαρκώς για το

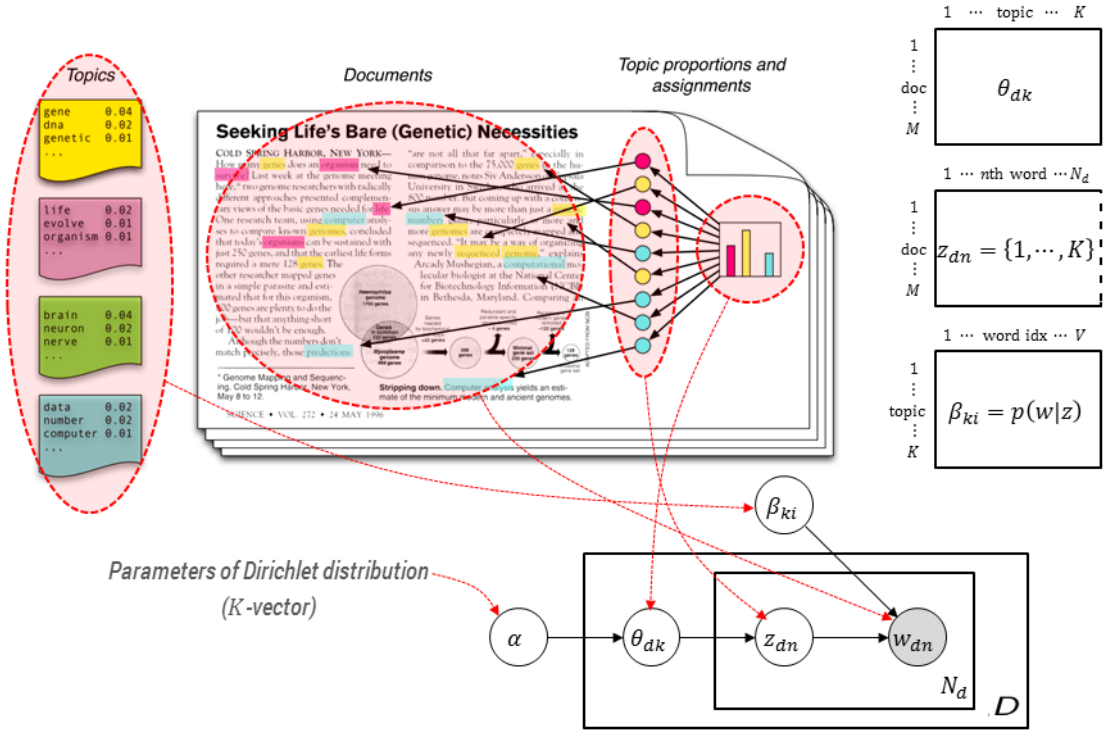
σκοπό που τη θέλουμε. Τα βέλη μας δείχνουν την εξάρτηση των μεταβλητών μεταξύ τους και είναι αυτά που μας βοηθούν να εκφράσουμε μαθηματικά την από κοινού κατανομή πιθανότητας των κρυφών και παρατηρήσιμων μεταβλητών.



Σχήμα 3.5: Graphical Model για την LDA μέθοδο. Ο κάθε κόμβος είναι μία τυχαία μεταβλητή και τα βέλη δείχνουν την εξάρτηση μεταξύ των μεταβλητών(η μεταβλητή στο τέλος του βέλους είναι εξαρτημένη στην μεταβλητή στην αρχή). Η γραμμοσκιασμένη μεταβλητή είναι οι λέξεις που παρατηρούμε και οι υπόλοιπες οι κρυφές μεταβλητές όπως τις αναφέρουμε.Οι ορθογωνικές πλάκες δείχνουν την διάσταση των μεταβλητών για τις οποίες ισχύει επαναληπτικά η αναπαριστώμενη μορφή.

Η μόνη διαφοροποίηση από αυτά που έχουμε αναφέρει είναι η παρουσία των δύο κόμβων α, η. Οι μεταβλητές αυτές, όπως εξηγείται και στο σχήμα, είναι υπερπαράμετροι του μοντέλου LDA. Πιο συγκεκριμένα, τις ορίζουμε εμείς για την εκπαίδευση του συστήματος και αποτελούν τους συντελεστές παραμετροποίησης των *Dirichlet* κατανομών από τις οποίες προκύπτουν οι κατανομές των λέξεων στα θέματα(β_k) και οι κατανομές των θεμάτων ανά κείμενο(θ_d). Θα αναφερθούμε στη συνέχεια συνοπτικά στη Dirichlet κατανομή, αλλά για την ώρα αρκεί να γνωρίζουμε ότι αποτελεί μία κατανομή της οποίας τα στοιχεία της είναι πάλι κατανομές. Δηλαδή, ανήκει σε μία οικογένεια κατανομών που παραμετροποιείται από ένα συντελεστή (εδώ α, η αντίστοιχα) και επιλέγοντας δείγματα από αυτή παίρνουμε κατανομές σε διαστάσεις όσες και η παράμετρος της(εδώ K, όσα και τα topics και $|V|$, όσο και η διάσταση του λεξιλογίου, αντίστοιχα).

Για καλύτερη κατανόηση των κρυφών μεταβλητών και τη σχέση τους με τα κείμενα, τα οποία είναι τα μόνα που έχουμε στη πραγματικότητα στη διάθεσή μας, παραθέτουμε και το **Σχήμα 3.6**. Σε αυτό φαίνονται οι συσχετίσεις των μεταβλητών που περιγράψαμε με το παράδειγμα



Σχήμα 3.6: Αντιστοιχία του Graphical Model με τις μεταβλητές του παραδείγματος. Παρατηρούμε, δεξιά στο σχήμα τις διαστάσεις των μεγεθών και στο υπόλοιπο κομμάτι στη πραγματικότητα οι μεταβλητές αυτές τι αντιπροσωπεύουν.(Έχουμε αφαιρέσει την πλάκα K γύρω από το β_{ki}, για ευκρίνεια.)

του σχήματος **3.3** και τη ποιοτική αναπαράσταση τους. Επίσης, φαίνονται και οι διαστάσεις των μεταβλητών καθώς και η σχέση εξάρτησης μεταξύ αυτών. Με δεδομένα λοιπόν όλα τα παραπάνω, η από κοινού πιθανότητα όλων των μεταβλητών προκύπτει σταδιακά ακολούθως. Αρχικά, για τη κατανομή των λέξεων στα topics, η οποία βρίσκεται στο $V-1$ *simplex*[1] του λεξιλογίου, βλέπουμε ότι η μεταβλητή β_k εξαρτάται μόνο από την υπερπαράμετρο η της Dirichlet κατανομής. Άρα, από αυτό το κομμάτι έχουμε την : $\prod_{k=1}^{K} p(\beta_k|\eta)$. Ομοίως, είναι και η θ_d η οποία εξαρτάται μόνο από την α υπεραπαράμετρο και είναι η πρώτη στην αλυσίδα των εξαρτημένων μεταβλητών. Αντίστοιχα, η μεταβλητή αυτή υπάρχει στο $K-1$ simplex των θεμάτων και αφορά όλα τα διαφορετικά κείμενα. Άρα θα προκύψει: $\prod_{d=1}^{D} p(\theta_d|\alpha)$. Στη συνέχεια της αλυσίδας είναι οι αναθέσεις $z_{d,n}$ οι οποίες εξαρτώνται από το θ_d, ενώ αφορούν όλα τα κείμενα και όλες τις λέξεις αυτών. Άρα, ο όρος που προκύπτει είναι : $\prod_{d=1}^{D}\prod_{n=1}^{N} p(z_{d,n}|\theta_d)$. Τέλος, για τις λέξεις που παρατηρούμε βλέπουμε ότι εξαρτώνται από δύο μεταβλητές και έχουν

[1] Ένα V-διάστατο διάνυσμα β κείται στο V-1-simplex αν $\beta_i \geq 0, \forall i$ και ικανοποιείται η συνθήκη $\sum_{i=1}^{V} \beta_i = 1$, επομένως μόνο τα $V-1$ είναι ανεξάρτητα εξ'ου και ο τίτλος.

πάλι τις ίδιες διαστάσεις με τη $z_{d,n}$ μεταβλητή. Άρα τελικά θα είναι στο σύνολο της:

$$p(\beta_k, \theta_d, z_{d,n}, w_{d,n}|\alpha, \eta) = p(\beta|\eta)p(\theta|\alpha)p(z|\theta)p(w|\beta_z) = \\ (\underbrace{\prod_{k=1}^{K} p(\beta_k|\eta)}_{\text{Dirichlet}(\eta)})(\underbrace{\prod_{d=1}^{D} p(\theta_d|\alpha)}_{\text{Dirichlet}(\alpha)})(\prod_{n=1}^{N} p(z_{d,n}|\theta_d)\underbrace{p(w_{d,n}|z_{d,n}, \beta_k)}_{\text{likelihood}})) \quad (3.1)$$

Έχοντας λοιπόν την από κοινού πιθανότητα από την (**3.1**), μπορούμε να βρούμε την posterior, όπως είχαμε αναφέρει ως εξής:

$$p(\beta_k, \theta_d, z_{d,n}|w_{d,n}, \alpha, \eta) = \frac{p(\beta_k, \theta_d, z_{d,n}, w_{d,n}|\alpha, \eta)}{p(w_{d,n}|\alpha, \eta)} \quad (3.2)$$

Δυστυχώς, η παραπάνω κατανομή δεν είναι δυνατόν να υπολογιστεί. Αυτό συμβαίνει λόγω του παρανομαστή $p(w_{d,n}|\alpha, \eta)$, ο οποίος αν αναλυθεί, περιθωριοποιώντας τις μεταβλητές τις α, η ώστε να εμφανιστούν οι θ_d, β_k, προκύπτει τελικά:

$$p(w|\alpha, \eta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int (\prod_{i=1}^{K} \theta_i^{a_i - 1})(\prod_{n=1}^{N} \sum_{k=1}^{K} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n}) \quad (3.3)$$

Η συνάρτηση αυτή όμως δεν γίνεται να υπολογιστεί λόγω της σύζευξης μεταξύ των θ, β κατά την άθροιση τους πάνω σε όλα τα δυνατά θέματα k[23]. Συγκεκριμένα, πρέπει να υπολογίσουμε την πιθανότητα να παραχθούν τα δεδομένα μας από όλα τα δυνατά topic μοντέλα, δηλαδή όλες τους πιθανούς συνδυασμούς λέξεων, σε όλα τα δυνατά θέματα, πράγμα αδύνατο να υπολογιστεί καθώς για ένα μόνο κείμενο πρέπει να δούμε K^N συνδυασμούς. Οπότε, όπως σε πολλά άλλα τέτοια προβλήματα θα προσπαθήσουμε να προσεγγίσουμε τη πραγματική κατανομή αυτή.

Approximate Posterior Inference

Υπάρχουν διάφοροι τρόποι για να προσεγγίσουμε τη συγκεκριμένη κατανομή. Αυτό γίνεται είτε μέσω *mean field variational methods*[15], είτε μέσω μεθόδων *expectation propagation*[67], είτε μέσω *collapsed variational methods*[95], είτε τέλος με τη μέθοδο που θα αναλύσουμε και εμείς η οποία είναι μία παραλλαγή της δειγματοληψίας Gibbs(*Collapsed Gibbs Sampling*). Αναλύουμε αυτή τη μέθοδο γιατί η βιβλιοθήκη που χρησιμοποιούμε, όπως περιγράφεται παρακάτω, αυτή την μέθοδο χρησιμοποιεί. Θα κάνουμε όμως και μία απαραίτητη αναφορά στη κατανομή *Dirichlet* και το θέμα των συζυγών πρότερων κατανομών(conjugacy),

καθώς είναι στοιχεία που θα χρησιμοποιήσουμε για να εξαγάγουμε το τελικό αποτέλεσμα.

Αρχικά, η generative process γίνεται τώρα πιο μαθηματικοποιημένη:

1. Για κάθε topic $k = 1...K$:

   (α′) $\beta_k \sim Dirichlet(\eta)$

2. Για κάθε κείμενο $d = 1...D$:

   (α′) $\theta_d \sim Dirichlet(\alpha)$

   (β′) Για κάθε λέξη $w_{d,n} \in d, n = 1...N$:

      i. Διάλεξε $z_{d,n} \sim Multinomial(\theta_d)$

      ii. Διάλεξε λέξη $w_{d,n}$ από την $p(w_{d,n}|z_{d,n}, \beta_{z_{d,n}})$, η οποία είναι $\sim Multinomial(\beta_{z_{d,n}})$

Η παραπάνω διαδικασία μας δίνει την από κοινού κατανομή, όπως είχε υπολογιστεί και στην εξίσωση (**3.1**) : $p(\beta_k, \theta_d, z_{d,n}, w_{d,n}|\alpha, \eta) = p(\beta|\eta)p(\theta|\alpha)p(z|\theta)p(w|\beta_z)$. Στη πραγματικότητα όμως έχοντας τα $z_{d,n}$ μπορούμε να παράγουμε τα β_k, θ_d, όπως θα δούμε παρακάτω, επομένως αυτό που πραγματικά ζητάμε να βρούμε είναι η: $p(Hid|Obs) = p(z_{d,n}|w_{d,n}, \alpha, \eta)$. Σύμφωνα με τον ορισμό της δεσμευμένης πιθανότητας, αυτή είναι:

$$p(z|w, \alpha, \eta) = \frac{p(z, w|\alpha, \eta)}{p(w|\alpha, \eta)} \propto p(z, w|\alpha, \eta) \tag{3.4}$$

Συνεχίζοντας και χρησιμοποιώντας τον ορισμό της *οριακής πιθανότητας*(marginal probability) έχουμε:(όπου οι δείκτες των μεταβλητών παραλείπονται όπου δεν χρειάζονται για ευκρίνεια)

$$p(z, w|\alpha, \eta) = \int\int p(z, w, \theta, \beta|\alpha, \eta) d\theta d\beta \tag{3.5}$$

Σύμφωνα όμως με την εξίσωση (**3.1**) η προηγούμενη εξίσωση γίνεται :

$$p(z, w|\alpha, \eta) = \int\int p(\beta|\eta)p(\theta|\alpha)p(z|\theta)p(w|\beta_z) d\theta d\beta \tag{3.6}$$

Χωρίζοντας τώρα τις μεταβλητές σύμφωνα με τα διαφορικά ολοκλήρωσης:

$$p(z, w|\alpha, \eta) = \int p(z|\theta)p(\theta|\alpha)d\theta \int p(w|\beta_z)p(\beta|\eta)d\beta \tag{3.7}$$

Παρατηρούμε σε αυτό το σημείο ότι και τα δύο ολοκληρώματα έχουν στο εσωτερικό τους γινόμενα πολυωνυμικών κατανομών με Dirichlet κατανομές. Σε αυτό το σημείο εισάγεται η έννοια των *συζυγών πρότερων κατανομών*(conjugate priors), η οποία χρησιμοποιείται πάρα πολύ συχνά σε θέματα Bayesian inference, καθώς μας επιτρέπει να προσεγγίζουμε posterior κατανομές χωρίς να κάνουμε ακριβής αριθμητικούς υπολογισμούς. Για να το εξηγήσουμε πλήρως, σύμφωνα με το κανόνα του Bayes ισχύει:

$$p(Hid|Obs) = \frac{p(Obs|Hid)p(Hid)}{p(Obs)} \propto p(Obs|Hid)p(Hid) \tag{3.8}$$

, όπου $p(Hid|Obs)$ η ύστερη κατανομή (posterior distribution), $p(Obs|Hid)$ η πιθανοφάνεια (likelihood), $p(Hid)$ η πρότερη κατανομή (prior distribution), $p(Obs)$ η σταθερά κανονικοποίησης (evidence). Αυτό που βλέπουμε τελικά είναι ότι η ύστερη κατανομή είναι ανάλογη των $\propto p(Obs|Hid)p(Hid)$, καθώς ο όρος $p(Obs)$ είναι κοινός για όλες τις Hid μεταβλητές. Αν η posterior κατανομή $p(Hid|Obs)$ ανήκει στην ίδια κλάση οικογένειας με την πρότερη κατανομή $p(Hid)$, τότε μεταξύ τους είναι συζυγείς κατανομές και η πρότερη κατανομή ονομάζεται συζυγής πρότερη κατανομή για τη πιθανοφάνεια $p(Obs|Hid)$. Η στρατηγική που ακολουθείται σε τέτοιες περιπτώσεις είναι να επιλέγουμε την πρότερη πιθανότητα να έχει τέτοια μορφή ώστε να είναι η συζυγής πρότερη ως προς την πιθανοφάνεια και επομένως η ύστερη κατανομή να είναι συζυγής με τη πρότερη και να διευκολύνει τους υπολογισμούς μας. Στη μοντελοποίηση που κάναμε προηγουμένως στη (**3.7**) βλέπουμε ότι έχουμε το γινόμενο πρότερων κατανομών Dirichlet ($p(\theta|\alpha) \sim Dir(a), p(\beta|\eta) \sim Dir(\eta)$), με πιθανοφάνειες πολυωνυμικών κατανομών $p(z|\theta) \sim Multi(\theta), p(w|\beta_z) \sim Multi(\beta_{z_k})$. Όμως η Dirichlet κατανομή είναι η συζυγής πρότερη της πολυωνυμικής, οπότε το γινόμενο τους θα είναι πάλι Dirichlet κατανομή απλά με διαφορετικό συντελεστή. Για να δούμε ποιος είναι αυτός θα κάνουμε και μία σύντομη αναφορά στη κατανομή Dirichlet.

Η *Dirichlet* κατανομή, συμβολίζεται με $Dir(\alpha)$, είναι μία οικογένεια κατανομών $K \geq 2$ διαστάσεων που παραμετροποιείται από το διάνυσμα $\alpha = [\alpha_1, \alpha_2, ..., \alpha_K], \alpha_i > 0$. Στην ουσία αποτελεί μία κατανομή κατανομών, δηλαδή τα δείγματα αυτής είναι κατανομές με διαφορετική μορφή ανάλογα τις τιμές του διανύσματος α. Στη περίπτωση μας θα ασχοληθούμε με συμμετρική(symmetric-exchangeable) κατανομή Dirichlet, για την οποία το διάνυσμα α έχει

σταθερή τιμή: $\alpha_1 = \alpha_2 = ... = \alpha_K$. Αν η θ ακολουθεί τη κατανομή Dirichlet τότε είναι :

$$\theta \sim Dir(\alpha) = p(\theta|\alpha) = \frac{\Gamma(Ka)}{\Gamma(a)^K} \prod_{k=1}^{K} \theta_k^{a_k-1} = \frac{1}{\Delta_K(a)} \prod_{k=1}^{K} \theta_k^{a_k-1} \tag{3.9}$$

, όπου $\Delta_K(a) = \frac{\Gamma(a)^K}{\Gamma(Ka)}$, η πολυδιάστατη επέκταση της συνάρτησης Beta. Μία ιδιότητα της Dirichlet που θα χρησιμοποιήσουμε είναι η αναμενόμενη τιμή :

$$\mathbb{E}[\theta_i|\alpha_i] = \frac{\alpha_i}{\sum\limits_i \alpha_i} \tag{3.10}$$

Η συνάρτηση *Γάμμα* πρόκειται για τη γενίκευση του παραγοντικού στους πραγματικούς αριθμούς σαν έννοια, δηλαδή ισχύει $\Gamma(x) = (x-1)!$. Επίσης, μία μαθηματική ιδιότητα που θα μας φανεί χρήσιμη στη συνέχεια, προκύπτει από το γεγονός ότι αφού η Dirichlet πρόκειται για συνάρτηση πυκνότητα πιθανότητας, αν την ολοκληρώσουμε στη μεταβλητή της, θ, τότε πρέπει να πάρουμε μονάδα:

$$\int \frac{1}{\Delta_K(a)} \prod_{k=1}^{K} \theta_k^{a_k-1} d\theta = 1 \tag{3.11}$$

Βγάζοντας λοιπόν εκτός ολοκληρώματος το κομμάτι με τη συνάρτηση $\Delta_K(a)$, καθώς δεν επηρεάζεται από τη μεταβλητή ολοκλήρωσης, έχουμε τελικά ένα τρικ που θα μας βοηθήσει στη συνέχεια:

$$\int \prod_{k=1}^{K} \theta_k^{a_k-1} d\theta = \Delta_K(a) \tag{3.12}$$

Συνήθως η Dirichlet χρησιμοποιείται ως *πρότερη κατανομή πιθανότητας*(prior probability distribution), καθώς αποτελεί τη συζυγή κατανομή της πολυωνυμικής κατανομής. Συγκεκριμένα στην περίπτωσή μας, αν έχουμε N διαφορετικά δείγματα $z_{d,n}$, τις λέξεις ενός κειμένου, τα οποία ανήκουν σε κάποια εκ των K διαφορετικών θεμάτων, τότε αν ορίσουμε ως $wt_{d,k}$ το διάνυσμα με μεταβλητές $wt_1, wt_2, ..., wt_K$ που περιέχουν το πλήθος αναθέσεων των λέξεων στο αντίστοιχο topic, για το συγκεκριμένο κείμενο, τότε η πολυωνυμική αυτή κατανομή γράφεται ως :

$$p(z|\theta) = \prod_{i=1}^{K} \theta_{d,k}^{wt_{d,k}}$$

. Οπότε το γινόμενο των δύο θα είναι:

$$p(z|\theta)p(\theta|\alpha) = \prod_{i=1}^{K} \theta_{d,k}^{nwt_{d,k}} \frac{1}{\Delta_K(a)} \prod_{k=1}^{K} \theta_{d,k}^{a_k-1} = \frac{1}{\Delta_K(a)} \prod_{k=1}^{K} \theta_{d,k}^{a_k+wt_{d,k}-1} \tag{3.13}$$

Συνεχίζοντας τώρα την εξίσωση (**3.7**) και με βάση την (**3.13**), θα υπολογίσουμε το πρώτο από τα δύο ολοκληρώματα. Πριν από αυτό όμως να τονιστεί ότι τα παραπάνω ισχύουν όταν μιλάμε για ένα συγκεκριμένο έγγραφο, το d σύμφωνα και με τους δείκτες των μεταβλητών. Τώρα που μας ενδιαφέρει το σύνολο της συλλογής μας κάνουμε την ίδια διαδικασία επαναληπτικά για όλα τα κείμενα, οπότε προστίθεται ο όρος $\prod_{d=1}^{D}$ και πλέον με $wt_{d,k}$ αναφερόμαστε στο συγκεκριμένο κείμενο d, ενώ με wt_d αναφερόμαστε στη στήλη του πίνακα μετρητών wt που περιέχει τα πλήθη εμφανίσεων όλων των διαφορετικών topics στο κείμενο d (Η εύκολη γενίκευση είναι δυνατή γιατί τα θ_d είναι ανεξάρτητα το ένα από το άλλο, για τα διάφορα d. Το ίδιο συμβαίνει και με τα $z_{d,n}$). Έχουμε λοιπόν:

$$\int p(z|\theta)p(\theta|\alpha)d\theta = \int \prod_{d=1}^{D} \frac{1}{\Delta_K(a)} \prod_{k=1}^{K} \theta_{d,k}^{a_k+wt_{d,k}-1} d\theta \tag{3.14}$$

Χρησιμοποιώντας τώρα το τρικ της εξίσωσης (**3.12**), μετατρέπουμε τον όρο της προηγούμενης εξίσωσης:

$$\int \prod_{k=1}^{K} \theta_{d,k}^{a_k+n_{d,k}-1} d\theta = \Delta_K(a + n_{d,k}) \tag{3.15}$$

Οπότε καταλήγουμε στη τελική μορφή:

$$\int p(z|\theta)p(\theta|\alpha) = \prod_{d=1}^{D} \frac{\Delta_K(a + wt_d)}{\Delta_K(a)} \tag{3.16}$$

Έπειτα, θα υπολογίσουμε με όμοιο τρόπο το δεύτερο ολοκλήρωμα. Πρώτα όμως θα αναφέρουμε τα επί μέρους τμήματα του γινομένου. Το $p(\beta|\eta) \sim Dir(\eta)$ οπότε, άμεσα προκύπτει, ότι για ένα συγκεκριμένο θέμα k η κατανομή πάνω στις λέξεις είναι:

$$\beta \sim Dir(\eta) = p(\beta|\eta) = \frac{\Gamma(V\eta)}{\Gamma(\eta)^V} \prod_{v=1}^{V} \beta_v^{\eta_v-1} = \frac{1}{\Delta_V(a)} \prod_{v=1}^{V} \beta_v^{\eta_v-1} \tag{3.17}$$

και άμεσα επεκτάσιμα για όλα τα θέματα(αφού είναι ανεξάρτητα τα β_k μεταξύ τους) :

$$\beta \sim Dir(\eta) = \prod_{k} p(\beta_k|\eta) = \prod_{k=1}^{K} \frac{1}{\Delta_V(\eta)} \prod_{v=1}^{V} \beta_{k,v}^{\eta_v-1} \tag{3.18}$$

Επίσης, παρόμοια με προηγουμένως προκύπτει η πολυωνυμική κατανομή $p(w|\beta_z)$. Αρκεί να ορίσουμε ένα πίνακα μετρητών tc, διαστάσεων $K \times V$, του οποίου το κελί $tc_{k,v}$ περιέχει το πλήθος των φορών που το topic k ανατέθηκε στη λέξη v. Γράφουμε το αποτέλεσμα απ'ευθείας

στο σύνολο των K topics:

$$p(w|\beta_z) = \prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{k,v}^{tc_{k,v}} \tag{3.19}$$

Συνεχίζοντας λοιπόν κατά τα γνωστά και με τις (**3.18**,**3.19**), το δεύτερο ολοκλήρωμα στην (**3.7**)

$$\int p(w|\beta_z)p(\beta|\eta)d\beta = \prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{k,v}^{tc_{k,v}}\frac{1}{\Delta_V(\eta)}\prod_{v=1}^{V}\beta_{k,v}^{\eta_v-1} = \prod_{k=1}^{K}\frac{1}{\Delta_V(\eta)}\prod_{v=1}^{V}\beta_{k,v}^{tc_{k,v}+\eta_v-1} \tag{3.20}$$

Και εφαρμόζοντας το ίδιο τρικ(**3.12**) με προηγουμένως καταλήγουμε στη μορφή:

$$\int p(w|\beta_z)p(\beta|\eta)d\beta = \prod_{k=1}^{K}\frac{\Delta_V(\eta+tc_k)}{\Delta_V(\eta)} \tag{3.21}$$

Άρα συνδυάζοντας τα δύο προηγούμενα(**3.16**, **3.21**) το τελικό αποτέλεσμα γίνεται:

$$p(z,w|\alpha,\eta) = \prod_{k}\frac{\Delta_V(\eta+tc_k)}{\Delta_V(\eta)}\prod_{d}\frac{\Delta_K(a+wt_d)}{\Delta_K(a)} \tag{3.22}$$

Έχοντας λοιπόν την προηγούμενη εξίσωση μπορούμε τώρα να κάνουμε δειγματοληψία Gibbs για να βρούμε τις τιμές της $p(z|w)$ με βάση τη γνώση των $p(z,w)$. Θα κάνουμε σε αυτό το σημείο μία μικρή αναφορά στις λεπτομέρειες της δειγματοληψίας για να γίνει κατανοητή διαισθητικά η λειτουργία της. Η δειγματοληψία Gibbs είναι μία ειδική περίπτωση *Markov Chain Monte Carlo-MCMC* προσομοίωσης[60, 55]. Η βασική ιδέα πίσω από αυτή είναι να προσεγγίσουμε τη ζητούμενη κατανομή, θέτοντας την ως την *κατανομή μόνιμης κατάστασης*(steady state distribution) της αλυσίδας Markov και καθώς θα έχουμε καταλήξει στη τελική μορφή, το λεγόμενο *burn-in* της αλυσίδας που είναι απαραίτητο για να εξαλείψουμε τα λάθη της αρχικοποίησης, τα δείγματα αυτής θα είναι δείγματα που προσεγγιστικά θα ανήκουν στη κατανομή που ζητάμε[39]. Ο Gibbs sampler είναι μία ειδική περίπτωση MCMC, όπου κάθε διάσταση x_i της ζητούμενης κατανομής, δειγματοληπτείται μία τη φορά, θεωρώντας ως γνωστές τις υπόλοιπες διαστάσεις, εφ'εξής x_{-i}. Δηλαδή, αυτό που θέλουμε είναι η:

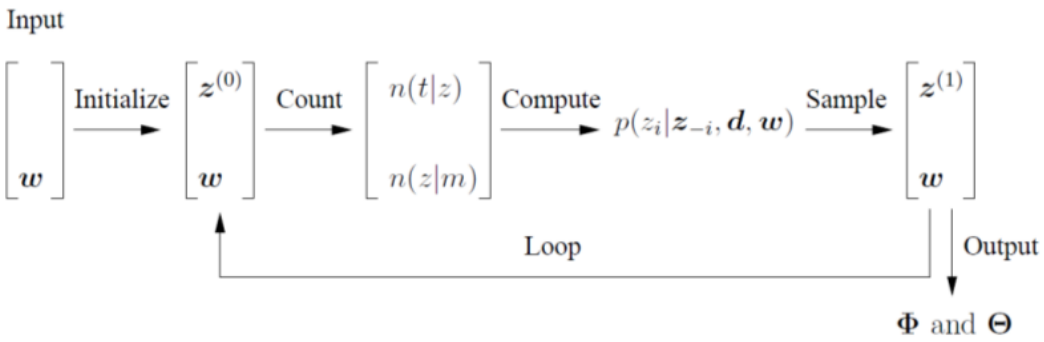
$$p(x_i|x_{-i}) = \frac{p(x_i)}{p(x_{-i})} = \frac{p(x)}{\int p(x)dx_i} \text{ με } x = [x_i, x_{-i}] \tag{3.23}$$

Στην περίπτωση που το μοντέλο μας περιέχει κρυφές μεταβλητές, όπως εδώ π.χ. η z, η

εξίσωση (**3.23**) γίνεται, με παρατηρήσιμη μεταβλητή πλέον τη x:

$$p(z_i|z_{-i},x)=\frac{p(z,x)}{p(z_{-i},x)}=\frac{p(z,x)}{\int p(z,x)dz_i} \tag{3.24}$$

Αν και δεν ξέρουμε πόσες επαναλήψεις θα χρειαστούν, δηλαδή για τη περίοδο burn in της αλυσίδας, είναι αποδεδειγμένο ότι θα έχουμε σύγκλιση με τη δειγματοληψία Gibbs στη κατανομή που θέλουμε. Στην LDA μέθοδο ενδιαφερόμαστε για τις κατανομές των λέξεων στα θέματα β_k, για τις κατανομές των θεμάτων στα κείμενα θ_d και για τις αναθέσεις των λέξεων των κειμένων στα θέματα, $z_{d,n}$. Ενώ μπορούμε πράγματι να κατασκευάσουμε αλγόριθμο που να τα υπολογίζει και τα 3 μέσω της Gibbs δειγματοληψίας, εντούτοις υπάρχει το πρόβλημα ότι το state space της αλυσίδας αποτελείται από τις μεταβλητές, με πλήρη διάσταση η κάθε μία, των κατανομών που προσεγγίζουμε. Οπότε για να μειώσουμε τον αριθμό επαναλήψεων (λιγότερα states, λιγότερες επαναλήψεις για σύγκλιση(convergence)) και επειδή έχοντας τα $z_{d,n}$ μπορούμε να βρούμε τα β_k,θ_d, όπως θα δείξουμε παρακάτω, επικεντρωνόμαστε στην εύρεση μόνο των $z_{d,n}$. Η προσέγγιση αυτή, ονομάζεται collapsed[72], ή Rao-Blackwellised[19] στη βιβλιογραφία. Πλέον δηλαδή μας ενδιαφέρει η πιθανότητα το θέμα k να ανατεθεί σε κάποια λέξη w_i(δηλαδή $z_i=k$), δοθέντων όλων των υπολοίπων αναθέσεων των λέξεων σε θέματα z_{-i} και τον παρατηρήσιμων μεταβλητών. Αυτό γράφεται ως: $p(z_i|z_{-i},w,\alpha,\eta)$. Η διαδικασία που περιγράψαμε προηγουμένως φαίνεται περιληπτικά στο **Σχήμα 3.7**.



Σχήμα 3.7: Εδώ είναι η Collapsed Gibbs Sampling διαδικασία για την LDA . Αρχικοποίηση των z, και κατασκευή των tc, wd πινάκων μετρητών($n(t|z), n(z|m)$ αντίστοιχα στην εικόνα), υπολογισμός με βάση την 3.27 και επανάληψη. Στο τέλος έχουμε σαν έξοδο τα β,θ που συμβολίζονται ως Φ,Θ στο σχήμα.

Για να υπολογίσουμε λοιπόν την παραπάνω κατανομή έχουμε:

$$p(z_i|z_{-i},w_{-i},\alpha,\eta)=\frac{p(z_i,z_{-i},w|\alpha,\eta)}{p(z_{-i},w|\alpha,\eta)}\propto\frac{p(z,w|\alpha,\eta)}{p(z_{-i},w_{-i}|\alpha,\eta)} \tag{3.25}$$

Ο αριθμητής της παραπάνω εξίσωσης είναι αυτός που φαίνεται στη (**3.22**). Ο παρανομαστής προκύπτει άμεσα καθώς είναι ίδιος σαν τύπος με τον αριθμητή μόνο που όλα τα μεγέθη δεν περιέχουν το i-οστό στοιχείο. Δηλαδή:

$$p(z_{-i}, w_{-i}|\alpha, \eta) = \prod_k \frac{\Delta_V(\eta + tc_k^{-i})}{\Delta_V(\eta)} \prod_d \frac{\Delta_K(a + wt_d^{-i})}{\Delta_K(a)} \tag{3.26}$$

Με τις (**3.25, 3.26**) και κάνοντας κάποιες αλγεβρικές απλοποιήσεις λόγων των συσχετισμών μεταξύ αυτών και των ιδιοτήτων της συνάρτησης *Γάμμα* που δεν θα παραθέσουμε για λόγους ευκρίνειας[104],το τελικό αποτέλεσμα προκύπτει ως:

$$\begin{aligned} p(z_i|z_{-i}, w) &= \prod_d \frac{\Delta_K(wt_d + \alpha)}{\Delta_K(wt_d^{-i} + \alpha)} \prod_k \frac{\Delta_V(tc_k + \eta)}{\Delta_V(tc_k^- i + \eta)} \\ &= \frac{\Gamma(wt_{d,k} + \alpha_k)\Gamma(\sum_{k=1}^K wt_{d,k}^{-i} + \alpha_k)}{\Gamma(wt_{d,k}^{-i} + \alpha_k)\Gamma(\sum_{k=1}^K wt_{d,k} + \alpha_k)} \frac{\Gamma(tc_{k,v} + \eta_v)\Gamma(\sum_{v=1}^V tc_{k,v}^{-i} + \eta_v)}{\Gamma(tc_{k,v}^{-i} + \eta_v)\Gamma(\sum_{v=1}^V tc_{k,v} + \eta_v)} \\ &= \frac{wt_{d,k} + \alpha_k - 1}{[\sum_{k=1}^K wt_{d,k} + \alpha_k] - 1} \frac{ct_{k,v} + \eta_v - 1}{[\sum_v = 1^V ct_{k,v} + \eta_v] - 1} \end{aligned} \tag{3.27}$$

Έχοντας λοιπόν τους μετρητές wt, ct από τη προηγούμενη διαδικασία και επειδή τα μεγέθη θ_k, β_k είναι οι ύστερες κατανομές και συζυγείς των πρότερων Dirichlet κατανομών άμεσα προκύπτει όπως δείξαμε και προηγουμένως:

$$p(\theta_d|z, w, \alpha) = Dir(\theta_d|wt_d + \alpha)$$

$$p(\beta_k|z, w, \eta) = Dir(\beta_k|ct_k + \eta)$$

Παίρνοντας την αναμενόμενη τιμή (expectation) των παραπάνω και σύμφωνα με την (**3.10**) για την Dirichlet κατανομή προκύπτουν τελικά οι τιμές:

$$\mathbb{E}[p(\theta_d|z, w, \alpha)] = \mathbb{E}[Dir(\theta_d|wt_d + \alpha)] = \hat{\theta_{d,k}} = \frac{wt_{d.k} + \alpha_k}{\sum_{k=1}^K (wt_{d,k} + \alpha_k)} \tag{3.28}$$

$$\mathbb{E}[p(\beta_k|z, w, \eta)] = \mathbb{E}[Dir(\beta_k|ct_k + \eta_v)] = \hat{\beta_{k,v}} = \frac{ct_{k,v} + \eta_v}{\sum_{v=1}^V (ct_{k,v} + \eta_v)} \tag{3.29}$$

Έχοντας λοιπόν τις εξισώσεις (**3.27,3.28,3.29**) είμαστε σε θέση να κατασκευάσουμε τον αλγόριθμο Collapsed Gibbs Sampling για τη μέθοδο μας. Αυτός φαίνεται στο (**Σχήμα 3.8**).

Algorithm 1: Collapsed LDA Gibbs Sampling Algorithm

Data: words $\in$ document collection **d**
Result: topic assignments $z_{d,n}$ and counts $tc_{k,v}, wt_{d,k}, sumk_k, sumd_d$
begin
  Randomly initialize $z_{d,n}$ and increment counters accordingly;
  while *not converged* **do**
    for $d = 0 \rightarrow D\text{-}1$ **do**
      for $n = 0 \rightarrow N_d\text{-}1$ **do**
        Firstly decrement counts for not accounting n-th element
        $wt_{d_{d,n},z_{d,n}} -= 1$;
        $tc_{z_{d,n},w_n} -= 1$;
        $sumk_{z_{d,n}} -= 1$;
        $sumd_d -= 1$;
        Sample new $\overline{z_{d,n}}$ according to updated probabilies from (3.27);
        Increment counts using generated $\overline{z_{d,n}}$
        $wt_{d_{d,n},\overline{z_{d,n}}} += 1$;
        $tc_{\overline{z_{d,n}},w_n} += 1$;
        $sumk_{\overline{z_{d,n}}} += 1$;
        $sumd_d += 1$;
      end
    end
  end
  Convergence criterion passed
  Read out lagged samples of θ_d, β_k according to (3.28,3.29)
end

Σχήμα 3.8: Τυπικός αλγόριθμος Collapsed Gibbs Sampling για την LDA

Βλέποντας τον αλγόριθμο, οι μεταβλητές που χρησιμοποιεί είναι 5, με τις 2 βοηθητικές. Η μία είναι η state space μεταβλητή της αλυσίδας η $z_{d,n}$ που ενσωματώνει τις αναθέσεις σε topics των λέξεων των κειμένων. Οι άλλες δύο είναι οι $wt_{d,k}, ct_{k,v}$, με διαστάσεις $D \times K, K \times V$ αντίστοιχα, εκ των οποίων η $wt_{d,k}$ μετράει τις φορές που οι λέξεις στο κείμενο d ανατέθηκαν στο topic k και η $ct_{k,v}$ τις φορές που η λέξη v ανατέθηκε στο θέμα k. Οι μεταβλητές $sumk_k, sumd_d$ είναι απλά τα αθροίσματα των γραμμών των δύο προηγούμενων πινάκων, καθώς χρησιμοποιούνται για τον υπολογισμό της (3.27).

Η διαδικασία που ακολουθείται είναι ως εξής: Αρχικά, αναθέτουμε τυχαία τις λέξεις στα topics (παίρνουμε τυχαία δείγματα της $z_{d,n} \sim Mult(\frac{1}{K})$) και με βάση αυτές τις τιμές αυξάνουμε τις τιμές των μετρητών μεταβλητών. Αυτή είναι η φάση της αρχικοποίησης. Στη συνέχεια περνάμε στη φάση burn-in της αλυσίδας Markov και κάνουμε επαναληπτικά την ίδια διαδικασία. Δηλαδή, για κάθε μία λέξη μειώνουμε αρχικά τα κελιά των διαφόρων μετρητών

που αφορούν στη λέξη αυτή(και το topic με το οποίο είναι συνδεδεμένη μέσω του υπάρχοντος $z_{d,n}$), για να μην εμπεριέχεται το n-οστό στοιχείο στον υπολογισμό της (3.27) για την νέα τιμή $\overline{z_{d,n}}$. Στη συνέχεια, αυξάνουμε τους μετρητές, στα νέα κελιά ανάλογα με τη τιμή του $\overline{z_{d,n}}$. Αυτή τη διαδικασία την επαναλαμβάνουμε μέχρι τη σύγκλιση της αλυσίδας στην steady state distribution.Υπάρχουν διάφορα κριτήρια για αυτό[55], αλλά στην περίπτωσή μας κοιτάμε τη διαφορά στο perplexity(ένα μέτρο που ποιοτικά εκφράζει την έκπληξη του μοντέλου καθώς βλέπει δείγματα της συλλογής μας) του μοντέλου από επανάληψη σε επανάληψη και όταν η διαφορά γίνει επαρκώς μικρή, θεωρούμε ότι έχουμε σύγκλιση. Τελικά, λαμβάνουμε αποτελέσματα για τα θ, β, z ανά τακτά χρονικά διαστήματα για να αποφύγουμε τη συσχέτιση μεταξύ των δειγμάτων[33], διαδικασία γνωστή ως lag-sampling ή thinning-out sampling.

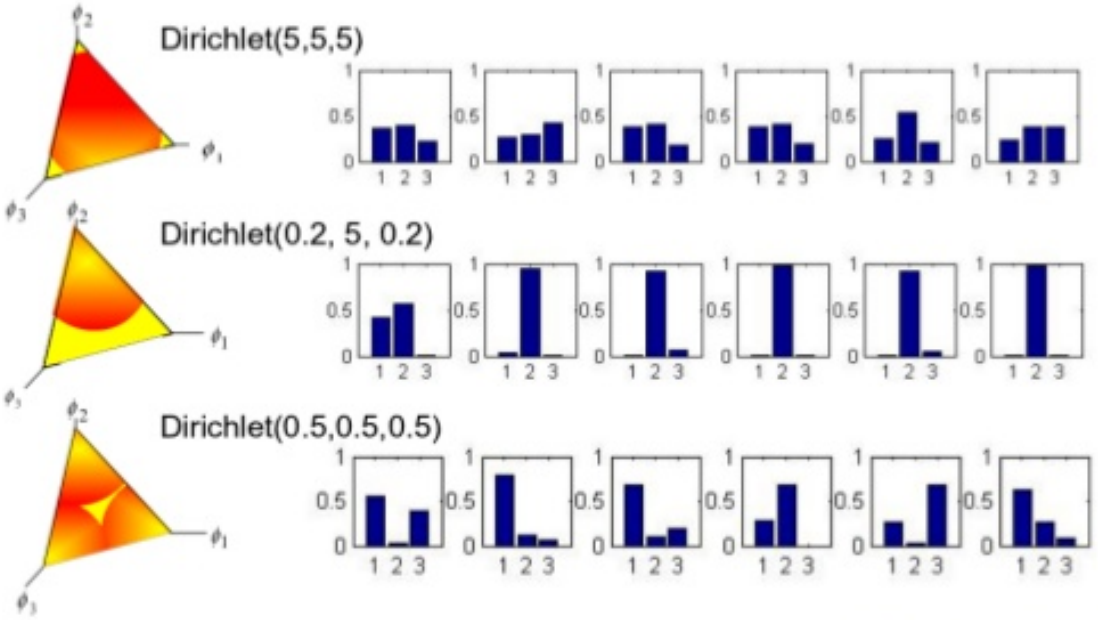
Ποιοτικά με την εξίσωση (**3.27**) αυτό που υπολογίζουμε με τη δειγματοληψία αυτή είναι η πιθανότητα μία συγκεκριμένη λέξη $w_{d,n}$ να ανήκει σε ένα θέμα j, δηλαδή την $p(z_{d,n} = j|z_{-(d,n)}, w)$, και οι παράγοντες που το επηρεάζουν είναι ακριβώς τα δύο αυτά κλάσματα. Το αριστερό κλάσμα εκφράζει τη πιθανότητα το θέμα j να εμφανιστεί στο κείμενο d σύμφωνα με τη κατανομή θεμάτων που υπάρχει ως τώρα και το δεξί κλάσμα τη πιθανότητα η λέξη $w_{d,n}$ να ανήκει στο θέμα αυτό. Όταν έχουμε αναθέσει πολλές λέξεις σε ένα topic, για ένα κείμενο πάντα, οι μετρητές στο δεξί κλάσμα αυξάνουν, αυξάνοντας έτσι τη πιθανότητα να αναθέσουμε μία νέα λέξη στο θέμα αυτό(στρατηγική γνωστή ως rich get richer). Συγχρόνως, όσο περισσότερο αναθέτουμε λέξεις στο συγκεκριμένο κείμενο σε ένα θέμα, τόσο περισσότερο πιθανό είναι και άλλες λέξεις να προστεθούν σε αυτό. Επομένως, ποιοτικά η επιλογή θέματος για κάθε λέξη γίνεται ανάλογα με το πόσο πιθανή είναι η λέξη να εμφανιστεί σε ένα θέμα και πόσο συχνά εμφανίζεται το συγκεκριμένο θέμα στο κείμενο που βρίσκεται η λέξη.

Τέλος, ένα σημείο που πρέπει να τονιστεί είναι η διαδικασία που ακολουθούμε με τη μείωση πρώτα των μετρητών κατά 1, sampling του $z_{d,n}$ και ξανά αύξηση των νέων κελιών. Αυτό συμβαίνει για να εκφραστεί ακριβώς η σημασία ότι η δειγματοληψία γίνεται για κάθε μεταβλητή, λαμβάνοντας υπ'όψιν μόνο τις υπόλοιπες μεταβλητές και όχι τη συγκεκριμένη ($z_{d,n}$). Δηλαδή αφαιρούμε 1 από το πλήθος των εμφανίσεων της συγκεκριμένης λέξης στο θέμα που άνηκε ως τώρα και 1 από το πλήθος εμφανίσεων του συγκεκριμένου topic στο κείμενο που βρίσκεται η λέξη. Στη συνέχεια παίρνουμε δείγμα από την $p(z_i|z_{-i}, w)$ για τη νέα τιμή του $\overline{z_{d,n}}$ και αυξάνουμε τους μετρητές στις κατάλληλες νέες θέσεις. Για αυτό το λόγο υπάρχει και το -1 στο τελικό τύπο της (**3.27**). Η ιδιότητα που μας επιτρέπει να το κάνουμε αυτό επαναληπτικά

για όλα τα z_i είναι η exchangeability των μεταβλητών αυτών, δηλαδή η αδιαφορία που παρουσιάζει η τελική κατανομή στη σειρά εμφάνισης των μεταβλητών. Στην περίπτωσή μας είναι το μοντέλο bag of words που περιγράψαμε.

Γενικές Παρατηρήσεις

Έχοντας ολοκληρώσει την περιγραφή του συστήματος θα κάνουμε μερικές γενικές παρατηρήσεις σχετικά με διάφορες πτυχές της μεθόδου. Αρχικά θα αναφερθούμε στις ιδιότητες της Dirichlet κατανομής που ως prior παίζει ρόλο στο τελικό αποτέλεσμα. Αναφέραμε πιο πάνω χωρίς περαιτέρω εξήγηση ότι χρησιμοποιούμε συμμετρικές τιμές για τα α, η. Αυτό οφείλεται στην επιρροή που έχουν στην οικογένεια κατανομών από στις οποίες κάνουμε δειγματοληψία. Για να γίνει αυτή πιο κατανοητή παραθέτουμε το **Σχήμα 3.9**, στο οποίο φαίνονται 3 τριδιάστατες κατανομές Dirichlet με διαφορετικά α, τοποθετημένες πάνω στο διδιάστατο simplex, από τις οποίες παράγονται διάφορες κατανομές πάνω σε αυτά τα τρία topics.



Σχήμα 3.9: Παραδείγματα τριδιάστατων κατανομών Dirichlet. Στα αριστερά φαίνονται οι τιμές των α και η αντίστοιχη κατανομή Dirichlet στο διδιάστατο simplex. Στα δεξιά φαίνονται διάφορα δείγματα από αυτές τις κατανομές.

Οι τιμές που παίρνει το διάνυσμα α είναι σαν δυνάμεις που μετακινούν την κατανομή, με μεγάλες τιμές $\alpha > 1$ να μετακινούν την κατανομή μακρυά από τα άκρα και προς το κέντρο, και κάνοντάς την πιο συγκεντρωμένη, προς το κέντρο(όπως φαίνεται στη πρώτη σειρά). Για ασύμμετρες τιμές στα α όπως η δεύτερη, βλέπουμε έντονη ασυμμετρία με την κατεύθυνση

στην πιο μεγάλη τιμή του α. Για $a < 1$ όμως βλέπουμε ότι η κατανομή εξωθείται προς τα άκρα, παράγοντας έτσι κατανομές πιο αραιές(sparse). Στην περίπτωσή μας τέτοιες κατανομές θέλουμε γιατί προτιμάμε κάθε κείμενο να ενεργοποιεί λίγα topics. Συγκεκριμένα, όσο μικρότερη είναι η τιμή των α τόσο πιο αραιή γίνεται η κατανομή των topics. Μειώνοντας δηλαδή την τιμή του α παίρνουμε κατανομές με μικρότερη εντροπία[73].

Αντίστοιχα, μικρότερη τιμή η σημαίνει ότι έχουμε πιο αραιή κατανομή των θεμάτων πάνω στις λέξεις, ενώ υψηλή σημαίνει διασπορά των λέξεων[94]. Επίσης, όπως είδαμε και προηγουμένως και από τους τύπους (**3.28,3.29**), οι μεταβλητές α, η μπορούν να λογιστούν ως οι πρότερες μετρήσεις για το πλήθος των εμφανίσεων των θεμάτων σε κείμενα και των λέξεων σε θέματα αντίστοιχα, προτού δούμε κάποιο από τα πραγματικά δεδομένα. Επιλέγοντας συμμετρικές τιμές για αυτά, θεωρούμε ότι κάθε topic έχει την ίδια πιθανότητα να ανήκει σε ένα κείμενο(για τη περίπτωση του α) και ότι κάθε λέξη έχει την ίδια πιθανότητα να ανατεθεί σε κάθε topic(για τη περίπτωση του η). Από την (**3.28**) είναι φανερό επίσης ότι όσο περισσότερα δεδομένα βλέπει ο αλγόριθμος, τόσο θα αυξάνει το $wt_{d.k} + \alpha_k$, οπότε τόσο θα αυξάνει και το θ_k, άρα πηγαίνουμε σε πιο συγκεντρωμένες, απ'ότι στο προηγούμενο βήμα της εκπαίδευσης, κατανομές. Δηλαδή, μετακινούμαστε από την prior εκτίμηση που είχαμε για τα δεδομένα μας $Dir(a)$(τι φανταζόμασταν ότι ισχύει) σε πιο σίγουρες εκτιμήσεις(τι ισχύει πραγματικά), όσο πιο πολλά δεδομένα έχουμε, πράγμα το οποίο είναι διαισθητικά πολύ ορθό.

Αναφερόμενοι στα προηγούμενα προσεγγίζουμε το θέμα επιλογής των υπερπαραμέτρων της μεθόδου οι οποίοι είναι τα α, η, K, δηλαδή η τιμή των Dirichlet prior και ο αριθμός των topics. Οι τρεις αυτοί παράμετροι έχουν σχέση μεταξύ τους και επηρεάζουν σημαντικά τη ποιότητα των αποτελεσμάτων. Όσον αφορά τα α, η, υπάρχουν διάφορες προσεγγίσεις που προωθούν ασύμμετρες τιμές για τους α συντελεστές[101] με βάση ότι το μοντέλο μας θέλουμε να αντιλαμβάνεται τις σημασιολογικά κοντινές συνεμφανίσεις των λέξεων. Δηλαδή, μας ενδιαφέρει όχι μόνο να εμφανίζονται συχνά κάποιες λέξεις, αλλά να χρησιμοποιούνται και στο ίδιο πλαίσιο-θέμα. Οπότε ασύμμετρες κατανομές θεμάτων θα οδηγούσαν σε πιο έντονες συσχετίσεις μεταξύ τους, όταν αυτές υπήρχαν, καθώς συγκεκριμένα θέματα θα ήταν μεν κοινά για τα περισσότερα κείμενα(αυτά με υψηλή τιμή α) αλλά όταν κάποιο κείμενο ενεργοποιούσε κάποιο άλλο θέμα, θα ήταν σημασιολογικά σχετικό με το συγκεκριμένο κείμενο. Από την άλλη, *(heuristic)* μέθοδοι αποτυπώνουν ως καλές τιμές τις: $\alpha = \frac{50}{K}, \eta = 0.01$[38]. Η υλοποίηση στο δικό μας σύστημα κάνει *βελτιστοποίηση υπερπαραμέτρων*(hyperparameter optimization)

για αυτά τα δύο μεγέθη[68, 9].

Όσον αφορά στον αριθμό των θεμάτων K, υπάρχουν πολλές διαφορετικές ιδέες για κάποιο φορμαλιστικό τρόπο επιλογής του. Υπάρχουν δύο διαφορετικές προσεγγίσεις κυρίως, είτε μετράμε τη ποιότητα των θεμάτων που παράγονται, είτε θέτουμε κάποιο εξωγενές μετρικό για το σύστημα μας. Ιδέες σαν τη πρώτη πολλές φορές εμπλέκουν και την ανθρώπινη παρατήρηση, όπως οι πολύ ενδιαφέρουσες ιδέες του word-topic intrusions όπου εξωτερικοί χρήστες κρίνουν τη ποιότητα των θεμάτων είτε τελικά[20] είτε κατά τη διάρκεια της εκπαίδευσης[47]. Ενδιαφέρουσα είναι και η πιο αυτοματοποιημένη ιδέα για τη συνεκτικότητα των θεμάτων μέσω του μετρικού topic coherence[16] ή την απόκλιση(divergence) από κάποια νόρμα κατανομών, ανάλογα με το πλήθος των θεμάτων[62, 37]. Κάποιες από αυτές τις ιδέες τις υλοποιήσαμε και εμείς στο πλαίσιο της εργασίας αλλά δεν τις κρατήσαμε ως μετρικά επειδή δεν ήταν κατάλληλα για το σκοπό που δημιουργήσαμε το σύστημά μας. Πάντως, μπορεί να συμπεράνει κανείς προσεγγιστικά ένα κατάλληλο πλήθος των θεμάτων, αναλόγως το πεδίο εφαρμογής του και την ομοιότητα των δεδομένων του. Έχει δειχθεί επίσης ότι το μικρότερο πλήθος γενικώς δίνει πιο εύρωστα αποτελέσματα, όπου το πόσο μικρό ή μεγάλο είναι το πλήθος εξαρτάται πάντα από το πλαίσιο της εργασίας μας[38]. Στη δική μας προσέγγιση αναφερόμαστε στο επόμενο κεφάλαιο με λεπτομέρεια με ποιο τρόπο διαλέξαμε τον αριθμό των θεμάτων.

Εναλλακτική προσέγγιση στο θέμα προέρχεται από ιδέες *non-parametric bayesian* μοντέλων. Οι πιο βασικές είναι από αυτές που εμπεριέχουν τις λεγόμενες *Hierarchical Dirichlet Processes*[96], στις οποίες δεν δίνουμε εμείς των αριθμό των θεμάτων αλλά προκύπτει στο στάδιο μάθησης του αλγορίθμου. Κοντινή ιδέα είναι και αυτή των *εμφωλευμένων*(nested) θεμάτων, όπου υπάρχει μία ιεραρχική δομή στα θέματα, με κάποια θέματα κοινά για το μεγαλύτερο πλήθος κειμένων[12]. Άλλη προσέγγιση είναι αυτή των *correlated topic models*[11] ή μοντέλων που προκύπτουν με *pachinko allocation*[56], όπου εντοπίζουν τη συσχέτιση κατά την εμφάνιση διαφόρων ζευγών topics. Μια ακόμα διαφορετική ιδέα είναι αυτή των *spherical topic models*, όπου επιτρέπει στις λέξεις να εμπεριέχουν και ένα βαθμό που δείχνει πόσο απίθανες είναι να συμπεριληφθούν σε ένα θέμα[80].

Κλείνοντας με τις βιβλιογραφικές αναφορές, να επισημάνουμε τις ρίζες της LDA μεθόδου. Πρώτη προσέγγιση έπειτα από την tf-idf, ήταν αυτή της *Latent Semantic Indexing-LSI*[22], όπου προβάλλεται ο πίνακας βαρών της tf-idf μεθόδου σε υποχώρο μικρότερης διάστασης, ο οποίος περιέχει το μεγαλύτερο μέρος της πληροφορίας. Η συνέχεια ήρθε με την *probabilistic*

Latent Semantic Indexing-pLSI[44], όπου μοντελοποιείται πιθανοτικά η παραγωγή των λέξεων από διάφορα θέματα, αλλά δε προσφέρει πιθανοτική ανάλυση στο επίπεδο των κειμένων, μειώνοντας τις δυνατότητες γενίκευσης της μεθόδου. Ενδιαφέρον παρουσιάζει ότι η σύλληψη της LDA έγινε σε τελείως διαφορετική θεματολογία, συγκεκριμένα στην ανάλυση πληθυσμών ως μείγμα των γονιδίων τους[77]. Τέλος, πλούσιες αναφορές υπάρχουν στη διαδικασία Gibbs sampling[30, 18, 71] και στην εφαρμογή της για την LDA μέθοδο[93, 40].

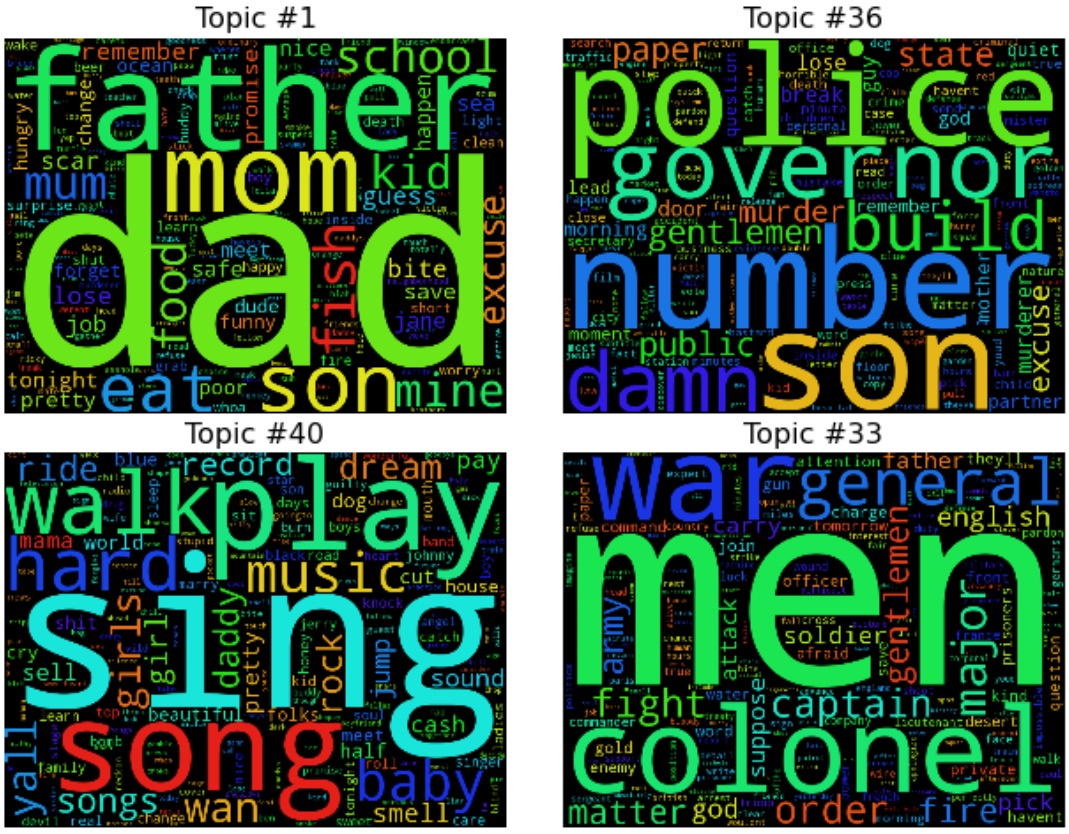
Ολοκληρώνοντας αυτή την ενότητα, θα κάνουμε μία ποιοτική αναφορά στην LDA και τα διάφορα πλεονεκτήματα της. Ο βασικός λόγος για τον οποίο έχουμε καλά αποτελέσματα με την LDA, είναι ότι μέσω των πρότερων Dirichlet κατανομών ωθούμε το σύστημα μας να μοντελοποιεί κάθε κείμενο με σχετικά λίγα θέματα. Άρα, επιδιώκουμε αραιές κατανομές θεμάτων στα θέματα, με μικρές τιμές $\alpha < 1$. Αυτό, σε συνδυασμό με την δυνατότητα που έχουμε να βρίσκουμε λέξεις που συνεμφανίζονται σε κείμενα, οπότε μπορούν ίσως να σχηματίσουν θέματα, μας οδηγεί σε μία ελάττωση της σημαντικότητας αυτών των συνυπάρξεων των λέξεων, οδηγώντας μας σε πιο συνεκτικά θέματα. Δηλαδή, δεν κοιτάμε μόνο για λέξεις που εμφανίζονται συχνά μαζί, όπως στις προγενέστερες μεθόδους, αλλά και την συνεμφάνιση θεμάτων σε κείμενα, πράγμα που μας οδηγεί σε θέματα με πιο συνεκτικά δεμένες λέξεις και καλύτερη αναπαράσταση των κειμένων.

Εκτός των άλλων μας προσφέρει μεγάλη μείωση της διάστασης του προβλήματός μας, καθώς μεταφέρει την αναπαράσταση των κειμένων από το $V-1$-simplex του λεξιλογίου στο $K-1$-simplex των θεμάτων. Αυτό είναι κατά κάποιο τρόπο μία μορφή Principal Component Analysis για τα δεδομένα μας. Επιπροσθέτως, η LDA μας δίνει ένα πολύ καλό τρόπο να αντιμετωπίσουμε το πρόβλημα της πολυσημίας, δηλαδή τη χρήση μίας λέξης με διαφορετικά νοήματα σε κάθε κείμενο(π.χ. "theatric play"-"play ball"). Αυτό γιατί μας επιτρέπει να ορίσουμε το πλαίσιο στο οποίο χρησιμοποιείται η λέξη, ακριβώς μέσω των θεμάτων. Επίσης, δεν είναι αποκλειστικά εφαρμόσιμη μέθοδος για κείμενο, αλλά μπορούμε να τη χρησιμοποιήσουμε για εικόνες, ήχους κ.α. αλλάζοντας πολύ λίγα πράγματα από τον αλγόριθμο(όπως τις πρότερες κατανομές). Τέλος, το τελικό αποτέλεσμα, δηλαδή η θεματική δομή των αντικειμένων που εξετάζουμε μπορεί να χρησιμοποιηθεί με πολλούς τρόπους. Μπορούμε να τη χρησιμοποιήσουμε για να εξετάσουμε ομοιότητα των αντικειμένων μέσω της θεματικής τους συνάφειας(document similarity), μπορούμε να ανασύρουμε πληροφορίες από αυτή τη δομή αναπαράστασης γνώσης(query answering και information retrieval) ή ακόμα μπορούμε να

οπτικοποιήσουμε τη σχέση μεταξύ των θεμάτων και των αντικειμένων που αφορούν σε μία πιο ποιοτική περιήγηση στο θεματικό χώρο αυτών.

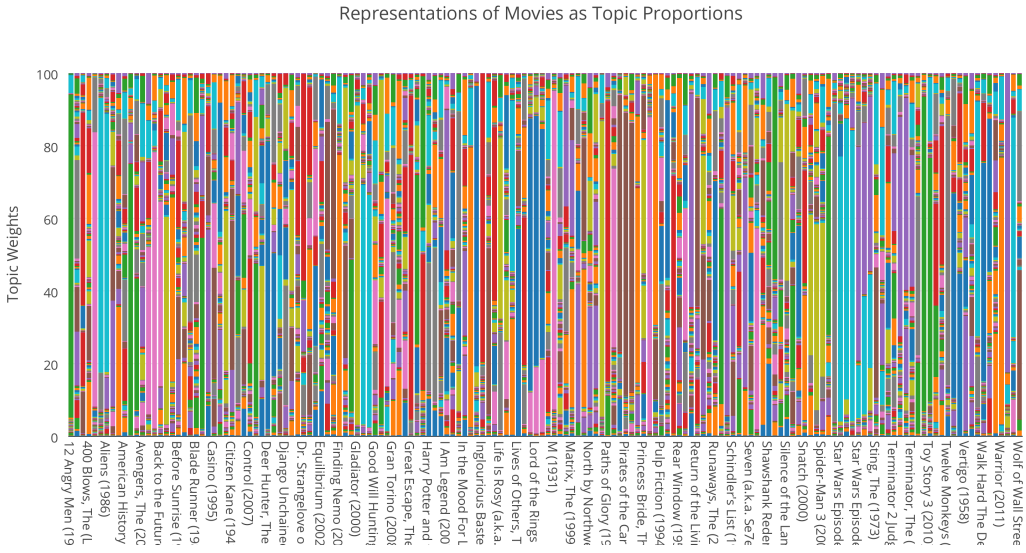
Παραδείγματα επί των υποτίτλων

Τελειώνοντας με το κεφάλαιο της LDA θα παραθέσουμε δύο παραδείγματα στα δεδομένα μας για να οπτικοποιήσουμε και τα λεγόμενά μας σε αυτή την ενότητα. Αρχικά, παραθέτουμε στο **Σχήμα 3.10** 4 διαφορετικά θέματα όπως αυτά προέκυψαν από την LDA μέθοδο στους υπότιτλους που είχαμε στη διάθεση μας. Σε αυτά τα νέφη λέξεων, οι λέξεις έχουν μέγεθος ανάλογο με τη συχνότητα εμφάνισης τους στα θέματα αυτά. Παρατηρεί κανείς ότι το σύστημα μας έχει δημιουργήσει θέματα αρκετά συνεκτικά από άποψη εννοιών των λέξεων. Το πάνω αριστερά νέφος για παράδειγμα, εκφράζει το θέμα της οικογένειας με λέξεις όπως father, mom, kid, school κ.α. να είναι οι πιο σημαντικές για το θέμα αυτό ενώ το κάτω δεξιά έχει συλλάβει έννοιες σχετικές με πόλεμο, με τις πιο σημαίνουσες λέξεις να είναι men, general, war, fire, order κ.α.



Σχήμα 3.10: 4 διαφορετικά θέματα του συστήματός μας. Το μέγεθος της κάθε λέξης είναι συνάρτηση της βαρύτητας της για το θέμα, δηλαδή της πιθανότητας εμφάνισης αυτής.

Τέλος, το επόμενο γράφημα(**Σχήμα 3.11**), απεικονίζει σε μορφή σταθμισμένων ιστογραμμάτων τις ταινίες με βάση τα ποσοστά συμμετοχής των θεμάτων σε κάθε μία από αυτές. Στον οριζόντιο άξονα είναι οι ταινίες της βάσης δεδομένων μας και στον κατακόρυφο είναι οι αναπαραστάσεις αυτών ως μείγματα θεμάτων. Κάθε ταινία εκφράζεται ως ένας συνδυασμός θεμάτων και το αντίστοιχο ιστόγραμμα απαρτίζεται από τα θέματα από τα οποία αποτελείται η ταινία. Κάθε θέμα έχει διαφορετικό χρώμα στο πίνακα και το μέγεθος του θέματος στο ιστόγραμμα της ταινίας εκφράζει το ποσοστό συμμετοχής του (ως επί τοις εκατό) στην ταινία αυτή.



Σχήμα 3.11: Αναπαράσταση των ταινιών ως μείγματα θεμάτων. Κάθε ταινία, στον οριζόντιο άξονα, εκφράζεται ως μείγμα των θεμάτων που την αποτελούν. Τα θέματα έχουν διαφορετικά χρώματα το καθένα και βρίσκονται το ένα πάνω στο άλλο ως ιστογράμματα. Η έκταση κάθε θέματος εκφράζει το ποσοστό συμμετοχής του θέματος στη ταινία αυτή.

Αν και εδώ έχουμε μία στατική εικόνα και δεν είναι δυνατή η διάδραση του χρήστη με το γράφημα, εντούτοις μπορεί να παρατηρήσει κανείς ότι σε σειρές ταινιών όπως οι ταινίες Star Wars ή The Lord of The Rings, ένα θέμα-χρώμα αποτελεί τον κύριο εκφραστή αυτών και είναι κοινό για όλες τις ταινίες της σειράς. Άρα μπορεί να συμπεράνει κάποιος ότι αυτό το θέμα έχει ως αντικείμενο κάτι σχετικό με τις ταινίες και επίσης να συμπεράνει ότι οι ταινίες αυτές ταιριάζουν μεταξύ τους λόγω της μεγάλης τιμής συμμετοχής του θέματος σε κάθε μία από αυτές. Αυτή ακριβώς η θεματική πλοήγηση στο χώρο των ταινιών και η συσχέτιση μεταξύ αυτών λόγω συγκεκριμένων θεμάτων είναι ένα από τα μεγάλα προτερήματα χρήσης αυτής της μεθόδου στο αντικείμενο που ασχολούμαστε.

3.3 Ανάλυση Ήχου

Σε αυτό το κεφάλαιο θα ασχοληθούμε με το κομμάτι της ανάλυσης του ήχου από τις ταινίες μας. Όπως αναφέραμε και προηγουμένως, η πληροφορία προερχόμενη από το ηχητικό σκέλος των ταινιών λαμβάνει βοηθητικό ρόλο, ως προς το κύριο σκέλος που είναι η ανάλυση κειμένου. Για αυτό το λόγο θα σκιαγραφήσουμε τα βασικά σκέλη της μεθόδου, χωρίς να μπούμε σε μεγάλες λεπτομέρειες. Να τονιστεί εδώ ότι ολόκληρη η ανάλυση του ήχου βασίστηκε στη βιβλιοθήκη pyAudioAnalysis[2] και σε μέρος της διδακτορικής διατριβής [3], αμφότερα του Δρ. Θοδωρή Γιαννακόπουλου. Τα βασικά μέρη με τα οποία θα ασχοληθούμε είναι το κομμάτι *εξαγωγής χαρακτηριστικών*(feature extraction) από τον ήχο και το κομμάτι *ταξινόμησης*(classification) του ήχου σε είδη μουσικής(music) και ηχητικών συμβάντων(sound).

3.3.1 Εξαγωγή Χαρακτηριστικών

Ξεκινώντας θα αναφερθούμε στο ηχητικό κομμάτι κάθε ταινίας και το τρόπο εξαγωγής διαφόρων χαρακτηριστικών από αυτό. Η εξαγωγή χαρακτηριστικών για ταινίες γίνεται τμηματικά σε διάφορα *πλαίσια*(frames) του ηχητικού κομματιού. Πιο συγκεκριμένα, έστω $x(n), n = 1, ..., L$ το ηχητικό σήμα μίας ταινίας. Η συνήθης πρακτική για τον υπολογισμό χαρακτηριστικών από αυτό είναι να χωρίζεται αυτό το σήμα σε συνεχόμενα πλαίσια(τα οποία μπορεί να είναι *επικαλυπτόμενα*(overlapping) ή και όχι) και σε αυτά τα πλαίσια να υπολογίζουμε διάφορα χαρακτηριστικά. Ο λόγος που γίνεται αυτό είναι γιατί θεωρούμε ότι εν γένει το σήμα ήχου δεν είναι στατικό, δηλαδή τα στατιστικά που σχετίζονται με αυτό αλλάζουν με το χρόνο, οπότε για να εξάγουμε κάποια συμπεράσματα για αυτό θεωρούμε ένα μικρότερο χρονικό διάστημα στο οποίο συμπεριφέρεται ως στατικό(quasistationary)[97]. Ορίζουμε ένα *παράθυρο*(window) μήκους N ως $w(n)$. Το πιο απλό παράθυρο είναι το τετραγωνικό το οποίο είναι μονάδα εντός του μήκους N και μηδέν εκτός αυτού. Δηλαδή:

$$w(n) = \begin{cases} 1 & , 0 \leq n \leq N-1 \\ 0 & , elsewhere \end{cases} \tag{3.30}$$

Η διαδικασία *παραθυροποίησης*(windowing) του σήματος, γίνεται πολλαπλασιάζοντας επαναληπτικά το σήμα $x(n)$ με το παράθυρο της επιλογής μας, καθώς αυτό μετακινείται στο χρόνο και διασχίζει όλο το ηχητικό σήμα. Δηλαδή, χωρίζουμε το ηχητικό σήμα σε πλαίσια με το

[2]https://github.com/tyiannak/pyAudioAnalysis
[3]http://cgi.di.uoa.gr/ tyiannak/phdText.pdf

i-οστό πλαίσιο να είναι: $x_i(n') = x(n)w(n - m_i)$, όπου m_i είναι η μετατόπιση στο χρόνο που αντιστοιχεί στο πλαίσιο αυτό. Προφανώς οι τιμές του m_i εξαρτώνται από το μέγεθος του παραθύρου(window size) και το βήμα (step) μετακίνησης αυτού. Πρέπει να είναι αρκετά μεγάλο, για να μπορούμε να εξάγουμε χαρακτηριστικά από αυτό αλλά από την άλλη αρκετά μικρό για να θεωρηθεί περίπου στατικό. Εδώ οι δύο διαφοροποιήσεις είναι οι εξής:

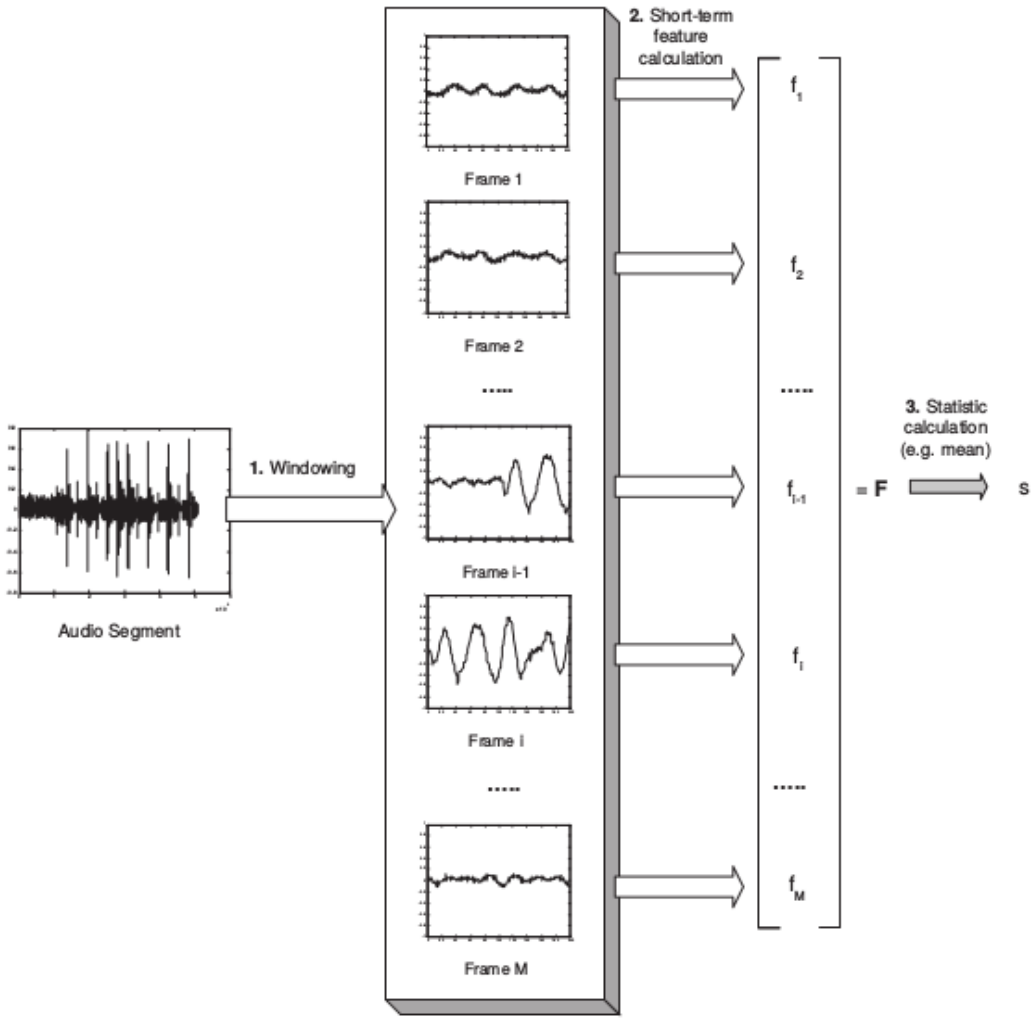
1. Σε βραχυπρόθεσμα παράθυρα επεξεργασίας (short-term processing windows) με συνήθη διάρκεια $10 - 50$msecs.

2. Σε μεσοπρόθεσμα παράθυρα επεξεργασίας (mid-term processing windows) με συνήθη διάρκεια $1 - 10$secs.

Ας θεωρήσουμε τα βραχυπρόθεσμα παράθυρα για αρχή. Αυτά χωρίζουν λοιπόν με τη παραπάνω μέθοδο το σήμα σε M διαφορετικά πλαίσια. Σε κάθε πλαίσιο από αυτά εμείς μπορούμε να υπολογίσουμε ένα χαρακτηριστικό έστω f, οπότε τελικά το σήμα μας θα αναπαρίσταται ως ένα διάνυσμα χαρακτηριστικών $F = [f_1, ..., f_M]$, το οποίο θα έχει τη τιμή του χαρακτηριστικού αυτού σε κάθε πλαίσιο. Όμως παράλληλα με τη βραχυπρόθεσμη επεξεργασία του σήματος, μπορούμε να κάνουμε και μεσοπρόθεσμη. Αυτό είναι δυνατόν αν αρχικά χωρίσουμε το σήμα σε μεσοπρόθεσμα πλαίσια, στη συνέχεια επιτελέσουμε την εξαγωγή των χαρακτηριστικών όπως παραπάνω σε κάθε βραχυπρόθεσμο πλαίσιο εντός του μεσοπρόθεσμου και τέλος να υπολογίσουμε ένα στατιστικό (όπως η μέση τιμή) για το διάνυσμα F που προέκυψε από τη βραχυπρόθεσμη επεξεργασία του σήματος. Άρα κάθε μεσοπρόθεσμο πλαίσιο τώρα εκπροσωπείται ως ένα στατιστικό των βραχυπρόθεσμων χαρακτηριστικών που το αποτελούν.

Αυτή η διαδικασία φαίνεται οπτικοποιημένη στο **Σχήμα 3.12**. Τώρα ας φανταστούμε ότι δεν υπολογίζουμε μόνο ένα χαρακτηριστικό ανά βραχυπρόθεσμο πλαίσιο αλλά παραπάνω, έστω K. Η μόνη αλλαγή είναι ότι το διάνυσμα F γίνεται $F = [f_{1,1..k}, ..., f_{M,1..K}]$ και αντίστοιχα για τα μεσοπρόθεσμα χαρακτηριστικά, πλέον υπολογίζουμε το αντίστοιχο στατιστικό για όλα τα διαφορετικά χαρακτηριστικά k. Αυτό γίνεται στη πραγματικότητα και στη δική μας προσέγγιση.

Συγκεκριμένα, τα χαρακτηριστικά που μπορούν να υπολογιστούν είναι δύο ειδών: είτε χαρακτηριστικά που εξάγονται απ'ευθείας από την αναπαράσταση του σήματος στο χρόνο, είτε εξάγονται στο χώρο των συχνοτήτων μετά το μετασχηματισμό του σήματος ήχου με βάση τον *Μετασχηματισμός Fourier Βραχέως Χρόνου*(Short Time Fourier Transform-SFTF)[88]. Τα



Σχήμα 3.12: Διαδικασία εξαγωγής μεσοπρόθεσμων χαρακτηριστικών. Στο σχήμα φαίνεται η διαδικασία που ακολουθείται, αρχικά με τη παραθυροποίηση σε μεσοπρόθεσμα πλαίσια, στη συνέχεια εξαγωγή των βραχυπρόθεσμων χαρακτηριστικών από αυτά και τέλος εξαγωγή του μεσοπρόθεσμου χαρακτηριστικού, ως στατιστικό των βραχυπρόθεσμων.

μεν πρώτα χαρακτηριστικά σχετίζονται με την ενέργεια κυρίως του σήματος, τα δε δεύτερα με ποιοτικές ιδιότητες του συχνοτικού περιεχομένου του σήματος. Σε αυτό το σημείο δεν θα μπούμε στη διαδικασία να αναλύσουμε όλα τα χαρακτηριστικά που θα εξάγουμε στο πλαίσιο της εργασίας, ούτε και την αιτιολογία της επιλογής τους. Είναι όλα ευρέως διαδεδομένα στη βιβλιογραφία και εκτενέστερη ανάλυση αυτών υπάρχει στη διδακτορική διατριβή που αναφέραμε και προηγουμένως.

Θα αρκεστούμε να τα παραθέσουμε ως πίνακα με μία πολύ σύντομη περιγραφή για το καθένα(**Πίνακας 3.1**). Όπως βλέπει κανείς εξάγουμε τελικώς 34 βραχυπρόθεσμα χαρακτηριστικά ανά σήμα ήχου. Τα μεσοπρόθεσμα χαρακτηριστικά, τώρα είναι δύο στατιστικά για κάθε

ένα βραχυπρόθεσμο χαρακτηριστικό, συγκεκριμένα η *μέση τιμή*(average value) και η *τυπική απόκλιση*(standard deviation), στο διάστημα του μεσοπρόθεσμου πλαισίου. Επομένως για κάθε μεσοπρόθεσμο πλαίσιο έχουμε $2 \times 34 = 68$ μεσοπρόθεσμα χαρακτηριστικά.

Feature ID	**Feature Name**	**Description**
1	Zero Crossing Rate	Ο ρυθμός που αλλάζει πρόσημο το σήμα κατά τη διάρκεια ενός συγκεκριμένου πλαισίου.
2	Energy	Το άθροισμα των τετραγώνων των τιμών του σήματος, κανονικοποιήμενο ως προς το μήκος του πλαισίου.
3	Entropy of Energy	Η εντροπία των κανονικοποιημένων ενεργειών των υπό πλαισίων. Είναι ένα μέτρο των απότομων αλλαγών.
4	Spectral Centroid	Το κέντρο βάρους του φάσματος.
5	Spectral Spread	Η διασπορά του φάσματος.
6	Spectral Entropy	Η εντροπία των κανονικοποιημένων φασματικών ενεργειών των υπο-πλαισίων.
7	Spectral Flux	Η τετραγωνική διαφορά ανάμεσα στα κανονικοποιημένα πλάτη των φασμάτων δύο διαδοχικών πλαισίων.
8	Spectral Rolloff	Η συχνότητα κάτω από την οποία βρίσκεται το 90% του πλάτους του φάσματος
9-21	MFCCs	Mel Frequency Cepstral Coefficients - Οι 13 πρώτοι συντελεστές Cepstrum του σήματος, στη κλίμακα συχνοτήτων Mel.
22-33	Chroma Vector	12-διάστατη αναπαράσταση της φασματικής ενέργειας, όπου το κάθε κελί αναπαριστά μία εκ των 12 κλάσεων τονικότητας, της δυτικού- τύπου μουσικής (ημιτονική απόσταση)
34	Chroma Deviation	Η τυπική απόκλιση των 12 προηγούμενων συντελεστών chroma.

Πίνακας 3.1: Το σύνολο των short-term χαρακτηριστικών

Για να γίνει βέβαια η παραπάνω ανάλυση των σημάτων ήχου θα πρέπει να ορίσουμε εμείς τα μήκη και τα βήματα των παραθύρων. Συγκεκριμένα εργαστήκαμε με short-term window, step : 50msec και για τα δύο και mid-term window, step : 10sec ομοίως και για τα δύο(δηλαδή και στο short-term και στο mid-term δεν έχουμε overlap μεταξύ των πλαισίων). Επίσης, στα τεχνικά μέρη της διαδικασίας έπειτα από την εξαγωγή του ήχου από τη ταινία, έγινε μετατροπή του σε μορφή .wav. Ο ρυθμός δειγματοληψίας ήταν 16kHz και έγινε μετατροπή, όπου χρειαζόταν, σε μονό κανάλι ήχου.

3.3.2 Ταξινόμηση Μουσικής και Ηχητικών Συμβάντων

Σύμφωνα με τα παραπάνω λοιπόν κάθε ταινία αναπαρίσταται ως ένας πίνακας διανυσμάτων ο οποίος έχει μέγεθος $M_d \times 68$, όπου M_d είναι το πλήθος μεσοπρόθεσμων πλαισίων ανά ταινία και 68 είναι τα μεσοπρόθεσμα χαρακτηριστικά για κάθε ένα πλαίσιο, όπως περιγράφηκαν προηγουμένως. Όμως εμείς δε θέλουμε οι ταινίες να αναπαριστώνται ως πίνακες χαρακτηριστικών, αλλά όπως και με την LDA μέθοδο, ως διανύσματα χαρακτηριστικών. Για να το πετύχουμε λοιπόν αυτό κάνουμε long-term averaging στα 68 αυτά διαφορετικά χαρακτηριστικά. Επομένως, παίρνουμε τη μέση τιμή κάθε χαρακτηριστικού πάνω σε όλα τα μεσοπρόθεσμα πλαίσια. Δηλαδή για κάθε χαρακτηριστικό i, αν $f_{i,1}, f_{i,2}, ..., f_{i,M_d}$ είναι στήλη i του παραπάνω πίνακα, εμείς παίρνουμε το μέσο όρο αυτής : $mean_{f_i} = \frac{\sum_{j=1}^{M_d} f_{i,j}}{M_d}$. Αυτό το κάνουμε για κάθε χαρακτηριστικό του παραπάνω πίνακα, οπότε επιτυγχάνουμε κάθε ταινία πράγματι να αναπαρίσταται ως ένα διάνυσμα χαρακτηριστικών 1×68, το οποίο περιέχει τη μέση τιμή των μεσοπρόθεσμων τιμών των αντίστοιχων χαρακτηριστικών.

Όμως θέλουμε μία πιο ποιοτική αναπαράσταση αυτών, καθώς δεν είναι διαισθητικά εύκολο να καταλάβουμε την ομοιότητα μεταξύ αυτών των διανυσμάτων, ώστε να συμπεράνουμε πράγματα για τη σχέση των ταινιών. Προς αυτό το σκοπό λοιπόν επιλέγουμε να ομαδοποιήσουμε τις ταινίες ως προς κάποια χαρακτηριστικά. Πιο συγκεκριμένα, μας ενδιαφέρουν τα διαφορετικά είδη της μουσικής κάθε ταινίας καθώς και τα διάφορα ηχητικά γεγονότα που ακούγονται σε αυτές. Θα ασχοληθούμε με κάθε μορφή πληροφορίας ξεχωριστά.

Ταξινόμηση Μουσικής

Αρχικά για το κομμάτι της μουσικής ορίζουμε εμείς αυθαίρετα 8 διαφορετικά είδη μουσικής ως πιθανές κλάσεις, που μπορεί να ανήκει ένας μουσικός ήχος. Αυτές οι 8 κλάσεις είναι

: Blues, Classical, Country, Electronic, Jazz, Rap, Reggae, Rock. Επιλέξαμε αυτές τις κλάσεις καθώς είναι είδη μουσικής που εμφανίζονται συχνά σε ταινίες. Θα μπορούσε βέβαια κάποιος να κάνει κάποια διαφορετική κατάτμηση του χώρου μουσικής σε είδη σύμφωνα με τη κρίση του. Μας ενδιαφέρει τώρα να εκφράσουμε κάθε ταινία ως μείγμα αυτών των ειδών μουσικής. Δηλαδή, στόχος μας είναι τελικά κάθε ταινία να αναπαρίσταται ως ένα διάνυσμα 8 διαστάσεων, που κάθε διάσταση θα αντιστοιχεί σε μία από αυτές τις κλάσεις και η τιμή της στη διάσταση αυτή θα δείχνει το ποσοστό από το οποίο αποτελείται η ταινία από αυτό το είδος μουσικής ως προς το σύνολο της. Για παράδειγμα μία ταινία με διάνυσμα χαρακτηριστικών μουσικής ως : $[0.5, 0, 0, 0, 0, 0, 0, 0.5]$ σημαίνει ότι αποτελείται κατά 50% από Blues μουσική και κατά 50% από Rock μουσική.

Για να το πετύχουμε αυτό θα πρέπει να φτιάξουμε ένα σύστημα το οποίο θα δέχεται στην είσοδο του τη ταινία σε μορφή ηχητικού σήματος και να μας δίνει στην έξοδο μία αναπαράσταση 1×8, όπως περιγράψαμε παραπάνω. Η μόνη διαφοροποίηση με την παραπάνω περιγραφή είναι ότι το ηχητικό σήμα της ταινίας, όπως εξάγεται αυτούσιο από τη ταινία, περιέχει πέρα από μουσική και ομιλία και διάφορους περιβαλλοντικούς ήχους. Οπότε, για να μπορέσουμε να κάνουμε ταξινόμηση των διαφορετικών μουσικών ήχων σε μία ταινία, πρέπει πρώτα να μπορούμε να ξεχωρίσουμε τη μουσική από τους υπόλοιπους ήχους της ταινίας. Για να το πετύχουμε αυτό, δίνουμε στο σύστημα μας ως είσοδο όλο το ηχητικό σήμα της ταινίας και ξεχωρίζουμε τις περιοχές του σήματος που αποτελούνται σίγουρα από μουσική. Για να γίνει αυτό δυνατό, χρησιμοποιείται μία προεκπαιδευμένη *Μηχανή Διανυσμάτων Υποστήριξης*(Support Vector Machine-SVM), η οποία είναι σε θέση να ξεχωρίσει ήχο μουσικής από οτιδήποτε άλλο(περισσότερα για τη λειτουργία των μηχανών διανυσμάτων υποστήριξης στη συνέχεια). Τις περιοχές λοιπόν του σήματος που αποτελούν πραγματικά μόνο μουσική και έχουν διάρκεια $\geq$ 10sec, τις επεξεργαζόμαστε όπως αναλύσαμε, με την εξαγωγή των ηχητικών χαρακτηριστικών και το long-term averaging αυτών.

Επισημαίνουμε, σε αυτό το σημείο ότι επιλέγουμε να χρησιμοποιήσουμε τα long-term χαρακτηριστικά κάθε ταινίας και όχι τα βραχυπρόθεσμα για το σκοπό της ταξινόμησης, για δύο λόγους: α) για να αποφύγουμε ταξινομήσεις που οφείλονται σε βραχυπρόθεσμα χαρακτηριστικά και δεν εκφράζουν το σύνολο του σήματος και β) για να εκμεταλλευτούμε το γεγονός ότι υπάρχει συνέχεια και συνάφεια περιεχομένου μεταξύ διαδοχικών βραχυπρόθεσμων πλαισίων μουσικής(δηλαδή ότι η μουσική είναι μία συνεχής διαδικασία). Έχουμε λοιπόν τώρα

στη διάθεση μας, το πλήθος των κομματιών μουσικής κάθε ταινίας, τα οποία αναπαριστώνται από τα 1×64 long term διανύσματα χαρακτηριστικών και θέλουμε κάθε ένα από αυτά να τα ταξινομήσουμε ως προς τα είδη της μουσικής που περιέχουν. Ο τρόπος με τον οποίο το πετυχαίνουμε αυτό είναι, όπως και προηγουμένως στο διαχωρισμό μουσικής και όλων των άλλων ήχων στη ταινία, μέσο μίας *μηχανής διανυσμάτων υποστήριξης*.

Οι μηχανές διανυσμάτων υποστήριξης αποτελούν μοντέλα *επιβλεπόμενης μάθησης*(supervised learning) που είναι κατάλληλα, μεταξύ άλλων, για ταξινόμηση[99, 21]. Με τον όρο επιβλεπόμενη μάθηση δηλώνουμε ότι, δίνουμε στο μοντέλο δεδομένα(διανύσματα χαρακτηριστικών για παράδειγμα) τα οποία έχουμε χαρακτηρίσει ότι ανήκουν σε κάποια κατηγορία(ας θεωρήσουμε για αρχή ότι υπάρχουν μόνο δύο διαφορετικές κλάσεις, για παράδειγμα μουσική και μη μουσική) εκ των προτέρων και το μοντέλο μας με τη γνώση αυτή μπορεί να ταξινομήσει νέα δεδομένα σε μία από τις κατηγορίες. Συγκεκριμένα, ο τρόπος με τον οποίο γίνεται αυτό είναι μέσω της αποτύπωσης των δεδομένων σε ένα χώρο, πολλές φορές μεγαλύτερης διάστασης απ'ότι ο αρχικός χώρος των διανυσμάτων χαρακτηριστικών, στον οποίο όσα δεδομένα ανήκουν σε μία κατηγορία μπορούν να ξεχωρίσουν από τα υπόλοιπα μέσω μιας διαχωριστικής γραμμής. Έτσι τα νέα δεδομένα που βλέπει το σύστημα μας κατηγοριοποιούνται ανάλογα σε ποια μεριά του χώρου αναπαριστώνται. Τα SVMs λειτουργούν έχοντας ως στόχο να μεγιστοποιήσουν το κενό ανάμεσα στη διαχωριστική γραμμή που χωρίζει τις διαφορετικές κατηγορίες του χώρου.

Δεν θα μπούμε σε πολλές περαιτέρω λεπτομέρειες για να μην πλατειάσουμε στο θέμα, αλλά θα αναφέρουμε δύο πτυχές που μας αφορούν. Η μία αφορά στο γεγονός ότι εμείς στο σχόλιο παραπάνω αναφέραμε ότι ο χώρος χωρίζεται σε δύο μόνο κλάσεις, ενώ στο πλαίσιο της ταξινόμησης που θέλουμε να κάνουμε, χρειαζόμαστε 8 διαφορετικές κλάσεις. Αυτό είναι ένα συνηθισμένο πρόβλημα και αν και υπάρχουν διάφοροι μέθοδοι να λυθεί[25] εμείς θα χρησιμοποιήσουμε τη διαδικασία γνωστή ως *One Versus All-OVA*. Σύμφωνα με αυτή τη μέθοδο φτιάχνουμε δυαδικούς ταξινομητές για κάθε κλάση, οι οποίοι θα βρίσκουν την πιθανότητα να ανήκει ένα δεδομένο σε κάποια κλάση ή όχι. Με αυτό το τρόπο, καταλήγουμε να έχουμε 8 διαφορετικές πιθανότητες για κάθε δεδομένο, οι οποίες αν κανονικοποιηθούν, εκφράζουν το ποσοστό συμμετοχής κάθε κλάσης στο δεδομένο αυτό. Αυτό ακριβώς το διάνυσμα των 8 πιθανοτήτων είναι που θέλουμε ως τελικό αποτέλεσμα

Η άλλη πτυχή του προβλήματος μας είναι η ιδέα των *soft margins*[21]. Σε περίπτωση που δεν υπάρχει υπερεπίπεδο που να είναι δυνατός ο απόλυτα σωστός διαχωρισμός των δύο κλάσεων, μπορούμε να τροποποιήσουμε τον αλγόριθμο μας, ώστε να διαλέγουμε το διαχωρισμό που κάνει τη καλύτερη δυνατή διαμέριση του χώρου, δηλαδή να επιτρέπει μερικά κακώς χωρισμένα δείγματα αλλά να μεγιστοποιεί πάλι την απόσταση των σωστά χωρισμένων δειγμάτων. Σε αυτή τη περίπτωση εισάγονται κάποιες υπερπαράμετροι που ορίζουν το βαθμό ποινής του αλγορίθμου για τα λάθη ταξινόμησης που κάνει και τελικά έχουμε μία ισορροπία μεταξύ του μικρότερου δυνατού αριθμού λαθών και του μεγαλύτερου περιθωρίου μεταξύ των δύο κλάσεων. Στην υλοποίηση μας, τις τιμές αυτών των υπερπαραμέτρων τις βελτιστοποιεί εσωτερικά ο αλγόριθμος εκπαίδευσης του συστήματος, κρατώντας ένα μέρος του *δείγματος εκπαίδευσης*(training set) ξεχωριστά για αυτό το σκοπό(είναι το λεγόμενο *cross-validation set*).

Αναφέραμε συνοπτικά τη μέθοδο που θα ακολουθήσουμε. Ειδικά, η υλοποίηση της μεθόδου αυτής που χρησιμοποιούμε είναι της βιβλιοθήκης *mlpy*[4]. Το μόνο που δε σχολιάσαμε είναι τα δεδομένα που θα δώσουμε για την εκπαίδευση του μοντέλου ταξινόμησης των διαφορετικών ειδών μουσικής. Προς αυτή την κατεύθυνση, χρησιμοποιήθηκαν συνολικά ≈ 9 ώρες μουσικού περιεχομένου καταγεγραμμένες από το ραδιόφωνο και χαρακτηρισμένες ως μία από τις 8 αυτές κλάσεις από κάποιον ανθρώπινο παρατηρητή. Δηλαδή, το σύστημα μας συνολικά έπαιρνε κάθε μία ηχογράφηση, η οποία είχε μία ταμπέλα ως ένα από τα 8 είδη μουσικής, εξήγαγε τα μεσοπρόθεσμα χαρακτηριστικά και στη συνέχεια μετέτρεπε τον αντίστοιχο πίνακα σε ένα long-term averaged feature vector με μέγεθος 1×68, το οποίο και τροφοδοτούσε στο σύστημα SVM μαζί με το χαρακτηρισμό της μουσικής για την εκπαίδευση του. Μετά το πέρας αυτής της διαδικασίας, έχουμε ένα σύστημα το οποίο είναι σε θέση όταν του δίνουμε ένα 68-διάστατο διάνυσμα χαρακτηριστικών που αφορά ένα κομμάτι μουσικής της ταινίας, να μας δίνει στην έξοδο του ένα διάνυσμα 8 διαστάσεων που εκφράζει το ποσοστό συμμετοχής κάθε κλάσης μουσικής στο κομμάτι ήχου αυτό.

Επομένως συνολικά, εμείς δίνουμε το ηχητικό σήμα κάθε ταινίας, αρχικά στο SVM μοντέλο ταξινόμησης του ήχου σε μουσική ή όχι, παίρνουμε τα κομμάτια μουσικής που μας ενδιαφέρουν, εφαρμόζουμε την διαδικασία εξαγωγής χαρακτηριστικών σε κάθε ένα από αυτά και τελικά δίνουμε αυτά τα διανύσματα στο μοντέλο SVM που τα ταξινομεί ως προς το είδος της μουσικής. Άρα τελικά για κάθε ταινία έχουμε ένα πλήθος 8-διάστατων διανυσμάτων συμμετοχής κάθε είδους μουσικής στα ηχητικά κομμάτια μουσικής της ταινίας. Το τελικό διάνυσμα 1×8

αναπαράστασης της ταινίας ως μείγμα διαφορετικών ειδών μουσικής, προκύπτει ως η πιθανοτική εκτίμηση του αθροίσματος των επιμέρους μουσικών κομματιών της ταινίας.

Ταξινόμηση Ηχητικών Συμβάντων

Η διαδικασία που ακολουθούμε για την ταξινόμηση του ηχητικού σήματος της μουσικής σε κλάσεις ηχητικών συμβάντων, είναι αρκετά πιο εύκολη καθώς δεν υπάρχει το στάδιο διαχωρισμού μουσικής και όχι, που υπήρχε προηγουμένως, αλλά δίνουμε στο σύστημα μας σας είσοδο, ολόκληρο το σήμα μουσικής. Αρχικά, ορίζουμε και πάλι 8 διαφορετικές κλάσεις ηχητικών συμβάντων που απαντώνται συχνά σε ταινίες: Music, Fight, Screams, Shots, Speech, Others1, Others2, Others3. Οι πρώτες 5 κλάσεις είναι φανερό τι συμβάντα στη ταινία μοντελοποιούν. Οι ήχοι με κωδικό πρόθεμα Others, συμπεριλαμβάνουν θορύβους που ακούγονται στο περιβάλλον της ταινίας. Ως Others1 έχουμε κωδικοποιήσει ήχους χαμηλής ενέργειας και σταθερού σχεδόν σήματος, όπως για παράδειγμα σιωπή ή χαμηλότονο θόρυβο στο ακουστικό φόντο. Ως Others2 έχουν μοντελοποιηθεί γεγονότα του περιβάλλοντος που αφορούν έντονες αλλαγές στο σήμα του ήχου, όπως ένα κεραυνός, ή μία πόρτα που κλείνει δυνατά. Τέλος, η κλάση Others3 περιλαμβάνει τους υπόλοιπους δυνατούς ήχους, κυρίως μηχανήματα ή αυτοκίνητα. Στη συνέχεια εφαρμόζουμε σε ολόκληρο το ηχητικό σήμα τη διαδικασία εξαγωγής χαρακτηριστικών, το οποίο μας δίνει τελικώς την αναπαράσταση της ταινίας ως ένα long-term averaged διάνυσμα 1×64.

Ομοίως λοιπόν με προηγουμένως κατασκευάζουμε ένα μοντέλο SVM με τις ίδιες ακριβώς ιδιότητες και τις ίδιες διαστάσεις στην είσοδο-έξοδο, με μόνες διαφοροποιήσεις ότι στην έξοδο έχουμε διαφορετικές κλάσεις ταξινόμησης και διαφοερικά είναι και τα δεδομένα εκπαίδευσής του. Για την εκπαίδευση αυτού του μοντέλου χρησιμοποιήθηκαν ≈ 5000 δείγματα ήχων που είχαν χαρακτηριστεί ως προς το ηχητικό γεγονός από ανθρώπινο παρατηρητή, δηλαδή περίπου 800 δείγματα ανά κλάση. Αυτά τα δείγματα προέκυψαν από 30 διαφορετικές ταινίες που καλύπτουν όλο το φάσμα των των ταινιών(θρίλερ, δράμα, κωμωδίες κ.α.) και έχουν διάρκεια από 0.5 έως 10 δευτερόλεπτα το κάθε ένα. Μετά το πέρας της διαδικασίας εκπαίδευσης λοιπόν, έχουμε στα χέρια μας ένα μοντέλο που με είσοδο ολόκληρο το ηχητικό σήμα της ταινίας, μας δίνει στην έξοδο ένα 8-διάστατο διάνυσμα με τις τιμές συμμετοχής κάθε ηχητικού συμβάντος, από τα παραπάνω, στη ταινία αυτή. Δηλαδή, προκύπτει ως το κανονικοποιημένο ιστόγραμμα εμφάνισης των συμβάντων στη διάρκεια ολόκληρου του ηχητικού σήματος της ταινίας.

3.4 Πίνακας Ομοιότητας

Έχοντας πλέον την αναπαράσταση των ταινιών σε μορφή χαρακτηριστικών διανυσμάτων στο topic-space(αντίστοιχα στο tf-idf χώρο ή στις κλάσεις ήχου και ηχητικών συμβάντων που περιγράψαμε παραπάνω), μπορούμε να εφαρμόσουμε όποιο μέτρο ομοιότητας θεωρούμε ότι εκφράζει καλύτερα την "απόσταση" μεταξύ των ταινιών σε αυτό το χώρο. Γενικώς, τα μέτρα απόστασης(ομοιότητας) χαρτογραφούν την ομοιότητα μεταξύ κάποιας συμβολικής αναπαράστασης των δεδομένων(εδώ είναι η αναπαράσταση στον αντίστοιχο διανυσματικό χώρο) με βάση τόσο τις ιδιότητες κάθε στοιχείου, όσο και τις ιδιότητες του ίδιου του μέτρου. Στη περίπτωση μας, εφαρμόσαμε το μετασχηματισμό *cosine similarity* για την ομοιότητα των χαρακτηριστικών των ταινιών σε όλους τους χώρους αναπαράστασης.

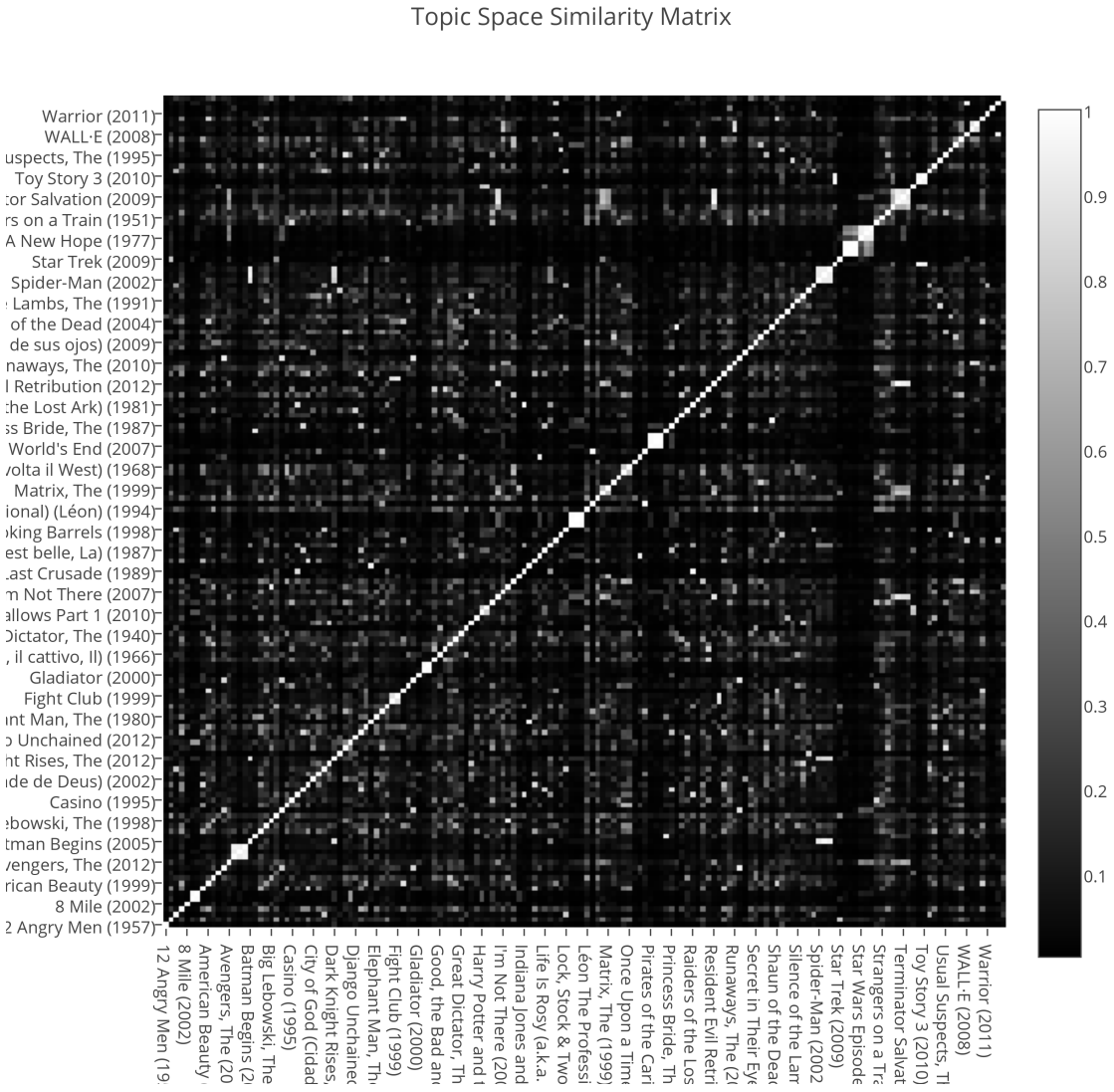
Στο μετασχηματισμό cosine similarity η ομοιότητα μεταξύ δύο στοιχείων ποσοτικοποιείται ως η γωνία θ μεταξύ των διανυσματικών αναπαραστάσεων στο χώρο χαρακτηριστικών που τα έχουμε αναλύσει[83]. Δηλαδή αν δύο ταινίες έχουν διανύσματα χαρακτηριστικών, σε όποιο χώρο βρισκόμαστε κάθε φορά, $\vec{m_a}, \vec{m_b}$ αντίστοιχα, τότε ο μετασχηματισμός cosine similarity αυτών είναι:

$$CosSim(\vec{m_a}, \vec{m_b}) = \theta = \frac{\vec{m_a}\vec{m_b}}{\|\vec{m_a}\|\|\vec{m_b}\|} = \frac{\sum_{i=1}^{N} m_a(i)m_b(i)}{\sqrt{\sum_{i=1}^{N} m_a(i)^2}\sqrt{\sum_{i=1}^{N} m_b(i)^2}} \tag{3.31}$$

, όπου με N συμβολίζεται η διάσταση του διανύσματος. Επειδή τα πιθανά διανύσματα που θα τροφοδοτήσουμε σε αυτό το μέτρο ομοιότητας αποτελούν είτε ποσοστά συμμετοχής κλάσεων-θεμάτων(για το topic,music,sound space) είτε συχνότητες λέξεων(για το tf-idf space), έχουμε μόνο θετικές τιμές στα κελιά των διανυσμάτων. Οπότε, ο μετασχηματισμός εκτείνεται από 0 εώς 1, με το 0 να εκφράζει την ορθογωνιότητα των διανυσμάτων, δηλαδή την μέγιστη δυνατή απόσταση τους και με 1 την παραλληλία τους, δηλαδή τη μέγιστη δυνατή ομοιότητα.

Είναι απλός, εύκολα κατανοητός και ταιριάζει στη μορφή των δεδομένων μας. Επίσης, είναι το πλέον χρησιμοποιούμενο μέτρο ομοιότητας για αναπαραστάσεις με την *tf-idf* μέθοδο και επιπροσθέτως εκφράζει σωστά και την απόσταση στο tag-space,όπως θα δούμε και στη συνέχεια. Έχοντας λοιπόν ένα μέτρο ομοιότητας για τα διανύσματα(ταινίες) μας, μπορούμε να βρούμε την ομοιότητα μεταξύ αυτών, υπολογίζοντας το cosine similarity κάθε ταινίας με κάθε άλλη.

Έτσι, παράγουμε λοιπόν το *πίνακα ομοιότητας* (similarity matrix) για τις ταινίες, στηριζόμενοι στα διανύσματα χαρακτηριστικών. Ο πίνακας αυτός εκφράζει την ομοιότητα κάθε ταινίας με κάθε άλλη ταινία στη βάση δεδομένων μας, ως μία συνεχή τιμή στο εύρος $0-1$, η οποία όσο μεγαλύτερη είναι τόσο περισσότερο ταιριάζει η ταινία. Για να το εκφράσουμε οπτικά αυτό μετασχηματίζουμε την συνεχή τιμή αυτή σε κλίμακα του γκρι, με το λευκό να αντιστοιχεί σε τιμές κοντά στο 1 και το μαύρο κοντά στο 0. Έτσι, προκύπτει ο πίνακας που φαίνεται ακολούθως(**Σχήμα 3.13**) και έχει παραχθεί από τις διανυσματικές αναπαραστάσεις των ταινιών στο topic-space .



Σχήμα 3.13: Topic Space Similarity Matrix για τη βάση των ταινιών

Όπως ήταν αναμενόμενο και παρατηρεί κανείς, η διαγώνιος είναι λευκή καθώς είναι η ομοιότητα κάθε ταινίας με τον εαυτό της. Επίσης, παρατηρεί κανείς και άλλες λεπτομέρειες. Για παράδειγμα, φαίνεται ένα "λευκό τετράγωνο" πάνω δεξιά στον πίνακα. Αυτό συμβαίνει γιατί είναι μαζεμένες σε εκείνο το σημείο οι ταινίες *Star Wars:I-VI* και η ομοιότητα που έχουν μεταξύ τους, προκαλεί αυτή τη συγκέντρωση λευκών κελιών. Αντίθετα, για τις ίδιες ταινίες παρατηρεί κανείς ότι τα υπόλοιπα κελιά είναι αρκετά σκουρόχρωμα, απόδειξη ότι δεν επιδεικνύουν μεγάλη συνάφεια, σύμφωνα με το μοντέλο μας, με τις υπόλοιπες ταινίες, πλην εξαιρέσεων.

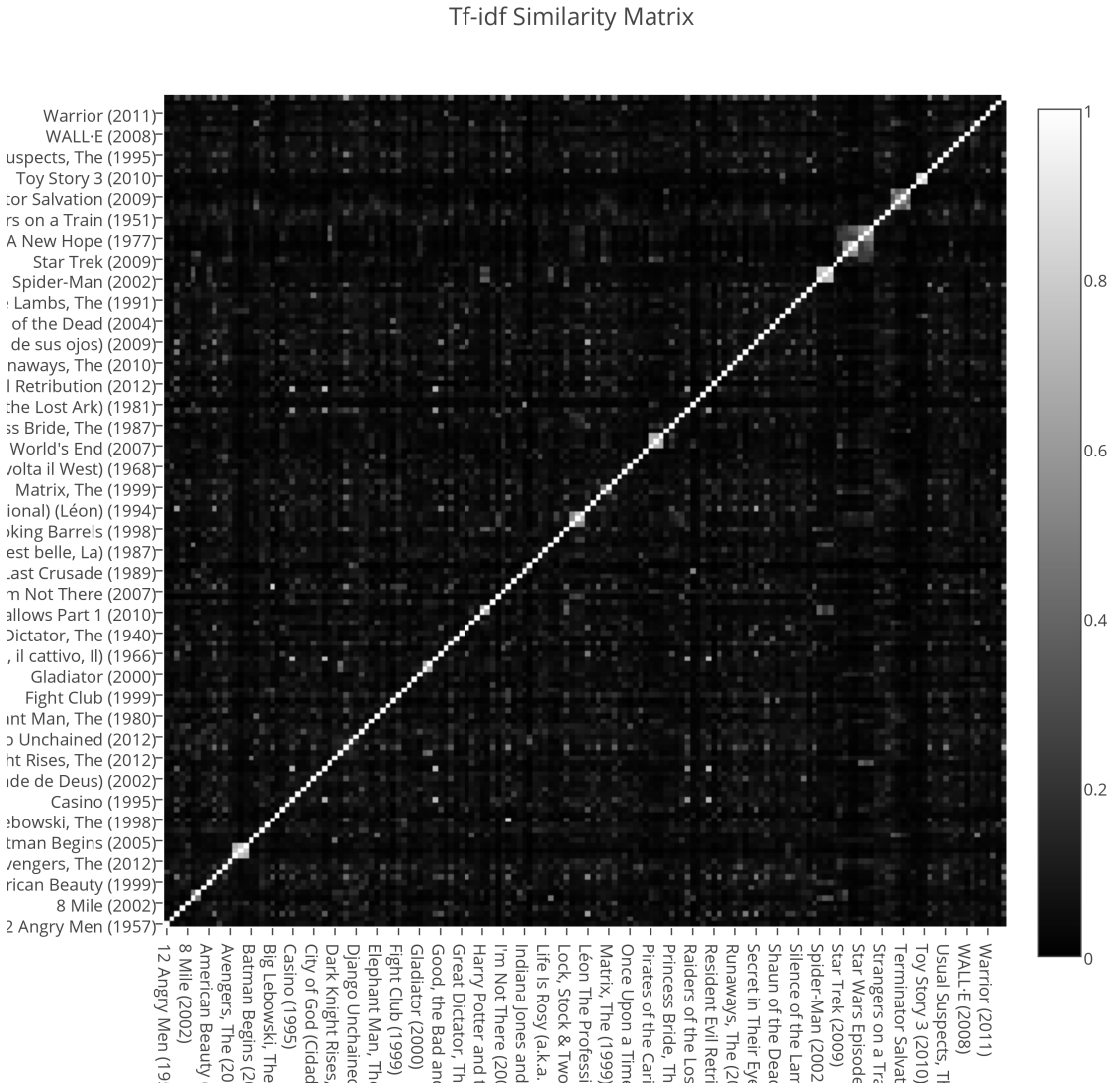
Στη συνέχεια και για λόγους πληρότητας, παραθέτουμε και το πίνακα ομοιότητας που προκύπτει από το *tf-idf* μοντέλο(**Σχήμα 3.14**). Παρατηρούμε ποιοτικές διαφορές στους δύο πίνακες, που θα συζητηθούν και στη συνέχεια, αφού παρουσιάσουμε και τον πίνακα ομοιότητας της βάσης αλήθειας, όπως θα αναφέρουμε στο επόμενο κεφάλαιο(**4.1.2**).

3.5 Σύντηξη Πληροφορίας

Σε αυτό το σημείο λοιπόν έχουμε στη διάθεση μας πίνακες ομοιότητας τόσο από την LDA μέθοδο, όσο και από τη μουσική και τα ηχητικά συμβάντα της ταινίας. Δηλαδή, έχουμε πληροφορία σχετικά με τη συνάφεια των ταινιών από τρεις διαφορετικές πηγές πληροφορίας, τα θέματα των υποτίτλων, το είδος της μουσικής της ταινίας και τα διαφορετικά ηχητικά γεγονότα που περιέχονται στη ταινία. Ο προβληματισμός που προκύπτει λοιπόν άμεσα είναι πως θα συνδυάσουμε τις πληροφορίες αυτές μεταξύ τους ώστε να κατασκευάσουμε ένα σύστημα εισηγήσεων που θα λαμβάνει υπ'όψιν του τις διαφορετικές πτυχές ομοιότητας από τα διάφορα κανάλια πληροφορίας.

Ο τρόπος συνδυασμού των καναλιών πληροφορίας απαντάται στη βιβλιογραφία ως *σύντηξη πληροφορίας*(data fusion)[103]. Αν και υπάρχουν πολλοί μέθοδοι και διάφορες προσεγγίσεις για το τρόπο που θα γίνει αυτό[5], όλες οι διαφορετικοί μέθοδοι χωρίζονται σε τρεις κυρίως κατηγορίες:

- Early-fusion, όπου έχουμε σύντηξη της πληροφορίας σε επίπεδο χαρακτηριστικών από τα διαφορετικά κανάλια πληροφορίας. Δηλαδή, αφού προχωρήσουμε στη διαδικασία εξαγωγής χαρακτηριστικών από κάθε διαφορετική πηγή πληροφορίας, συνενώνουμε, με διάφορους τρόπους, τα διανύσματα χαρακτηριστικών αυτά σε ένα και έχουμε μία μόνο



Σχήμα 3.14: Tf-idf Similarity Matrix για τη βάση των ταινιών

διαδικασία εκπαίδευσης, η οποία μας δίνει το τελικό σύστημα ταξινόμησης, παλινδρόμησης ή ότι άλλο θέλουμε[106, 91].

- Late-fusion, όπου έχουμε σύντηξη της πληροφορίας σε επίπεδο αποτελεσμάτων για κάθε κανάλι πληροφορίας. Δηλαδή, μετά την εξαγωγή των χαρακτηριστικών, εκπαιδεύουμε διαφορετικά συστήματα γνώσης για κάθε πηγή πληροφορίας και οι έξοδοι αυτών συνδυάζονται με διάφορους τρόπους, ώστε να προκύψει το τελικό αποτέλεσμα[6, 48].

- Hybrid-fusion, όπου προφανώς έχουμε ένα συνδυασμό των παραπάνω μεθόδων. Αν και λιγότερο χρησιμοποιούμενη διαδικασία, ενδείκνυται σε πολλές εφαρμογές καθώς μας επιτρέπει να κρατάμε τα θετικά στοιχεία και των δύο προηγούμενων μεθόδων[75, 7].

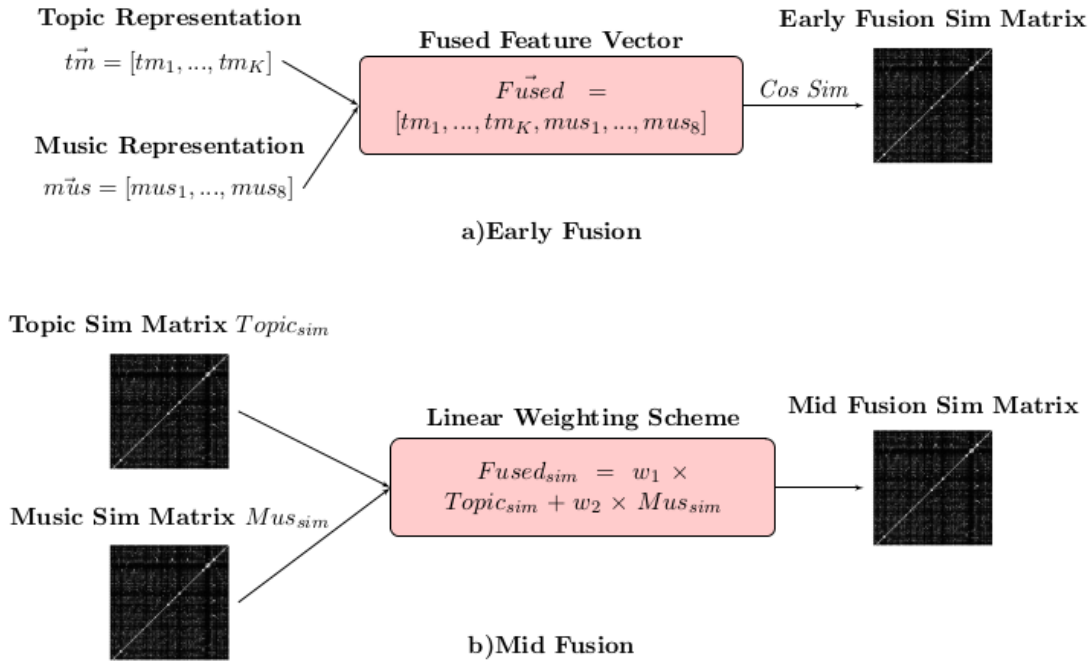
Στο πλαίσιο της εργασίας μας υλοποιήσαμε μονάχα δύο Late-Fusion τρόπους σύντηξης, καθώς η διερεύνηση της βέλτιστης μεθόδου δεν αποτελεί το βασικό στόχο αυτής. Και οι δύο τρόποι είναι αρκετά άμεσοι στην υλοποίηση τους. Ο πρώτος τρόπος συνίσταται αρχικά στη δημιουργία ενός επεκταμένου διανύσματος χαρακτηριστικών για κάθε ταινία, το οποίο θα περιλαμβάνει τα διανύσματα χαρακτηριστικών σειριακά τοποθετημένα για κάθε πηγή πληροφορίας που μας ενδιαφέρει. Δηλαδή, αν θέλαμε να ενώσουμε τη πληροφορία από το topic modeling σύστημα και την μουσική(music-classes) της ταινίας, τότε αν $\vec{tm} = [tm_1, tm_2, ..., tm_K]$ είναι η αναπαράσταση της ταινίας στο χώρο των K θεμάτων και $\vec{mus} = [mus_1, mus_2, ..., mus_8]$ η αναπαράσταση της ταινίας στο χώρο των κλάσεων 8 μουσικής που ορίσαμε προηγουμένως, τότε η ενωμένη πληροφορία θα ήταν ως το υπερδιάνυσμα χαρακτηριστικών: $\vec{Fused} = [\vec{tm}, \vec{mus}] = [tm_1, tm_2, ..., tm_K, mus_1, mus_2, ..., mus_8]$. Έχοντας λοιπόν αυτό το επεκταμένο διάνυσμα χαρακτηριστικών για κάθε ταινία, εφαρμόζουμε κατά τα γνωστά το μετασχηματισμό cosine similarity μεταξύ των ταινιών και μας προκύπτει ένας πίνακας ομοιότητας.

Αυτός ο πίνακας ομοιότητας τώρα έχει προκύψει από τη συνύπαρξη, με την ίδια βαρύτητα, των αναπαραστάσεων της ταινίας στους διάφορους χώρους πληροφορίας, στο παράδειγμα μας των κλάσεων της μουσικής και των θεμάτων. Βέβαια, ο τρόπος αυτός επεκτείνεται και σε οποιοδήποτε άλλο συνδυασμό χώρων, όπως κλάσεις μουσικής και κλάσεις ηχητικών γεγονότων ή στο τριπλό συνδυασμό και του θεματικού χώρου, από το κείμενο των ταινιών, και των χώρων κλάσεων μουσικής και θορύβων, από το ηχητικό σκέλος της ταινίας. Επειδή, έχουμε συνδυασμό των αναπαραστάσεων των ταινιών, οι οποίες είναι σε μορφή διανυσμάτων θα αναφερόμαστε σε αυτό το τρόπο εφ'εξής ως early fusion. Αλλά δε θα πρέπει να συγχέεται με το παραπάνω ορισμό της πραγματικής early-fusion, καθώς εκεί όπως τονίσαμε έχουμε μόνο μία διαδικασία εκπαίδευσης ενώ στη δική μας υλοποίηση, που είναι late-fusion κάθε κανάλι πληροφορίας έχει ξεχωριστή διαδικασία εκπαίδευσης για να μας δώσει τη τελική αναπαράσταση της ταινίας στον αντίστοιχο χώρο.

Ο δεύτερος τρόπος είναι και αυτός σε επίπεδο αποτελεσμάτων κάθε καναλιού πληροφορίας και πιο συγκεκριμένα ασχολούμαστε κατ'ευθείαν με το πίνακα ομοιότητας κάθε διαφορετικής αναπαράστασης. Δηλαδή, αν θέλουμε να συνδυάσουμε πάλι το topic μοντέλο και το μοντέλο αναπαράστασης σε κλάσεις μουσικής των ταινιών, τότε αν ο πίνακας ομοιότητας του topic μοντέλου είναι $Topic_{sim}$ και ο πίνακας ομοιότητας από τη συσχέτιση μεταξύ των κλάσεων μουσικής είναι $Music_{sim}$, τότε ο τελικός πίνακας ομοιότητας προκύπτει ως γραμμικός συν-

δυασμός αυτών: $Fused_{sim} = w_1 \times Topic_{sim} + w_2 \times Music_{sim}$, όπου $\vec{w} = [w_1, w_2]$, είναι ένα διάνυσμα βαρών που επιλέγουμε εμείς και τα στοιχεία του αθροίζουν στη μονάδα. Αντίστοιχα, αν θέλουμε να ενώσουμε τη πληροφορία από 3 πίνακες ομοιότητας, προσθέτοντας στο προηγούμενο συνδυασμό και τον πίνακα ομοιότητας από την αναπαράσταση στο χώρο με τα ηχητικά συμβάντα, τότε το διάνυσμα βαρών θα είχε ένα ακόμα στοιχείο κ.ο.κ.

Τη συγκεκριμένη μέθοδο θα την ονομάζουμε από εδώ και πέρα ως mid-fusion, καθώς ενώ είναι late-fusion διαδικασία είμαστε μετά το στάδιο της συνένωσης των διανυσμάτων, όπως περιγράφηκε στη προηγούμενη early-fusion μέθοδο, αλλά υπάρχουν και πιο late στάδια σύντηξης πληροφορίας(όπως για παράδειγμα στη φάση του information retrieval), εξ'ού και η ονομασία mid-fusion. Για να βρούμε τους βέλτιστους συνδυασμούς στα διάφορα mid-fusion σχήματα, δοκιμάσαμε εξαντλητικά όλες τις τιμές με απλή *αναζήτηση πλέγματος*(grid search) στο χώρο των βαρών και κρατήσαμε αυτούς που μας έφερναν τα καλύτερα τελικά αποτελέσματα, όπως αυτά αναλύονται στο επόμενο κεφάλαιο(**4**) και αναγράφονται στο πίνακα αποτελεσμάτων(**4.1**).



Σχήμα 3.15: Early και Mid fusion διαγράμματα. Στο πάνω σχήμα **(a)**, φαίνεται η πορεία που ακολουθούμε στην early-fusion, με τη δημιουργία του επεκταμένου διανύσματος και τη μετέπειτα παραγωγή του πίνακα ομοιότητας. Στο κάτω σχήμα **(b)** φαίνεται η διαδικασία γραμμικού συνδυασμού των πινάκων ομοιότητας για τη παραγωγή του τελικού πίνακα.

Τέλος, παρατίθενται τα σχεδιαγράμματα των δύο μεθόδων στο **Σχήμα 3.15** για οπτικοποίηση των δύο μεθόδων και καλύτερη κατανόηση τους.

Κεφάλαιο 4

Ανάκτηση Πληροφορίας και Πειραματική Αποτίμηση

Σε αυτό το κεφάλαιο θα περιγράψουμε τη διαδικασία που ακολουθήσαμε στο πλαίσιο της διπλωματικής για την υλοποίηση του εισηγητικού συστήματος για ταινίες. Αρχικά, θα αναφερθούμε στα δεδομένα που αποτελούν τη βάση της εργασίας μας καθώς και τον τρόπο παραγωγής της βάσης αλήθειας για το χαρακτηρισμό των τελικών αποτελεσμάτων αποτελεσμάτων μας(**4.1**). Δευτερευόντως, θα αναφέρουμε σύντομα σε μερικές τεχνικές λεπτομέρειες σχετικά με την υλοποίηση της μεθόδου(**4.2**). Ακολούθως, θα αναλύσουμε τον τρόπο με τον οποίο γίνεται η ανάκτηση πληροφορίας για συστάσεις σχετικά με τις ταινίες **(4.3)**. Στη συνέχεια, θα περιγράψουμε τις προσεγγίσεις που είχαμε ως προς ποια μέτρα απόδοσης είναι κατάλληλα για να περιγράψουν την ποιότητα των εισηγήσεων μας **(4.4)**. Τέλος, θα παρουσιάσουμε τα αποτελέσματα της μεθόδου μας και κάποια συμπεράσματα **(4.5)**, τα οποία θα επεκταθούν στο επόμενο κεφάλαιο **(5)**.

4.1 Περιγραφή Δεδομένων και Βάση Αλήθειας

Στη συγκεκριμένη ενότητα θα περιγράψουμε πρώτα το σετ δεδομένων που χρησιμοποιήσαμε(**4.1.1**) και έπειτα θα αναφερθούμε στη *βάση αλήθειας*(ground truth)(**4.1.2**) που δημιουργήθηκε ως μέτρο σύγκρισης για τις εισηγήσεις μας.

4.1.1 Περιγραφή Δεδομένων

Ξεκινώντας, θα αναφερθούμε στα δεδομένα και στο τρόπο απόκτησης τους. Επιλέξαμε μία συλλογή 160 ταινιών, ως τη βάση δεδομένων για το εισηγητικό μας σύστημα. Επειδή θέλουμε οι ταινίες να είναι γνωστές στο μέσο χρήστη, επιλέξαμε το μεγαλύτερο πλήθος από αυτές από τη λίστα των *Top 250 Movies* [1] από τη δημοφιλή ιστοσελίδα ταινιών *IMDb*. Η συγκεκριμένη λίστα έχει συνταχθεί με βάση τις προτιμήσεις των χρηστών, όπως αυτοί τις έχουν εκφράσει στην ιστοσελίδα αυτή, και η σειρά των ταινιών υπολογίζεται με βάση μία συγκεκριμένη φόρμουλα, ώστε να αποτελεί μία αληθινή εκτίμηση *Bayes* (*true 'Bayesian estimator'*), της σειράς προτίμησης των χρηστών. Επίσης, επιλέχθηκαν ειδικά κάποιες ταινίες από αυτές, όπως *σίκουελ*(sequel) άλλων ταινιών ή κάποια ειδικής κατηγορίας (π.χ. *γουέστερν*-western), έτσι ώστε οι περισσότερες από τις ταινίες να έχουν τουλάχιστον μία σχετική ταινία στη βάση, η βάση ταινιών να είναι αρκετά πλήρης όσον αφορά στη θεματολογία και στο είδος των ταινιών (ώστε να μην υπάρχει *προκατάληψη*(bias) ως προς τα μεταδεδομένα, είδος, ηθοποιούς κτλ.), αλλά και να είναι και ποιοτικά ελέγξιμα τα αποτελέσματα. Τα αρχεία των ταινιών είναι σε *μορφή*(format) μία εκ των { *.avi, .mkv, .mp4* }.

Έτσι παραδειγματικά θα αναφέρουμε μερικές ταινίες για διάφορα από τα πιο ευρέως διαδεδομένα είδη:

- ⋆ **Action** : *The Avengers (2012), Die Hard (1988), Elite Squad The Enemy Within (2010), Indiana Jones and the Last Crusade (1989)*
- ⋆ **Animation** : *Finding Nemo (2003), Grave of the Fireflies (Hotaru no haka) (1988), The Lion King (1994), Princess Mononoke (Mononoke-hime) (1997)*
- ⋆ **Comedy** : *The Great Dictator (1940), Groundhog Day (1993), Intouchables (2011), The Big Lebowski (1998)*
- ⋆ **Drama** : *The 400 Blows (Les quatre cents coups) (1959), Forrest Gump (1994), Good Will Hunting (1997), The Green Mile(1999)*
- ⋆ **Music** : *24 Hour Party People (2002), 8 Mile (2002), Ray (2004), Walk the Line (2005)*
- ⋆ **Mystery** : *Memento (2000), The Prestige (2006) ,The Secret in Their Eyes (2009) ,Strangers on a Train (1951)*

[1] http://www.imdb.com/chart/top

- ⋆ **Romance** : *American Beauty (1999), Before Sunrise (1995), Casablanca (1942), In the Mood For Love (Fa yeung nin wa) (2000)*
- ⋆ **Sci-fi** : *Alien (1979), Back to the Future (1985), Blade Runner (1982), Star Trek (2009), Star Wars Episode I - The Phantom Menace (1999)*
- ⋆ **Thriller** : *Seven (a.k.a. Se7en) (1995), Shutter Island (2010), The Silence of the Lambs (1991), The Sixth Sense(1999)*
- ⋆ **War** : *Apocalypse Now (1979), Boot, Das (Boat, The) (1981), Inglourious Basterds (2009), Saving Private Ryan (1998)*
- ⋆ **Western** : *For a Few Dollars More (1965), The Good, the Bad and the Ugly (1966), Once Upon a Time in the West (C'era una volta il West) (1968)*

Στη συνέχεια, βρήκαμε τους αγγλικούς υποτίτλους για τις συγκεκριμένες ταινίες και βεβαιωθήκαμε ότι ήταν σωστοί και συγχρονισμένοι. Τους προμηθευτήκαμε από την μεγαλύτερη ανοιχτή βάση υποτίτλων[2] όπου οι χρήστες αναρτούν ελεύθερα τον υποτιτλισμό κάθε ταινίας. Οι υπότιτλοι είναι σε μορφή(*format*) .srt .

Ακολούθως, τους επεξεργαστήκαμε, όπως περιγράφηκε στην **Επεξεργασία Κειμένου (3.2.2)**, με αποτέλεσμα κάθε ταινία να περιγράφεται πλέον ως μία λίστα από ζεύγη. Κάθε ζεύγος, αποτελείται από ένα δείκτη, που μας υποδεικνύει σε ποια λέξη του λεξικού αντιστοιχεί αυτή η καταχώρηση και ένα μετρητή, ο οποίος αναφέρει πόσες φορές εμφανίζεται η λέξη αυτή στους υπότιτλους της ταινίας. Σε αυτή τη μορφή θα δοθεί το *σώμα*(corpus) των δεδομένων μας για την υλοποίηση της μεθόδου **Latent Dirichlet Allocation(LDA)** .

Σε αυτό το σημείο και ξεχωριστά από την εφαρμογή της μεθόδου *LDA* , το *corpus* των υποτίτλων μεταβάλλεται με βάση το μετασχηματισμό *term frequency-inverse document frequency (tf-idf)* [84, 63] όπως αυτός περιγράφηκε στη παράγραφο **(3.2.2)**. Στη συνέχεια θα χρησιμοποιήσουμε αυτή τη διανυσματική αναπαράσταση για να υπολογίσουμε την ομοιότητα μεταξύ των υποτίτλων κάθε ταινία άμεσα, χωρίς τη μεταφορά μας στο *χώρο θεμάτων*(topic-space) δηλαδή, υπολογίζοντας την απόσταση μεταξύ των διανυσμάτων. Αυτό το κάνουμε για να έχουμε και μία ακόμα baseline μέθοδο ως μέτρο σύγκρισης για τη χρησιμότητα του topic μοντέλου.

[2]http://www.opensubtitles.org/en/search

4.1.2 Παραγωγή Βάσης Αλήθειας

Για να αξιολογήσουμε την ποιότητα των εισηγήσεών μας, πέρα από τα κατάλληλα μέτρα απόδοσης που θα περιγραφούν παρακάτω, χρειαζόμαστε και ένα σημείο αναφοράς, τη *βάση αλήθειας*(ground truth). Στη συγκεκριμένη εφαρμογή, η βάση αλήθειας θα μας υποδεικνύει "πόσο ταιριάζει" μία ταινία με κάποια άλλη στη βάση δεδομένων μας και επομένως για κάθε ταινία ποιες είναι οι καλύτερες συστάσεις που μπορούν να γίνουν σε κάποιο χρήστη. Για να γίνει αυτό, θα πρέπει πρώτα να ασχοληθούμε με τη παραγωγή της βάσης αλήθειας (*ground truth generation*).

Καθώς δεν υπάρχει κάποιος επίσημος ορισμός του "πόσο ταιριάζει" μία ταινία με κάποια άλλη, ο μόνος τρόπος που μπορούσε να γίνει αυτό είναι λαμβάνοντας υπ'όψη τις γνώμες ατόμων σχετικά με κάθε ταινία. Για να είναι όσο το δυνατόν πιο αντικειμενική και καθολικά αποδοχούμενη η βάση αλήθειας, απαιτείται μεγάλος αριθμός ατόμων που θα εκφράσουν τις προτιμήσεις τους για κάθε ταινία και κυρίως, τη συσχέτιση μεταξύ αυτών. Για το σκοπό αυτό, χρησιμοποιήσαμε το *σετ δεδομένων Tag Genome*(Tag Genome data set) το οποίο βασίζεται στην ιστοσελίδα MovieLens[3]. Το MovieLens είναι μία ιστοσελίδα συστάσεων για ταινίες η οποία επιτρέπει στους χρήστες να θέτουν *ετικέτες* (tags) σε κάθε ταινία. Το Tag Genome data set είναι μία συλλογή δεδομένων, η οποία περιγράφει κάθε ταινία με τη βοήθεια μιας ομάδας *ετικετών*(tags). Για κάθε μία από ένα πλήθος 9734 ταινιών, υπάρχει μία συσχέτιση κάθε ταινίας με κάθε ένα από τα 1128 tags υπό τη μορφή μίας συνεχούς τιμής από 0 εώς 1. Δηλαδή, κατά βάση είναι μία διανυσματική αναπαράσταση κάθε ταινίας σε ένα 1128-διάστατο χώρο από tags (εφ'εξής tag-space). Όσο πιο μεγάλη συνάφεια έχει μία ταινία με ένα tag, τόσο πιο μεγάλη τιμή έχει στην αντίστοιχη κατεύθυνση στο χώρο αυτό. Αυτές οι συσχετίσεις υπολογίστηκαν από τις προτιμήσεις των χρηστών, όπως αυτές εκφράζονταν με τη πάροδο των χρόνων σε reviews, σχόλια ή απ'ευθείας tagging των ταινιών. Στη πραγματικότητα βέβαια, υπήρξε μεγάλη επεξεργασία αυτών των δεδομένων για να έρθουν στην τελική τους μορφή[100], αλλά αυτή η ανάλυση δεν ανήκει στην παρούσα εργασία.

Συγκεκριμένα, μας παρέχονται 3 αρχεία:

- Ένα αρχείο που περιέχει τις σχέσεις tag↔tag id **(1)**
- Ένα αρχείο που περιέχει τις σχέσεις movie title↔movie id **(2)**

[3]http://www.movielens.org

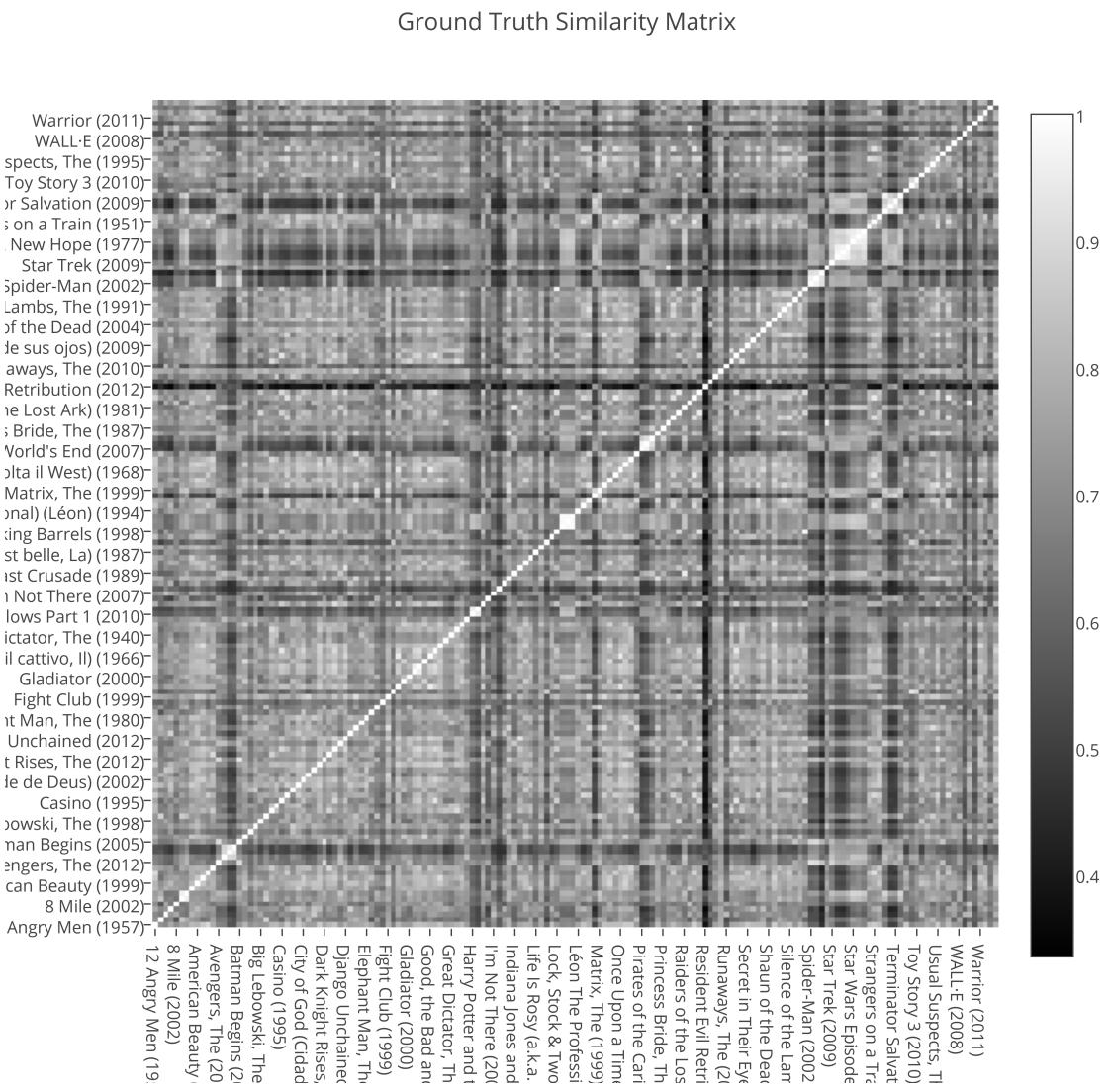
- Ένα αρχείο που περιέχει τις σχέσεις movie id↔tag id↔id relevance **(3)**

Για να τα συνδυάσουμε έπρεπε να κάνουμε την εξής διαδικασία: Να αντιστοιχίσουμε τους τίτλους των ταινιών της βάσης μας, με τους αντίστοιχους δείκτες(id) από το αρχείο **(2)**. Αυτό το κάναμε αυτοματοποιημένα, συγκρίνοντας την *απόσταση Levenshtein*(Levenshtein distance) κάθε τίτλου της βάσης μας με κάθε τίτλο από το tag genome datasetκαι κρατώντας την ελάχιστη. Η *Levenshtein distance* είναι μια από τα πιο διαδεδομένα μετρικά απόστασης μεταξύ συμβολοσειρών. Περιληπτικά, υπολογίζει την απόσταση μεταξύ δύο συμβολοσειρών με βάση τον ελάχιστο αριθμό αλλαγών στους χαρακτήρες της μία συμβολοσειράς(εισαγωγές, αντικαταστάσεις, αφαιρέσεις και μεταθέσεις), ώστε να είναι ίδια με την άλλη.[54]. Εδώ ο τίτλος με την ελάχιστη απόσταση από τον δικό μας τίτλο, είναι προφανώς ο ίδιος τίτλος της ταινίας, οπότε έχουμε βρει το ζητούμενο id για αυτή τη ταινία.(Αυτό προαπαιτεί οι τίτλοι των ταινιών να ακολουθούν ένα συγκεκριμένο μοτίβο, πράγμα που φροντίσαμε να ισχύει) Στη συνέχεια από το αρχείο **(3)**, έχοντας το δείκτη της ταινίας, βρίσκουμε τη συνάφεια της με όλα τα tags και μπορούμε να δημιουργήσουμε τη διανυσματική αναπαράσταση της ταινίας στο tag space χώρο.

Κατ'ουσίαν, αυτά τα tags αποτελούν περιγραφείς των ταινιών. Είναι διάφορες λέξεις: ουσιαστικά (*guns, airplane, wine*) , επίθετα (*awesome, bad, dark*), χρονολογίες (*9/11, 1980*), ονόματα και τόποι (*iran, jesus, superman*) κ.α. τα οποία οι χρήστες πιστεύουν ότι αποτελούν χαρακτηριστικά των ταινιών. Είναι σημαντικό να προσέξει κανείς ότι υπάρχουν και tags που αποτελούν μεταδεδομένα για τις ταινίες. Για παράδειγμα, υπάρχει το tag *tarantino* , το οποίο έχει πολύ υψηλές τιμές σε ταινίες που έχει σκηνοθετήσει ο *Quentin Tarantino*. Στη δικιά μας βάση δεδομένων είναι μεταξύ των άλλων και οι ταινίες του, *Reservoir Dogs(1992)* και *Django Unchained(2012)*, οι οποίες δεν έχουν μεγάλη συνάφεια όσον αφορά το σενάριο, δηλαδή το κομμάτι των διαλόγων που το σύστημα μας έχει ως κύρια είσοδο, αλλά στη βάση αλήθειας θα εμφανίσουν να έχουν υψηλή συσχέτιση μεταξύ τους λόγω της συγκεκριμένης ετικέτας. Αυτό είναι μονάχα ένα παράδειγμα του πλουραλισμού σε ετικέτες υψηλού γνωσιακού επιπέδου που δεν είναι δυνατό να ενσωματώσει με κανένα τρόπο το σύστημα μας. Παρόμοιας μορφής είναι και άλλες ετικέτες, όπως *oscar (best picture), adapted from game* . Οι δυνατότητες λοιπόν που έχουν οι χρήστες για να εκφράσουν τις συσχετίσεις μεταξύ των ταινιών είναι πολύ μεγαλύτερες, ή έστω πολύ διαφορετικές, από αυτές που εκφράζει το σύστημα μας. Αυτό είναι πολύ σημαντικό καθώς εξηγεί διάφορες ποιοτικές διαφορές μεταξύ των συσχετίσεων που θα παραχθούν από το σύστημα μας, σε σχέση με τη βάση αλήθειας, όπως θα δούμε και στη

συνέχεια.

Έχοντας λοιπόν την αναπαράσταση των ταινιών σε διανυσματική μορφή στο tag space, μπορούμε να εφαρμόσουμε το μετασχηματισμό *cosine similarity*, όπως και προηγουμένως. Παράγουμε λοιπόν τον *πίνακα ομοιότητας*(similarity matrix) για τις ταινίες, στηριζόμενοι στη βάση αλήθειας, όπως και με το topic space. Έτσι, προκύπτει ο πίνακας που φαίνεται ακολούθως(**Σχήμα 4.1**) και έχει παραχθεί από τις διανυσματικές αναπαραστάσεις των ταινιών στο tag space, δηλαδή είναι σύμφωνος με τη βάση αλήθειας. Αν γυρίσει κανείς πίσω στο πίνακα ομοιότητας για το topic space μοντέλο(**Σχήμα 3.13**) θα παρατηρήσει διαφορές, τις οποίες θα αναλύσουμε στη συνέχεια(**4.5**).



Σχήμα 4.1: Ground Truth Similarity Matrix για τη βάση των ταινιών

4.2 Θέματα Υλοποίησης

Να σημειωθεί εδώ ότι η υλοποίηση τόσο της *tf-idf* μεθόδου, όσο και διάφορες υλοποιήσεις για την *LDA* ανήκουν στη βιβλιοθήκη Gensim [79]. Η βιβλιοθήκη αυτή παρέχει είτε ολοκληρωμένες υλοποιήσεις διαφόρων μεθόδων επεξεργασίας κειμένου είτε κατάλληλα *περιτυλίγματα κώδικα*(wrappers) προς άλλες έτοιμες υλοποιήσεις. Στην περίπτωσή μας θα χρησιμοποιήσουμε το *wrapper* για την *LDA* μεθόδο, όπως υλοποιήθηκε από τον *Andrew McCallum*. Η βιβλιοθήκη ονομάζεται *MAchine Learning for LanguagE Toolkit - MALLET* [64] και αποτελεί μία πολύ γρήγορη και αποτελεσματική εφαρμογή της *δειγματοληψίας Gibbs* (Gibbs sampling) για *Bayesian inference* , όπως αυτή περιγράφηκε συνοπτικά στη παράγραφο **(3.2.3)**. Να αναφερθεί εδώ ότι δοκιμάσαμε και τη γηγενή υλοποίηση του Gensim για την *LDA* μέθοδο [43], η οποία βασίζεται στη *μεταβολή συμπερασματολογία*(variational inference). Στην ίδια λογική είναι βασισμένη και η υλοποίηση μέσω της βιβλιοθήκης *Vowpal Wabbit* [3], για την οποία παρέχεται *wrapper* από το Gensim πακέτο. Παρ΄όλα ταύτα και οι δύο υλοποιήσεις δεν μας επέστρεφαν καλά αποτελέσματα. Ένα λόγος ίσως είναι ότι αυτές μέθοδοι βασίζονται στο λογισμό μεταβολών για σύγκλιση σε συγκεκριμένο μορφή των *θεμάτων*(topics), και όταν έχουμε σχετικά λίγα δεδομένα η μέθοδοι δεν είναι τόσο εύρωστοι (robust). Τέλος, δοκιμάσαμε και τη προσέγγιση με την ιεραρχική δόμηση των θεμάτων, η οποία βασίζεται στον αλγόριθμο *Hierarchical Dirichlet Process* [102], η οποία, ομοίως με τις άλλες προσεγγίσεις, δεν απόδωσε θέματα με νόημα και συνεκτικότητα.

4.3 Ανάκτηση Πληροφορίας

Στη συγκεκριμένη ενότητα θα αναφερθούμε στη διαδικασία *ανάκτησης πληροφορίας* (information retrieval) για συστάσεις ταινιών από το σύστημα μας. Ως ανάκτηση πληροφορίας νοείται η διαδικασία κατά την οποία απαιτείται κάποιου είδους πληροφορία, σε σχέση με μία συλλογή πηγών πληροφορίας. Στη περίπτωση του εισηγητικού συστήματος που προτείνουμε, ως ανάκτηση πληροφορίας νοείται η ανάσυρση από τη συλλογή μας, σχετικών ταινιών ως προς μία συγκεκριμένη ταινία, που αποτελεί και το αντικείμενο αναζήτησης. Δηλαδή, έχοντας αναπαραστήσει την συνάφεια των ταινιών σε κάποιο γνωσιακό χώρο μέσω του αντίστοιχου πίνακα ομοιότητας, ζητείται να βρεθούν οι σχετικές ταινίες για κάθε ταινία στη βάση μας. Με γνώμονα λοιπόν αυτή την απαίτηση πληροφορίας, πρέπει να αντιμετωπίσουμε ένα πλήθος ζητήματα όπως πως ορίζεται το πόσο μοιάζει μία ταινία με μία άλλη, πόσες ταινίες πρέπει να προτείνουμε για κάθε ταινία και διάφορα άλλα σημαντικά θέματα. Με αυτά θα ασχοληθούμε

στη παρούσα ενότητα καθώς και σε ένα βαθμό και στην (**4.4**).

Έχοντας το πίνακα ομοιότητας λοιπόν, όπως αυτός προέκυψε από κάποιο μοντέλο(LDA, tf-idf, music, mid-fusion κ.ο.κ), μπορούμε να βρούμε για κάθε ταινία με ποιες ταιριάζει περισσότερο, σύμφωνα με τη μοντελοποίηση που της έχουμε κάνει. Συγκεκριμένα, αρκεί να πάμε στη γραμμή ή στήλη του πίνακα που αντιστοιχεί στην ταινία αυτή και να δούμε τις τιμές συνάφειας που έχει με όλες τις υπόλοιπες. Από εκεί και πέρα το θέμα είναι πως θα διαλέξει κανείς το κατάλληλο τρόπο για το ποιες εισηγήσεις είναι ορθές να γίνουν.

Το πρώτο βήμα προφανώς είναι για κάθε ταινία, να ταξινομήσουμε την αντίστοιχη γραμμή, ή στήλη, του πίνακα ομοιότητας, με τις μεγαλύτερες τιμές, εξαιρουμένης της τιμής της διαγωνίου, να μας ενδιαφέρουν περισσότερο. Δηλαδή, για κάθε ταινία έχουμε στη διάθεσή μας ένα διάνυσμα με 159 τιμές συνάφειας, χωρίς αυτή της διαγωνίου, ως προς τις υπόλοιπες ταινίες της βάσης δεδομένων μας. Ο προβληματισμός τώρα είναι στο *ερώτημα*(query):"Για την x ταινία, ποιες ταινίες μου προτείνεις;", με ποιο τρόπο θα απαντήσει κανείς έχοντας το παραπάνω διάνυσμα. Μία προσέγγιση είναι για κάθε x ταινία που μπορεί να ερωτηθεί, να επιστρέφει τις N πιο συναφείς ταινίες με βάση το παραπάνω διάνυσμα, όπου N είναι ένας προκαθορισμένος αριθμός. Δηλαδή, θα βρίσκει ποια κελιά έχουν τις N πρώτες μεγαλύτερες τιμές στο προαναφερθέν διάνυσμα και θα επιστρέφει τους αντίστοιχους τίτλους ταινιών, με βάση τους δείκτες των κελιών αυτών. Η συγκεκριμένη προσέγγιση όμως εμφανίζει αδυναμίες στις εξής τρεις περιπτώσεις:

- Η πρώτη περίπτωση συμβαίνει όταν οι τιμές συνάφειας τους διανύσματος ομοιότητας της ζητούμενης ταινίας είναι πολύ μικρές για όλες τις υπόλοιπες ταινίες. Αυτό μπορεί να συμβαίνει συνήθως γιατί το σύστημα μας δεν έχει μοντελοποιήσει καλά την συγκεκριμένη ταινία, οπότε αδυνατεί να βρει παρόμοιες με αυτή ή γιατί πράγματι δεν υπάρχουν σχετικές ταινίες στη βάση των ταινιών μας. Αυτό έχει σαν αποτέλεσμα οι πρώτες N ταινίες που θα επιστρέψει αυτό το σύστημα *ανάκτησης πληροφορίας*(information retrieval), να μην είναι κατά πάσα πιθανότητα "καλές" εισηγήσεις για τη συγκεκριμένη ταινία. Ένα πιο ευέλικτο σύστημα θα μπορούσε να αντιληφθεί ότι η συγκεκριμένη ταινία δεν έχει συναφείς ταινίες, είτε πράγματι αυτές δεν υπάρχουν στη βάση δεδομένων είτε επειδή έχει γίνει λάθος μοντελοποίηση(πράγμα που το σύστημα δεν γνωρίζει βέβαια), οπότε και να αποφεύγει να κάνει λανθασμένες εισηγήσεις.

- Η άλλη περίπτωση είναι η συνέχεια της προηγούμενης και αφορά σε περιπτώσεις στις

οποίες οι τιμές συνάφειας του διανύσματος που είναι αρκετά υψηλές ώστε να δικαιολογούν ομοιότητα μεταξύ των ταινιών είναι λιγότερες από N. Αυτό θα οδηγούσε σε μερικές σωστές εισηγήσεις και σε μερικές λανθασμένες πιθανότατα. Αντίθετα, ένα σύστημα με πιο δυναμικό τρόπο επιλογής εισηγήσεων θα μπορούσε να το αποφύγει αυτό, επιλέγοντας εν γνώσει του να επιστρέψει λιγότερες από N εισηγήσεις, για τις οποίες όμως θα ήταν αρκετά πιο βέβαιο ότι είναι πράγματι συναφείς ταινίες.

- Τέλος, μία διαφορετική περίπτωση από τις άλλες δύο είναι η ακόλουθη. Σε περίπτωση που μία ταινία έχει παραπάνω από N σχετικές ταινίες στη βάση δεδομένων μας και έστω ότι το σύστημα μας έχει μοντελοποιήσει σωστά αυτές τις ταινίες, οπότε οι τιμές συνάφειας μεταξύ αυτών πράγματι καταδεικνύουν αυτή την ομοιότητα, θα έπρεπε το σύστημα να μπορεί να προσφέρει στο χρήστη και παραπάνω εισηγήσεις. Κάτι τέτοιο δεν συμβαίνει στη περίπτωση του *προκαθορισμένου κατωφλιού*(hard threshold) N ταινιών, ενώ ένα ευμετάβλητο όριο εισηγήσεων θα μπορούσε να αντιμετωπίσει αυτή τη περίπτωση.

Από την παραπάνω ανάλυση κατέστη σαφές ότι χρειαζόμαστε έναν πιο δυναμικό τρόπο επιλογής του αριθμού των "καλών" εισηγήσεων για κάθε ταινία. Το μόνο εργαλείο που έχουμε στα χέρια μας, με βάση τον πίνακα ομοιότητας πάντα, είναι οι τιμές συνάφειας που περιέχονται σε αυτό. Οπότε θα πρέπει να αξιοποιήσουμε αυτές τις τιμές ώστε να μας καταδεικνύουν ποιες είναι καλές εισηγήσεις για ποιες ταινίες. Επομένως, αυτή η αλληλουχία σκέψης μας οδηγεί στο να ορίσουμε κάποιο είδος *κατώφλι*(threshold), το οποίο αν μία τιμή συνάφειας μεταξύ δύο ταινιών το ξεπεράσει, θεωρούμε ότι οι δύο ταινίες αυτές είναι επαρκώς σχετικές, ώστε να τις αλληλοπροτείνουμε. Σε αυτή την περίπτωση υπάρχουν δύο προσεγγίσεις, οι οποίες εμφανίζουν θετικά και αρνητικά και πρέπει κανείς να βρει την ισορροπία ανάμεσα σε αυτά. Πιο συγκεκριμένα:

- Η πρώτη είναι να ορίσουμε ένα *οικουμενικό κατώφλι*(global threshold) για όλες τις ταινίες. Δηλαδή, όσες τιμές συνάφειας είναι πάνω από αυτό το ολικό κατώφλι, να θεωρούμε ότι θα μας δώσουν σχετικά σωστές εισηγήσεις οπότε να τις δεχτούμε. Αν και είναι μία αρκετή λογική προσέγγιση, στην περίπτωση που μία ταινία το σύστημα μας δεν την έχει μοντελοποιήσει πολύ σωστά και επομένως δεν έχει υψηλές τιμές συνάφειας στο αντίστοιχο διάνυσμα ομοιότητας που να ξεπερνάνε το κατώφλι αυτό, αλλά εντούτοις οι υψηλότερες, ως προς την ταξινόμηση, τιμές συνάφειας, οι οποίες ξανατονίζουμε δεν περνάνε αυτό το κατώφλι, αντιστοιχούν σε καλές εισηγήσεις, τότε αυτές τις περιπτώσεις

τις χάνουμε. Δηλαδή, όταν το σύστημα έχει αποτύχει στην ακριβή αναπαράσταση της ταινίας στο χώρο των θεμάτων, που εξετάζουμε, αλλά παρ'όλα ταύτα έχει ενσωματώσει την σχετική σειρά ομοιότητας της εν λόγω ταινίας ως προς τις άλλες, τότε τέτοιες περιπτώσεις θα χαθούν, καθώς οι μικρές τιμές συνάφειας θα μας εμποδίσουν από το να κάνουμε οποιαδήποτε εισήγηση.

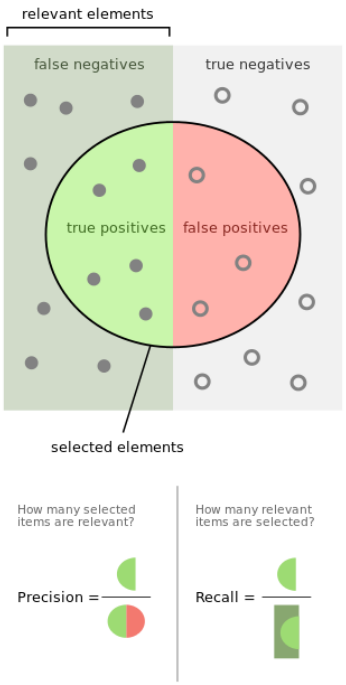
- Στη δεύτερη περίπτωση υπάρχει ένα δυναμικό κατώφλι για κάθε ταινία. Το πως ορίζεται το κάθε κατώφλι είναι βασική ιδέα πίσω από αυτή την ανάλυση. Σε περιπτώσεις όπως η δική μας, που έχουμε μία ταξινομημένη λίστα από τιμές συνάφειας, οι οποίες εκφράζουν συσχέτιση με το αντικείμενο για το οποίο γίνεται ερώτημα, η πλέον διαδεδομένη είναι το κατώφλι να ορίζεται ως ένα ποσοστό της μέγιστης τιμής συνάφειας για τη συγκεκριμένη ταινία. Δηλαδή, έστω max η μέγιστη τιμή συνάφειας σε ένα διάνυσμα ομοιότητας μίας ταινίας(υπενθυμίζουμε ότι διάνυσμα ομοιότητας είναι η γραμμή ή στήλη, της αντίστοιχης ταινίας στο πίνακα ομοιότητας που περιέχει τις τιμές συνάφειας της συγκεκριμένης ταινίας με όλες τις άλλες ταινίες στη βάση ταινιών μας). Όλες οι ταινίες, οι τιμές συνάφειας των οποίων με την συγκεκριμένη ταινία ξεπερνούν ένα ποσοστό αυτής της μέγιστης τιμής, π.χ. 60% του max, δηλαδή $0.6 * max$, θεωρούνται ικανές εισηγήσεις για την συγκεκριμένη ταινία. Όμως και αυτή η ιδέα αντιμετωπίζει κάποια προβλήματα. Στην περίπτωση που το σύστημά μας έχει χάσει τελείως μία ταινία, δηλαδή είναι πολύ λανθασμένη η μοντελοποίησή της και η σχέση της με τις άλλες ταινίες δεν εκφράζει την πραγματική συνάφεια με αυτές, θα ήταν καλύτερο να μην επιστρέψει κάποια εισήγηση λανθασμένη, αλλά με αυτή τη μέθοδο πάντα θα επιστρέφει τουλάχιστον μία σύσταση(προφανώς είναι η ταινία στην οποία αντιστοιχεί η τιμή συνάφειας max που πάντα θα είναι μεγαλύτερη από το δυναμικό κατώφλι το οποίο είναι ποσοστό αυτής). Για λόγους πληρότητας, να αναφέρουμε ότι υπάρχουν και διάφορες άλλες επιλογές για το δυναμικό κατώφλι, εκ των οποίων οι πιο διαδεδομένες είναι η *μέση τιμή*(mean value) (η οποία είναι ευάλωτη σε διανύσματα ομοιότητας με πολύ ακραίες τιμές) και η *ενδιάμεση τιμή*(median value)(η οποία είναι μία καλύτερη εναλλακτική από τη μέση τιμή, ως προς το θέμα των ακραίων τιμών) του διανύσματος ομοιότητας. Αν το διάνυσμα ομοιότητας για μία ταινία είναι : $[x_1, x_2, ..., x_K]$, όπου K το πλήθος των ταινιών (εδώ 160) οι δύο παραπάνω επιλογές είναι :

$$mean_x = \frac{\sum_{i=1}^{K} x_i}{K} \tag{4.1}$$

και:

$$median_x = \begin{cases} \frac{x_{\frac{K}{2}} + x_{\frac{K}{2}+1}}{2} & , K \mod 2 = 0 \\ x_{\frac{K-1}{2}+1} & , K \mod 2 = 1 \end{cases} \tag{4.2}$$

Η ανάλυση που κάναμε παραπάνω μπορεί να γίνει και πιο φορμαλιστικά, με τις έννοιες *precision* και *recall*[98]. Σε ένα πρόβλημα ανάκτησης πληροφορίας στο οποίο μπορούμε να αποφανθούμε ποια είναι σχετικά και ποια όχι ως αποτελέσματα, ως *precision* ορίζουμε το κλάσμα των εισηγήσεων που επιστρέψαμε τα οποία είναι πράγματι σχετικά και ως *recall* το κλάσμα των πραγματικά σχετικών εισηγήσεων το οποίο εμείς αναγνωρίσαμε και επιστρέψαμε. Στην ακόλουθη εικόνα (**Σχήμα 4.2**)[4] γίνεται εποπτικά ξεκάθαρος ο ορισμός των δύο αυτών μέτρων. Σε γενικές γραμμές, θα θέλαμε να έχουμε όσο το δυνατόν μεγαλύτερες τιμές και για τα δύο αυτά μετρικά, καθώς υψηλό *precision* συνεπάγεται υψηλή ακρίβεια αποτελεσμάτων, ενώ υψηλό *recall* σημαίνει ότι επιστρέψαμε πολλά από τα σχετικά αποτελέσματα.



Σχήμα 4.2: Precision και Recall

Το πρόβλημα όμως έγκειται, όπως το αναλύσαμε και ποιοτικά παραπάνω με τα παραδείγματα, ότι άνοδος του ενός μέτρου συνήθως συνεπάγεται μείωση του άλλου.[36] Δηλαδή στην εφαρμογή μας με τις εισηγήσεις, αν θέλουμε να αυξήσουμε την ακρίβεια του συστήματος μας, το

[4] https://en.wikipedia.org/wiki/Precision_and_recall

precision, θα πρέπει να επιστρέφουμε εισηγήσεις μόνο όταν είμαστε απόλυτα σίγουροι ότι είναι σωστές, οπότε σε γενικές γραμμές να μειώσουμε τον αριθμό των εισηγήσεων που επιστρέφουμε, πράγμα που οδηγεί σε μείωση του *recall*. Αντίστροφα, για να αυξήσουμε το *recall* πρέπει να αυξήσουμε τον αριθμό των επιστρεφόμενων αποτελεσμάτων πράγμα που συνήθως οδηγεί σε μείωση του *precision*. Οπότε πρέπει να έρθουμε σε ένα *συμβιβασμό*(tradeoff) ανάμεσα στα δύο μεγέθη. Υπάρχουν διάφορα μετρικά που χρησιμοποιούνται που τα συνδυάζουν, όπως το *F-1 score* ή ο συντελεστής *Matthews correlation coefficient*, που δεν είναι του παρόντος να επεκταθούν, καθώς εμάς μας ενδιαφέρει κυρίως η επιλογή σωστού κατωφλίου για ανάκτηση πληροφορίας. Σε τέτοιου είδους εφαρμογές γενικώς είναι προτιμότερος ένας συνδυασμός με υψηλό *recall* και μικρό *precision*, σε σχέση με μικρό *recall* και υψηλό *precision*, διότι μπορεί ο χρήστης να επιλέξει από μεγαλύτερο εύρος εισηγήσεων, ακόμα και αν μερικές από αυτές δεν είναι σωστές, πράγμα που ο ίδιος μπορεί να το ξεχωρίσει, παρά να ψάξει εξ'αρχής όλη τη βάση δεδομένων γιατί δεν είχε αρκετές εισηγήσεις.[107] Προς αυτό το σκοπό επιλέξαμε να συνδυάσουμε τις προσεγγίσεις με ένα δυναμικό και ένα οικουμενικό κατώφλι. Συγκεκριμένα, ορίζουμε ένα οικουμενικό κατώφλι συνάφειας με αρκετά χαμηλή τιμή, για να αποκλείσουμε τις ταινίες που δεν έχουν μοντελοποιηθεί καθόλου σωστά, και το συνδυάζουμε με ένα δυναμικό κατώφλι, διαφορετικό για κάθε ταινία, ως ποσοστό της μέγιστης τιμής του διανύσματος συνάφειας για αυτή τη ταινία. Το οικουμενικό κατώφλι και το ποσοστό ορίστηκαν αντίστοιχα **0.2** και **80%**, έπειτα από πειραματικές δοκιμές και επιλογή του καλύτερου συνδυασμού για το σύστημα μας. Έτσι για κάθε ταινία, κοιτάμε το διάνυσμα ομοιότητάς της και όσες ταινίες ξεπερνούν το παραπάνω συνδυασμό δυναμικού και σταθερού κατωφλιού, τις θεωρούμε ως καλές εισηγήσεις και τις επιστρέφουμε στο χρήστη. Το αν αυτές είναι πράγματι καλές επιλογές, θα φανεί στη συνέχεια και αυτό μας οδηγεί στο επόμενο κεφάλαιο, καθώς για να κριθεί μία εισήγηση ως καλή ή όχι, απαιτείται ένα αντικειμενικό κριτήριο αλήθειας(στη περίπτωση μας είναι η βάση αλήθειας που κατασκευάσαμε προηγουμενως) και για να μετρηθεί η ποιότητα της εισήγησης, ένα κατάλληλο μέτρο.

4.4 Μέτρα Απόδοσης

Σε αυτή την ενότητα θα ασχοληθούμε με τα *μέτρα απόδοσης*(evaluation measures) που σχεδιάστηκαν για την αξιολόγηση του συστήματός μας. Έχει γίνει μια αναφορά σε θέματα απόδοσης της *LDA* μεθόδου και στο (*3.2.3*), αλλά εδώ ενδιαφερόμαστε από τη σκοπιά ανάκτησης πληροφορίας. Υπάρχουν κυρίως δύο τρόποι με τους οποίους μετριέται η απόδοση ενός

topic-model συστήματος.

Ο πρώτος είναι μέσω υπολογισμού της *πιθανοφάνειας*(likelihood) σε ένα δείγμα από τα δεδομένα μας το οποίο έχουμε φροντίσει να μη δει το σύστημά μας(heldout sample)[15, 14]. Για αυτές τις περιπτώσεις, υπάρχουν διάφορα μετρικά για το πως η πληροφορία που έχει λάβει το σύστημα μας από ένα corpus, μπορεί να εφαρμοστεί σε δεδομένα που δεν έχουμε ξαναδεί. Αν και αυτά τα μετρικά έχουν καλές δυνατότητες γενίκευσης σε άλλα μοντέλα και διευκολύνουν συγκρίσεις μεταξύ αυτών, εντούτοις δεν ενδιαφέρονται για το τρόπο αναπαράστασης των δεδομένων εσωτερικά σε κάθε μοντέλο.[16] Αντιθέτως, στην περίπτωσή μας ενδιαφερόμαστε αρκετά για τον τρόπο αναπαράστασης των δεδομένων καθώς είναι ένα κέρδος που έχουμε, αυτή η υψηλότερου γνωσιακού επιπέδου εκπροσώπηση των ταινιών ως μείγματα θεμάτων. Επίσης, δεδομένου του μικρού αριθμού ταινιών στη βάση δεδομένων μας, το σύστημά μας δεν έχει απεριόριστες δυνατότητες γενίκευσης, έτσι οι μετρήσεις για το likelihood και το perplexity, σχετικά με νέα δεδομένα δεν θα είναι πολύ διαφωτιστικές. Αυτό γιατί ο σκοπός μας δεν είναι απαραίτητα να δούμε αν το σύστημά μας μπορεί να αναπαραστήσει αρκετά καλά καινούργια unseen δεδομένα, αλλά να προσφέρει ένα διαφορετικό τρόπο αναπαράστασης ταινιών ο οποίος και θα οδηγεί σε διαφορετικού τύπου εισηγήσεις και θα επιτρέπει την εξερεύνηση των ταινιών από διαφορετική σκοπιά(αυτή των θεμάτων). Παρ'όλα ταύτα να τονιστεί εδώ ότι δοκιμάσαμε στα πλαίσια της διπλωματικής να χρησιμοποιήσουμε ως *test set*, μία συλλογή 30 υποτίτλων, αλλά για τους λόγους που αναφέραμε παραπάνω δεν είχαμε ικανοποιητικά αποτελέσματα, ώστε να χρησιμοποιήσουμε μία τέτοια μέθοδο.

Έτσι στρεφόμαστε στο δεύτερο τρόπο υπολογισμού της απόδοσης του συστήματός μας, αυτόν σχετικά με την επιτυχία στην ανάκτηση πληροφορίας, καθώς ταιριάζει και περισσότερο στην εφαρμογή μας. Σε αυτή τη περίπτωση, πρέπει να βρει κάποιος μία μεθοδολογία αντιπαραβολής αποτελεσμάτων για να κρίνει αν οι επιστρεφόμενες εισηγήσεις είναι αρκετά καλές. Το μέτρο σύγκρισης προφανώς είναι η βάση αλήθειας(ground truth, ή αλλιώς golden rule) που περιγράψαμε προηγουμένως. Εμείς χρησιμοποιήσαμε κυρίως δύο τέτοια μέτρα.

4.4.1 Μέση Βαθμολογία Εισηγήσεων (Mean Recommendations Score)

Έχοντας λοιπόν το πίνακα ομοιότητας του συστήματος που μας ενδιαφέρει (πρωτίστως του topic μοντέλου, αλλά ομοίως είναι και για το tf-idf και τα fusion μοντέλα) και του πίνακα ομοιότητας της βάσης αλήθειας, θέλουμε να βαθμολογήσουμε τις εισηγήσεις που μας επιστρέφει,

με τον τρόπο που περιγράψαμε στο (**4.3**), το μοντέλο μας. Η βαθμολόγηση αυτή αφορά τη συνάφεια που έχουν οι εισηγήσεις μας με την ταινία. Αλλά αυτή η συνάφεια δεν εκφράζεται πραγματικά στον πίνακα ομοιότητας του μοντέλου, αυτή είναι η εκτίμηση του μοντέλου μας, αλλά στο πίνακα ομοιότητας της βάσης αλήθειας. Δηλαδή, μας ενδιαφέρουν οι αριθμητικές τιμές συνάφειας στη βάση αλήθειας, για τις εκάστοτε εισηγήσεις κάθε διαφορετικής ταινίας. Αυτό το πετυχαίνουμε ως εξής: Αρχικά και για την εκάστοτε ταινία, βρίσκουμε τις πιο συναφείς ταινίες από το διάνυσμα ομοιότητας του μοντέλου που εξετάζουμε. Έπειτα, κρατώντας τους δείκτες θέσης αυτών των ταινιών, πηγαίνουμε στο αντίστοιχο διάνυσμα ομοιότητας για τη συγκεκριμένη ταινία, του πίνακα αλήθειας αυτή τη φορά, και κρατάμε τις τιμές συνάφειας για κάθε μία από αυτές τις εισηγήσεις. Παίρνουμε τον μέσο όρο των τιμών για αυτές τις εισηγήσεις και αυτή είναι η βαθμολογία που παίρνει το σύστημά μας για τη συγκεκριμένη ταινία. Επαναλαμβάνουμε αυτή τη διαδικασία για όλες τις ταινίες που έχουμε στη διάθεσή μας, οπότε έχουμε μία βαθμολογία για κάθε ταινία. Η *Μέση Βαθμολογία Εισηγήσεων*(Mean Recommendations Score-**MRS**), είναι ο μέσος όρος αυτών των βαθμολογιών. Πιο φορμαλιστικά μπορεί να οριστεί ως :

$$\mathbf{MRS} = \frac{\sum_{i=1}^{N} \frac{\sum_{j=1}^{K_i} GSM(i, index(j))}{K_i}}{N} \tag{4.3}$$

, όπου στη παραπάνω εξίσωση είναι: N το πλήθος της βάσης ταινιών(160 εδώ), K ο αριθμός εισηγήσεων για κάθε ταινία(διαφορετικός από ταινία σε ταινία λόγω του δυναμικού τρόπου ορισμού του κατωφλιού), $GSM(i,:)$ το διάνυσμα ομοιότητας του πίνακα αλήθειας για τη ταινία i και $index(j)$ οι δείκτες θέσης των ταινιών-εισηγήσεων πάνω στο διάνυσμα ομοιότητας για τη συγκεκριμένη ταινία.

4.4.2 Μέση Απώλεια Κατάταξης (Mean Ranking Loss)

Με το προηγούμενο μέτρο μετράμε κυρίως το βαθμό συνάφειας των εισηγήσεών μας, όπως αυτός υπολογίζεται από τη βάση αλήθειας. Αυτές οι τιμές συνάφειας όμως δεν ενσωματώνουν την έννοια της σειράς κατάταξης των εισηγήσεων μας, συγκρινόμενες πάντα με τη βάση αλήθειας που σχεδιάσαμε. Γι'αυτό σχεδιάσαμε το ακόλουθο μέτρο της *Μέσης Απώλειας Κατάταξης*(Mean Ranking Loss-**MRL**). Αφού σκοπό έχουμε να προσφέρουμε καλύτερες εισηγήσεις, επιλέξαμε αυτό το μέτρο ως κριτήριο βελτιστοποίησης του συστήματος για να επιλέξουμε το κατάλληλο μοντέλο. Η βάση της σκέψης ήταν να μετράμε πόσο απέχει σε σειρά κατάταξης η κάθε εισήγηση από την ιδανική. Δηλαδή, αν για παράδειγμα για μία ταινία

επιστρέφαμε τρεις εισηγήσεις θα θέλαμε αυτές οι εισηγήσεις να ήταν αντίστοιχα οι $[1, 2, 3]$ σε σειρά κατάταξης στη βάση αλήθειας. Αυτό όμως δεν είναι πολύ πιθανό να συμβεί, οπότε εμείς θέλουμε να μετρήσουμε την απώλεια που έχει το σύστημά μας σε περίπτωση που οι επιστρεφόμενες τιμές απέχουν από το ιδανικό. Αρχικά, κοιτάμε οι εισηγήσεις που έκανε το σύστημα μας, σε τι σειρά κατάταξης βρίσκονται στη λίστα των "σωστών" εισηγήσεων, όπως αυτές προκύπτουν από το διάνυσμα ομοιότητας της βάσης αλήθειας για αυτή τη ταινία. Έπειτα, μετράμε την απόσταση κάθε εισήγησης από την αντίστοιχη ιδανική, όσον αφορά τη σειρά κατάταξης. Για να επιστρέψουμε στο παράδειγμα μας με τις 3 εισηγήσεις, έστω ότι οι εισηγήσεις που επιστρέψαμε είχαν την μορφή $[2, 5, 8]$, δηλαδή η 1η δική μας εισήγηση είναι σύμφωνα με τη βάση αλήθειας η 2η πιο σχετική εισήγηση, η 2η δική μας εισήγηση είναι στη πραγματικότητα η 5η και η 3η εισήγηση μας είναι η 8η πιο συναφής σύμφωνα πάντα με το διάνυσμα ομοιότητας του πίνακα της βάσης αλήθειας για αυτή τη ταινία. Η ιδανική περίπτωση όπως προείπαμε θα ήταν να επιστραφεί η λίστα εισηγήσεων με σειρά κατάταξης $[1, 2, 3]$. Η απώλεια ως προς αυτή τη λίστα, τη λίστας που επέστρεψε το μοντέλο μας, υπολογίζεται ως το άθροισμα της απόστασης κάθε στοιχείου από την ιδανική του θέση, δηλαδή για τη 1η εισήγηση είναι $2 - 1 = 1$, για τη 2η είναι $5 - 2 = 3$ και για τη 3η είναι $8 - 3 = 5$, άρα συνολικά η απώλεια του συστήματος για τη συγκεκριμένη ταινία είναι $rankingloss = 1 + 3 + 5 = 9$. Επειδή όμως αυτή η απώλεια είναι για κάθε εισήγηση αθροιστικά, διαιρούμε με το πλήθος των εισηγήσεων για να βρούμε τη μέση απώλεια. Άρα για τη συγκεκριμένη ταινία είναι $meanrankingloss = \frac{9}{3} = 3$.

Σε αυτό το σημείο θα κάνουμε μία μικρή παρένθεση για να εξηγήσουμε καλύτερα στη συνέχεια την εξέλιξη αυτού του μέτρου. Όπως προείπαμε το μέτρο αυτό θα αποτελούσε κριτήριο βελτιστοποίησης του μοντέλου, τόσο για τον αριθμό των topics, όσο και για ποιο μοντέλο ακριβώς, δοθέντων διάφορων μοντέλων με συγκεκριμένο αριθμό topics, θα επιλέξουμε. Αν αφήσουμε όμως το μέτρο όπως έχει, θα δούμε ότι τα μοντέλα που τα πηγαίνουν καλύτερα σε τέτοια περίπτωση είναι μοντέλα τα οποία προωθούν μικρό αριθμό εισηγήσεων ανά ταινία, καθώς έτσι είναι πιο πιθανό να μειώσουν τις απώλειες κατάταξης(με γνώμονα ότι τα μοντέλα μας δεν είναι ιδανικά, όσο περισσότερες εισηγήσεις κάνουν, τόσο περισσότερο αυξάνει η πιθανότητα να έχουν κάποια κακής ποιότητας εισήγηση και επομένως να έχουν αυξημένες απώλειες κατάταξης). Αλλά αυτό είναι αντιπαραγωγικό για το σκοπό μας καθώς θα θέλαμε έναν επαρκή αριθμό εισηγήσεων ανά ταινία. Αποφασίσαμε λοιπόν να βάλουμε μία *ποινή*(penalty) για κάθε πρόταση που έχει λιγότερες από $K = 4$ εισηγήσεις. Η τιμή αυτή προέκυψε παρατηρώντας τη βάση ταινιών και τις συσχετίσεις μεταξύ των ταινιών και πειραματιζόμενοι με διάφορες τι-

μές, καθώς 4 καλές εισηγήσεις είναι πολλές φορές δύσκολο να βρεθούν σε τόσο μικρό σετ δεδομένων. Συγκεκριμένα, σε κάθε πρόταση που έχει λιγότερες από 4 εισηγήσεις, θα προσθέτουμε 100 στη συνολική απώλεια κατάταξης, για κάθε μία εισήγηση λιγότερη από τις 4. Δηλαδή, στο προηγούμενο παράδειγμα η τελική απώλεια κατάταξης για τη συγκεκριμένη ταινία θα ήταν $mean\_ranking\_loss = \frac{9+100}{4} = 27.25$(προσθέσαμε μία φορά 100 καθώς είχαμε 3 εισηγήσεις αντί για 4, αν είχαμε 2 εισηγήσεις θα προσθέταμε 200 κ.ο.κ). Αυτό το κάνουμε για να τιμωρήσουμε τα μοντέλα τα οποία μας προσφέρουν λίγες εισηγήσεις για να αποφύγουν υψηλές απώλειες κατάταξης. Η τιμή 100 επιλέχτηκε πειραματικά.(αρκεί βασικά να είναι λίγο μεγαλύτερη από 80 που είναι η μέση τιμή απώλειας για ένα μοντέλο τυχαίων εισηγήσεων)

Η τελευταία τροποποίηση στο μέτρο αυτό, προέρχεται από την λογική παρατήρηση ότι μας ενδιαφέρει περισσότερο να είναι πετυχημένες οι πρώτες εισηγήσεις και λιγότερο οι τελευταίες. Για παράδειγμα, αν για μία ταινία μας επιστρέψει το σύστημα μας 7 εισηγήσεις δεν θα πρέπει να μας πειράζει τόσο αν η 6η και η 7η δεν είναι καλές, ενώ οι άλλες έχουν πετύχει. Δηλαδή, η βαρύτητα των τελευταίων εισηγήσεων στη ποιότητα του συστήματος δεν θα έπρεπε να είναι η ίδια με των πρώτων. Για το σκοπό αυτό αποφασίσαμε να προσθέσουμε ένα *σχέδιο βαρών*(weighting scheme), ώστε να υπολογίσουμε πραγματικά το *σταθμισμένο μέσο όρο*(weighted arithmetic mean) των απωλειών. Ο τρόπος που υπολογίσαμε τα βάρη-συντελεστές βαρύτητας ήταν αρχικά μία σταδιακή *απομείωση*(decay) της βαρύτητας κάθε βάρους. Συγκεκριμένα, για K εισηγήσεις, η πρώτη εισήγηση θα είχε L βαθμούς βαρύτητας, η δεύτερη $L-1$ κ.ο.κ μέχρι τη τελευταία εισήγηση η οποία θα έχει 1 βαθμό βαρύτητας. Αυτό το σχήμα θα έχει σαν αποτέλεσμα να υπάρχουν συνολικά $1+2+..+L-1+L = \frac{L(L+1)}{2}$ βαθμοί βαρύτητας. Η τιμή λοιπόν κάθε βαθμού βαρύτητας προκύπτει ως $w = \frac{1}{\frac{L(L+1)}{2}}$, καθώς θέλουμε να αθροίζουν στη μονάδα. Άρα το διάνυσμα βαρών για K εισηγήσεις τελικά είναι : $weights = [1*w, 2*w, ..., (L-1)*w, L*w]$. Κάνοντας δοκιμές όμως με αυτό το τρόπο υπολογισμού των βαρών βλέπαμε ότι και πάλι δινόταν παραπάνω έμφαση, απ'ότι θα θέλαμε, σε μικρής σημασίας εισηγήσεις. Επίσης, εκ των πραγμάτων τα συστήματά μας δεν έχουν συχνά περισσότερες από 4 εισηγήσεις για κάθε ταινία, επομένως προτιμήσαμε ένα ασύμμετρο σχήμα υπολογισμού των βαρών που να ανταποκρίνεται στις πραγματικές ανάγκες του συστήματος δεδομένου του μέσου πλήθους εισηγήσεων ανά ταινία. Συγκεκριμένα, τα βάρη ορίζονται με

βάση το πλήθος K των εισηγήσεων για κάθε ταινία ως εξής :

$$weights = \begin{cases} [1] & , K = 1 \\ [0.6, 0.4] & , K = 2 \\ [0.5, 0.3, 0.2] & , K = 3 \\ [0.4, 0.3, 0.2, 0.1] & , K = 4 \\ [0.4, 0.3, 0.2, x, x, ..., x] & , K > 4 \end{cases} \tag{4.4}$$

, όπου στη περίπτωση που έχουμε πάνω από 4 εισηγήσεις, από τη 4η και μετά κάθε μία λαμβάνει την ίδια μικρή βαρύτητα $x = \frac{0.1}{K-3}$.

Έτσι για να επανέλθουμε στα προηγούμενα, η σταθμισμένη πλέον μέση απώλεια κατάταξης για αυτή τη ταινία είναι :

$$\begin{aligned} mean\_ranking\_loss =& \frac{\sum_{i=1}^{K} weights[i] * recommendation\_rank[i]}{K} \\ =& \frac{1 * 0.4 + 3 * 0.3 + 5 * 0.2 + 100 * 0.1}{4} = 12.3 \end{aligned} \tag{4.5}$$

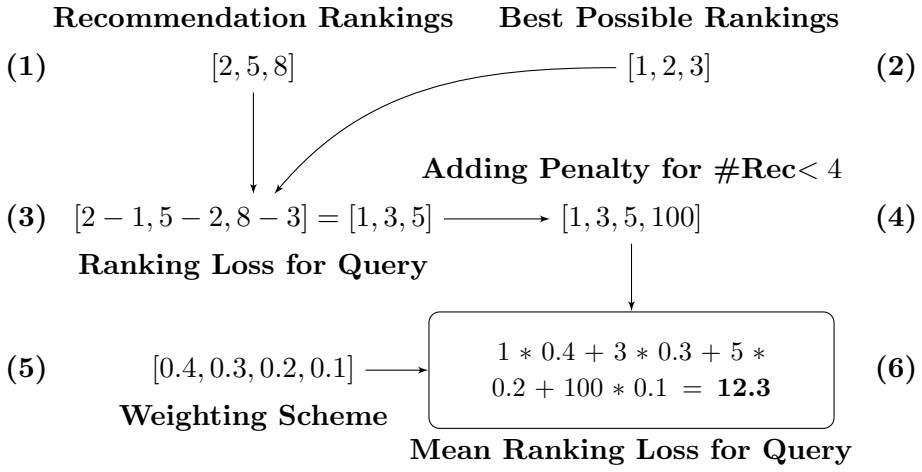
, όπου K ο αριθμός των εισηγήσεων για τη συγκεκριμένη ταινία, $weights$ ο κατάλληλος πίνακας βαρών και $recommendation\_rank$, η λίστα με τη σειρά κατάταξης κάθε εισήγησης. Το παραπάνω μέγεθος όπως εξηγήσαμε αφορά μία ταινία μόνο, άρα τελικά η *Μέσης Απώλειας Κατάταξης*(Mean Ranking Loss-**MRL**) υπολογίζεται ως μέσος όρος όλων των ταινιών ως εξής:

$$\mathbf{MRL} = \frac{\sum_{i=1}^{N} mean\_ranking\_loss[i]}{N} \tag{4.6}$$

, όπου $mean\_ranking\_loss[i]$ η μέση απώλεια κατάταξης κάθε ταινίας.

Τα παραπάνω φαίνονται και γραφικά στο **Σχήμα 4.4.2**.

Κλείνοντας αυτή τη παράγραφο, θα αναφέρουμε απλώς ότι στους παρακάτω πίνακες αποτελεσμάτων πέρα από τα δύο αυτά μέτρα, όπου όπως αναφέραμε πιο βασικό είναι η *Μέσης Απώλειας Κατάταξης*(Mean Ranking Loss-**MRL**), εμφανίζονται και άλλες δύο ποιοτικές μετρήσεις. Η μία είναι η *ενδιάμεση τιμή*(median value) της σειράς κατάταξης για κάθε μία από τις πρώτες εισηγήσεις(προφανώς, αφού αφορά το σύνολο της επίδοσης του μοντέλου πάνω



Σχήμα 4.3: Σχηματικό παράδειγμα για το MRL μέτρο απόδοσης.

στην βάση των ταινιών μας, πρόκειται για τη μέση τιμή των ενδιάμεσων τιμών των τριών πρώτων εισηγήσεων). Η άλλη είναι(πάλι πρόκειται για μέσες τιμές, καθώς μιλάμε για την επίδοση του συστήματος σε όλες τις ταινίες) το ποσοστό των τριών πρώτων εισηγήσεων που είναι στο Top 10 και Top 20 των πραγματικών εισηγήσεων σύμφωνα με τη βάση αλήθειας. Δηλαδή, το ποσοστό πάνω σε όλες τις πρώτες εισηγήσεις του συστήματός μας που βρίσκεται εντός των 10 και 20 καλύτερων εισηγήσεων αντίστοιχα, σύμφωνα με τη βάση αλήθειας, το ποσοστό πάνω σε όλες τις δεύτερες εισηγήσεις του συστήματός μας που βρίσκεται εντός των 10 και 20 και ομοίως για τις τρίτες εισηγήσεις. Αυτά είναι πρόσθετα ποιοτικά μέτρα που μας δείχνουν την επίδοση του συστήματος μας. Τέλος, να αναφέρουμε εδώ ότι μερικές ενδιαφέρουσες εναλλακτικές στο δικό μας μέτρο της *Μέσης Απώλειας Κατάταξης*, απαντώνται στη βιβλιογραφία ως Top-K Lists ranking[28, 37] ή Indefinite Lists ranking[105, 52].

4.5 Πειραματικά Αποτελέσματα και Συμπεράσματα

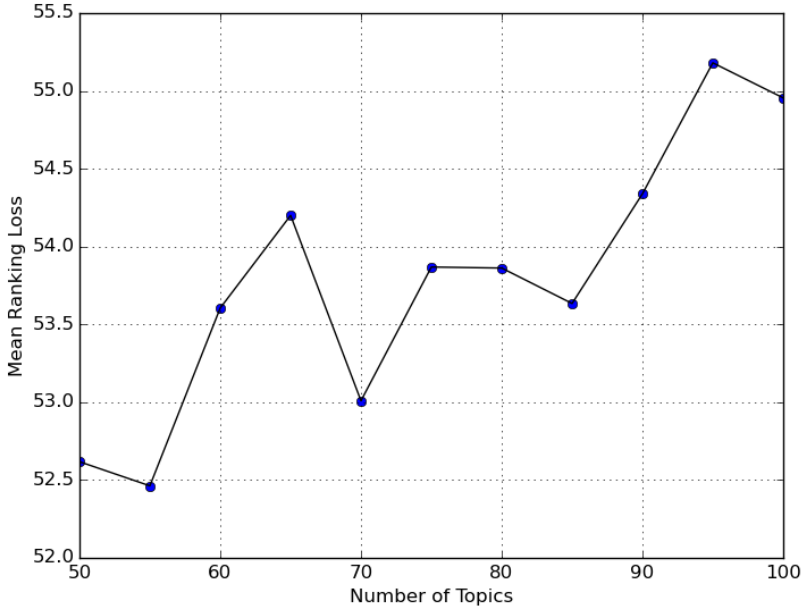
Σε αυτή την ενότητα θα αναφερθούμε στα πειραματικά αποτελέσματα, στο πλαίσιο της περιγραφής που έγινε προηγουμένως. Επίσης, θα παρουσιάσουμε διαφορές ανάμεσα στα μοντέλα (topic, tf-idf, fusion, κτλπ.) και θα αναφερθούμε και σε συγκεκριμένα παραδείγματα ανάκτησης πληροφορίας.

Αρχικά, παίρνοντας τη σκυτάλη από τη προηγούμενη ενότητα θα αναφερθούμε στην επιλογή του αριθμού των topics. Όπως περιγράψαμε και στο θεωρητικό σκέλος αλλά και στα μέτρα απόδοσης, υπάρχουν διάφοροι τρόποι υπολογισμού της ποιότητας ενός μοντέλου topic. Εμάς

δεν μας ενδιαφέρει η καλή γενίκευση του μοντέλου, οπότε perplexity ή likelihood measures που αφορούν held out δεδομένα, αν και εύκολα κατανοητά και εφαρμόσιμα δεν είναι κατάλληλα για την εφαρμογή μας(έτσι και αλλιώς η συγκεκριμένη υλοποίηση δεν είναι online, δηλαδή σε περίπτωση που θέλουμε να προσθέσουμε νέες ταινίες στο topic μοντέλο πρέπει να εφαρμόσουμε από την αρχή τον αλγόριθμο).Θέλουμε λοιπόν το σύστημά μας να προσφέρει καλή αναπαράσταση των ταινιών στο topic space, ώστε να είναι δυνατή η *περιήγηση*(browsing) των ταινιών μέσω αυτού του χώρου, αλλά και να προσφέρει καλές εισηγήσεις σε ερωτήματα του χρήστη. Ορμώμενοι από τα παραπάνω σχεδιάσαμε τη *Μέση Απώλεια Κατάταξης*(**MRL**) και αυτό είναι το μέτρο ως προς το οποίο θα βελτιστοποιήσουμε τον αριθμό των topic.

Για το σκοπό αυτό εκπαιδεύσαμε διάφορα topic μοντέλα, με διαφορετικό αριθμό θεμάτων και υπολογίσαμε το MRL τους. Συγκεκριμένα, τα μοντέλα είχαν από 50 έως 100 topics, με βήμα 5, δηλαδή $50, 55, .., 100$. Επίσης, για κάθε διαφορετικό αριθμό θεμάτων, εκπαιδεύαμε 5 τέτοια μοντέλα και παίρναμε τους μέσους όρους για το MRL για τη συγκεκριμένη ομάδα μοντέλων, έτσι ώστε να είναι πιο εύρωστη η επιλογή μας(το model averaging μας επιτρέπει να να έχουμε καλύτερα αποτελέσματα καθώς έχουμε πρόσβαση σε πολλαπλά μέγιστα της posterior). Το αποτέλεσμα από αυτή τη διαδικασία φαίνεται στην επόμενη εικόνα(**Σχήμα 4.4**), όπου βλέπουμε μία βύθιση στα 55 topics, οπότε επιλέγουμε το συγκεκριμένο αριθμό. Το πεδίο $50 - 100$ επιλέχτηκε έτσι καθώς, πιο πολλά από 100 κατακερματίζουν πάρα πολύ το topic space, δεδομένου ότι έχουμε μόνο 160 ταινίες, και αυτό οδηγεί σχεδόν σε ένα topic ανά ταινία, ενώ λιγότερα από 50 δημιουργεί πολύ μπλεγμένα topics και δεν οδηγεί σε καλή μοντελοποίηση. Έχοντας βρει το κατάλληλο αριθμό από topics, εκπαιδεύσαμε άλλη μία φορά μία ομάδα από μοντέλα, με σταθερό αριθμό 55 θεμάτων και επιλέξαμε το καλύτερο, πάλι ως προς το MRL.

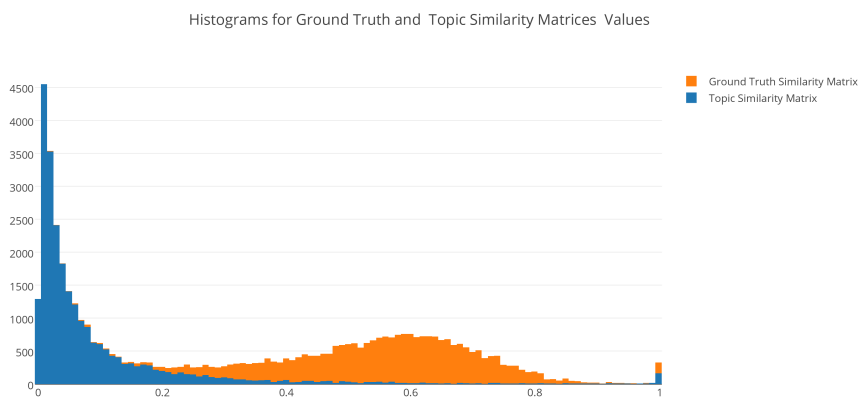
Σε αυτό το σημείο και πριν τα τελικά αποτελέσματα, θα αναφερθούμε σύντομα στις ποιοτικές διαφορές ανάμεσα στους πίνακες ομοιότητας της βάσης αλήθειας(**Σχήμα 4.1**) και του topic μοντέλου(**Σχήμα 3.13**). Όπως μπορεί να παρατηρήσεις κανείς ο ground truth πίνακας ομοιότητας είναι πολύ πιο έντονα λευκός απ'ότι ο αντίστοιχος του topic space. Δηλαδή, οι τιμές συνάφειας μεταξύ των ταινιών στη πρώτη περίπτωση είναι σε γενικές γραμμές πολύ μεγαλύτερες απ'ότι στη δεύτερη. Αυτό μπορεί κανείς να το παρατηρήσει και από το διπλανό colorscale στα γραφήματα καθώς, στη περίπτωση της βάσης αλήθειας το μαύρο εκφράζει τιμές πιο κοντά στο ≈ 0.4 παρά στο ≈ 0.1 όπως είναι με το topic space. Αυτό συμβαίνει γιατί οι δύο



Σχήμα 4.4: Mean Ranking Loss του LDA μοντέλου, για διάφορες τιμές πλήθους topic

αυτοί χώροι αναπαριστούν αρκετά διαφορετικά τα δεδομένα. Μη ξεχνάμε, ότι στο topic space οι ταινίες αναπαριστώνται ως διανύσματα 55 διαστάσεων, ενώ στο tag space ως διανύσματα 1128 διαστάσεων. Επίσης, οι ταινίες έχουν μεγάλο βαθμό συμμετοχής σε $1-3$ topics περίπου, ενώ στο tag space συμμετέχει κάθε ταινία σε πολύ μεγαλύτερο αριθμό από tags και λόγω του πλουραλισμού που διακατέχει το χώρο των ετικετών(όπως π.χ. μεταδεδομένα), κάθε ταινία μπορεί να ενεργοποιεί ένα μεγάλο αριθμό απο tags, διευκολύνοντας έτσι το συσχετισμό της με άλλη ταινία. Αυτό γίνεται εμφανές και από το επόμενο γράφημα(**Σχήμα 4.5**), το οποίο απεικονίζει τα ιστογράμματα των τιμών συνάφειας για αυτές τις δύο αναπαραστάσεις.

Όπως βλέπει κανείς, οι τιμές συνάφειας στη βάση αλήθειας ακολουθούν πιο ομαλά μία κανονική κατανομή με μέση τιμή συνάφειας 0.6 περίπου. Αυτό δικαιολογεί την ανάλυση που κάναμε προηγουμένως, οπότε και τώρα φαίνεται ότι κατά μέσο όρο και θεωρητικά όχι τόσο σχετικές ταινίες, μπορεί να έχουν υψηλή τιμή συνάφειας(γύρω από τη μέση τιμή). Από την άλλη, οι τιμές συνάφειας του δικού μας μοντέλου ακολουθούν την εκθετική κατανομή, με μία μικρή κορυφή κοντά στο 1. Αυτό δείχνει ότι οι μεγαλύτερο πλήθος ταινιών μεταξύ τους έχουν μικρή τιμή συνάφειας, αλλά ξεχωρίζουν οι λίγες αυτές που "ταιριάζουν" μεταξύ τους, σύμφωνα πάντα με το σύστημά τους. Δηλαδή, το δικό μας σύστημα κάνει πολύ πιο επιθετική κατηγοριοποίηση της σχέσης μεταξύ των ταινιών, προτιμώντας είτε έντονες συσχετίσεις είτε



Σχήμα 4.5: Ιστογράμματα των τιμών συνάφειας του LDA και του ground truth μοντέλου

πολύ μικρές τιμές συνάφειας. Αυτό βοηθάει στο διαχωρισμό σχετικών και μη ταινιών, αλλά από την άλλη δεν δίνει τόσες πολλές επιλογές ώστε να επιστραφούν στο χρήστη παραπάνω ταινίες(προφανώς γιατί σε κάθε ταινία θα υπάρχουν μόνο $2-3$ ταινίες που περνάνε το κατώφλι ομοιότητας για κάθε τέτοια ταινία και οι υπόλοιπες θα έχουν πολύ χαμηλή τιμή συνάφειας).

Θα παρουσιάσουμε τον πίνακα(**Πίνακας 4.1**) με τα τελικά αποτελέσματα για τα διάφορα μοντέλα με τα οποία ασχοληθήκαμε. Έχουμε προσθέσει στη λίστα και ένα μοντέλο τυχαίας επιλογής εισηγήσεων(Random) για λόγους πληρότητας στην αντιπαραβολή που κάνουμε. Αρχικά, να εξηγήσουμε ότι ως E.F, εννοείται μοντέλο Early Fusion και με M.F, μοντέλο Mid Fusion, όπως αυτά εξηγήθηκαν στα προηγουμένως(**3.5**). Επίσης, τα αρχικά A, M, T που χρησιμοποιούνται σε μοντέλα σύντηξης, σημαίνουν Audio, Music, Topic αντίστοιχα και δείχνουν ποια modalities συμμετέχουν στο συγκεκριμένο μοντέλο σύντηξης. Ακόμα, οι τιμές που εμφανίζονται στα μοντέλα M.F-Mid Fusion δίπλα από τα διάφορα modalities, δείχνουν το ποσοστό συμμετοχής κάθε πίνακα ομοιότητας στο τελικό πίνακα ομοιότητας του μοντέλου.

Αναφερόμενοι στα αποτελέσματα τώρα, μπορεί να παρατηρήσει κανείς αρχικά ότι το topic μοντέλο είναι καλύτερο από το μοντέλο με τη tf-idf αναπαράσταση σε όλους τους τομείς. Επίσης, βλέπει κανείς στη συνέχεια ότι τα μοντέλα audio, music δεν είναι καλά από μόνα τους. Παρ'όλα ταύτα, παρατηρεί κανείς στη συνέχεια ότι υπάρχουν καλά μοντέλα σύντηξης που βελτιώνουν συγκεκριμένους τομείς επίδοσης του σκέτου topic μοντέλου, τόσο Early Fusion όσο και Mid Fusion.

Συγκεκριμένα, παρατηρούμε ότι το μοντέλο με Early Fusion (Audio, Topic) και το μον-

	Mean Ranking Score	**Mean Ranking Loss**	**Median (1st Rec)**	**Median (2nd Rec)**	**Median (3rd Rec)**	**1st Rec Top10** %	**2nd Rec Top10** %	**3rd Rec Top10** %	**1st Rec Top20** %	**2rd Rec Top20** %	**3rd Rec Top20** %
Ground Truth	0.803	0.0	1.0	2.0	3	100.0	100.0	100.0	100.0	100.0	100.0
Topic	**0.661**	**47.8**	**13.0**	17.5	22	43.8	40.6	32.9	57.5	50.8	48.0
Tf-idf	0.627	53.8	18.0	28.0	39	39.4	29.7	27.4	51.9	41.4	37.0
Music	0.512	74.0	62.0	62.5	61	13.1	8.6	23.3	25.0	17.2	30.1
Audio	0.520	70.8	54.0	55.0	72	11.9	12.5	5.5	24.4	24.2	15.1
E.F(M,T)	0.655	48.5	**12.0**	25.0	**18**	44.4	32.8	**38.4**	54.4	46.9	52.1
E.F(A,T)	**0.668**	**46.3**	14.0	**15.5**	23	44.4	39.8	32.9	54.4	**57.0**	49.3
E.F(A,M)	0.529	70.0	56.0	54.5	71	19.4	13.3	12.3	31.3	28.1	20.6
E.F(A, M, T)	0.657	47.9	13.0	17.0	32	43.8	40.6	23.3	56.9	54.7	39.7
M.F(T 0.7, M 0.3)	0.664	47.0	13.5	17.5	18	45.0	41.4	35.6	56.9	51.6	**52.1**
M.F(T 0.5, A 0.5)	0.662	47.5	13.0	17.0	25	43.8	**43.0**	34.3	58.1	51.6	46.6
M.F(A 0.9, M 0.1)	0.541	67.8	51.0	49.5	74.0	16.9	14.8	13.7	33.1	27.3	23.3
M.F(T 0.5, A 0.3, M 0.2)	**0.665**	**47.1**	**12.5**	21.5	19	**45.0**	42.2	35.6	**58.1**	49.2	50.7
Random	0.106	82.9	81.0	79.0	73	5.6	4.7	2.7	14.4	11.7	6.9

Πίνακας 4.1: Πειραματικά αποτελέσματα για διάφορα μοντέλα

τέλο με Mid Fusion (Topic 0.5, Audio 0.3, Music 0.2) υπερέχουν των υπολοίπων συνολικά, το μεν πρώτο με καλύτερα MRL, MRS και το δε δεύτερο με καλύτερες σειρές κατάταξης των εισηγήσεων. Παρατηρώντας, όμως ότι η βελτίωση ως προς τα MRL, MRS, είναι σχετικά μικρή θα προτιμήσουμε ως βέλτιστο μοντέλο το Mid Fusion (Topic 0.5, Audio 0.3, Music 0.2).

Επίσης, να τονιστεί ότι ο παραπάνω πίνακας έχει παραχθεί με πλήθος εισηγήσεων για κάθε ταινία, όσες και οι εισηγήσεις του topic μοντέλου. Δηλαδή, βλέπουμε πόσες εισηγήσεις κάνει το topic μοντέλο σε κάθε ταινία και τόσες κάνουν και τα υπόλοιπα μοντέλα. Αυτό το κάναμε για να είναι πιο συγκεκριμένη η σύγκριση μεταξύ των μοντέλων. Το μέσο πλήθος εισηγήσεων είναι ≈ 3 ανά ταινία.

Στη συνέχεια, θα παρουσιάσουμε ένα γράφημα(**Σχήμα 4.6**) σχετικά με τις επιδόσεις των μοντέλων Topic, tf-idf, μαζί με τα Ground Truth, Random για σύγκριση, στο οποίο θα απεικονίζεται το *ranking score*, κάθε μοντέλου, ανά ταινία. Όπως μπορεί να παρατηρήσει

κανείς, σε γενικές γραμμές πράγματι το topic μοντέλο είναι καλύτερο από το tf-idf μοντέλο. Υπάρχουν μερικές ταινίες που έχει καλύτερο σκορ το tf-idf μοντέλο και αυτές είναι που δεν έχουν μοντελοποιηθεί αρκετά καλά από το σύστημά μας. Ακόμα, υπάρχουν συγκεκριμένες ταινίες που όλα τα μοντέλα δεν τα πηγαίνουν τόσο καλά (όπως το *I'm Still Here (2010)*), καθώς είναι πιο ιδιαίτερες σαν ταινίες. Επιπρόσθετα, μπορεί να δει κανείς ότι το tf-idf τα πάει καλά, όπως και το topic βέβαια, σε σίκουελ ταινιών, καθώς μπορεί εύκολα να εντοπίσει κοινές λέξεις, ιδιαίτερες για τις συγκεκριμένες ταινίες όπως ονόματα(Legolas στις ταινίες The Lord of the Rings), αντικείμενα(lighsaber στις ταινίες Star Wars) κ.ο.κ, στα κείμενα του σεναρίου τους. Από την άλλη το topic μοντέλο τα πάει καλύτερα συγκριτικά σε θεματικές ταινίες που μπορεί να μην αποτελούν μέρος σειράς αλλά έχουν κοντινά θεματικές ταινίες στη βάση, όπως το Goodfellas (1990).

Σχήμα 4.6: Ranking Scores ανά ταινία για τα μοντέλα Topic, Tf-idf, Ground Truth, Random

Συνεχίζοντας, θα παρουσιάσουμε μερικά παραδείγματα εισηγήσεων, που επιδεικνύουν τη διαφοροποίηση στις εισηγήσεις που προσφέρει μία θεματική αναπαράσταση των ταινιών, καθώς και τη βελτίωση που παρουσιάζει το topic μοντέλο με τη σύντηξη με τα audio, music μοντέλα. Αρχικά, για το topic μοντέλο μερικές από τις πιο ιδιαίτερες εισηγήσεις είναι οι εξής:

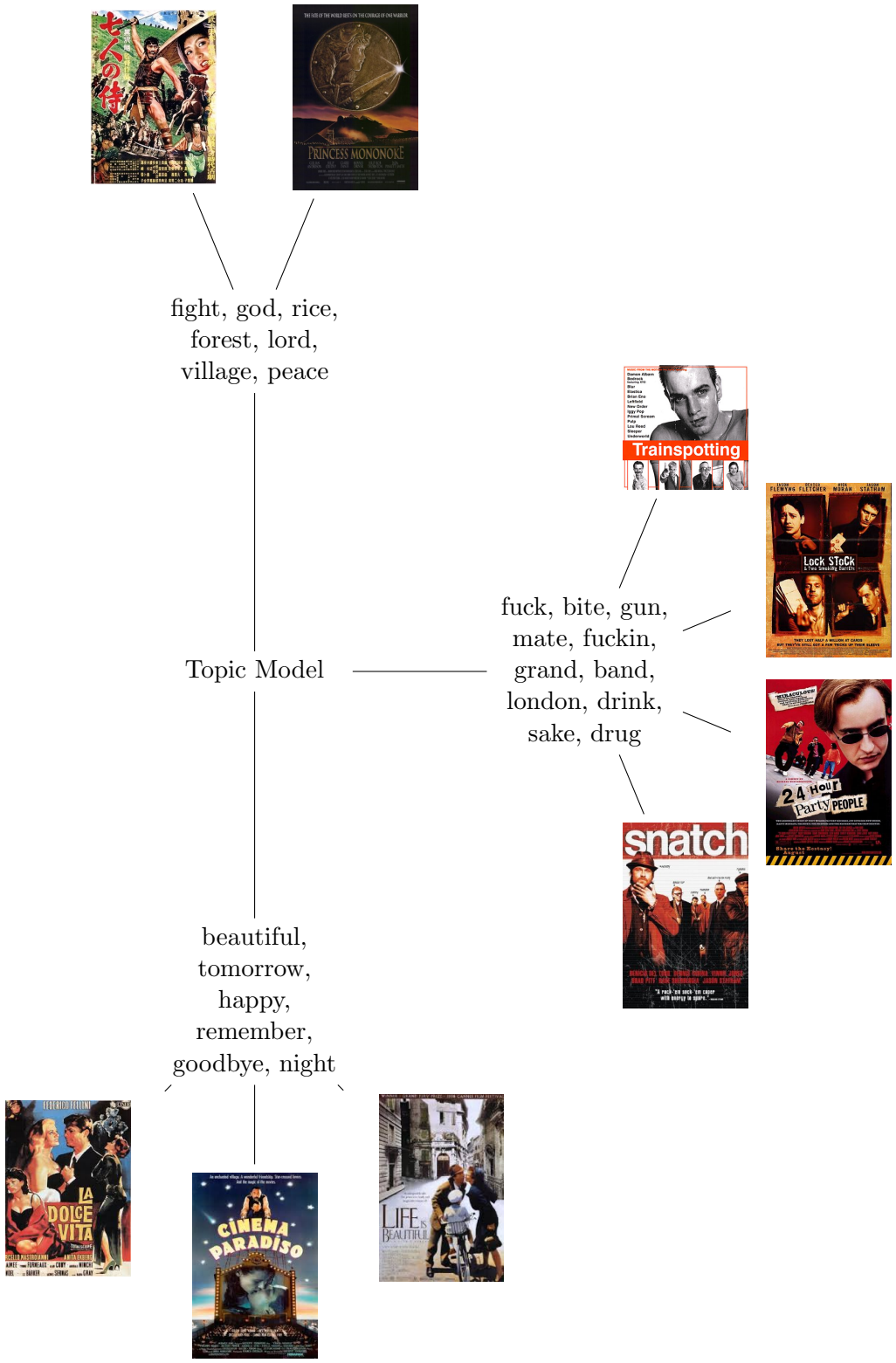
1. Για το *Princess Mononoke*, το σύστημα εισηγείται την ταινία *Seven Samurai*, έχοντας πιάσει το θέμα της αγροτικής κοινωνίας στην Ιαπωνία, με πολεμική χροιά στη ταινία. Συγκεκριμένα, το topic το οποίο μοιράζονται σε μεγάλο βαθμό περιέχει λέξεις όπως : fight, god, rice, forest, lord, village, peace κ.α. Είναι ένα ζευγάρι ταινιών που τα άλλα συστήματα συστάσεων, όπως το tf-idf ή κάποιο βασισμένο σε μεταδεδομένα ή απόψεις χρηστών, δεν το βρίσκει εύκολα. Αν κάποιος όμως, βλέποντας μία από αυτές τις δύο ταινίες ήθελε κάτι στη παραπάνω θεματολογία, το σύστημα μας του δίνει τη δυνατότητα να εξερευνήσει ποιες άλλες ταινίες έχουν σχέση με το συγκεκριμένο θέμα.

2. Για το *Lock, Stock & Two Smoking Barrels*, έχουμε ως εισηγήσεις τις ταινίες : *Trainspotting, 24 Hour Party People, Snatch*, οι οποίες αντιπροσωπεύουν την undergound-παράνομη σκηνή, σε pub στην Αγγλία, κατά βάση στου Λονδίνο. Γι'αυτό και το κοινό τους θέμα, επικεντρώνεται σε βρισιές οι οποίες είναι κομμάτι αναπόσπαστο των διαλόγων και του ύφους των ταινιών, όπως : fuck, bite, gun, mate, fuckin, grand, band, london, drink, sake, drug.

3. Για το *Cinema Paradiso*, οι πρώτες προτάσεις είναι *La Dolce Vita, La vita è bella.* Ταινίες που δεν ταιριάζουν έντονα σε σενάριο, αλλά είναι δραματικές ταινίες Ιταλικής παραγωγής με ελαφριά κωμωδία και παρόμοιο νοσταλγικό ύφος. Αυτό αποτυπώνεται και στο κοινό τους θέμα με λέξεις όπως : beautiful, tomorrow, happy, remember, goodbye, night

Τα παραπάνω φαίνονται και σχηματικά στο **Σχήμα 4.7**.

Στη συνέχεια θα αναφέρουμε μερικές βελτιώσεις που παρατηρούμε στις εισηγήσεις του απλού topic μοντέλου με την σύντηξη και των audio ή/και music μοντέλων. Αν και βλέπουμε διαφορές σε πολλές ταινίες, θα αναφέρουμε αυτές που επιδεικνύουν την επίδραση που έχει η σύντηξη των μοντέλων. Αυτές είναι ταινίες, που είτε δεν είχαν μοντελοποιηθεί πολύ σωστά ή είναι δύσκολο το topic μοντέλο να βρει σχετικές ταινίες με αυτές καθώς δεν ταιριάζουν από άποψη κειμένου. Αυτές είναι:



Σχήμα 4.7: Σχηματικό διάγραμμα συνάφειας ταινιών με βάση συγκεκριμένα θέματα(topics)

1. Πρώτα θα αναφερθούμε στο *Lion King*, το οποίο όντας animation είναι ιδιαίτερα δύσκολη ταινία να της γίνουν καλές προτάσεις από ένα topic μοντέλο σύστημα. Αυτό συμβαίνει γιατί, αυτό που κατά βάση δίνει τη ταυτότητα σε ένα animation film είναι η πολύ ζωντανή και ζωηρά χρωματισμένη εικόνα και ο ιδιαίτερος ήχος του. Είναι πολύ δύσκολο για ένα σύστημα αναπαράστασης γνώσης να αντιληφθεί μονάχα από τους διαλόγους ότι πρόκειται για ταινία animation και να προτείνει αντίστοιχα τέτοιες ταινίες. Γι'αυτό και το topic μοντέλο εισηγείται την ταινία *The Wizard of Oz*, το οποίο στη βάση αλήθειας είναι η 4η πιο σχετική ταινία. Η συνάφεια προκύπτει κυρίως από την ύπαρξη του λιονταριού και στης δύο ταινίες που υπάρχει στους διαλόγους. Το μοντέλο όμως με Mid Fusion του Topic και του Audio μοντέλου, με ποσοστά συμμετοχής 50%, έχει ως τρίτη εισήγηση τη ταινία Nemo, το οποίο είναι και αυτό animated και αποτελεί τη 2η πιο σχετική ταινία σύμφωνα με τη βάση αλήθειας.

2. Άλλο παράδειγμα αποτελεί η ταινία *Pulp Fiction*. Για αυτή, σύμφωνα με τη βάση αλήθειας η πιο σχετική ταινία είναι το *Reservoir Dogs*. Αυτό συμβαίνει όχι γιατί αυτές οι δύο ταινίες μοιάζουν πολύ ως σενάριο, απλά αποτελούν δύο από τις πιο γνωστές ταινίες του *Quentin Tarantino*, ο οποίος έχει αρκετά ιδιαίτερο κινηματογραφικό ύφος. Αυτό είναι μία άλλη περίπτωση στην οποία αδυνατεί το topic μοντέλο να εντοπίσει ομοιότητες, λόγω μη δυνατότητας αναγνώρισης τέτοιων μέσω των διαλόγων. Συγκεκριμένα, το topic μοντέλο μόνο του μας προτείνει τη ταινία *Taxi Driver*. η οποία είναι η 12η πιο σχετική ταινία. Με το Mid Fusion μοντέλο ,με συμμετοχές Topic 50%, Audio 30% και Music 20%, έχουμε πλέον ως δεύτερη εισήγηση πράγματι το *Reservoir Dogs*. Σε αυτή τη περίπτωση οι audio, music ομοιότητες των δύο οδήγησαν σε καλύτερη εισήγηση το σύστημα μας.

3. Τέλος, θα αναφέρουμε και σαν παράδειγμα μία ταινία που απλώς απέτυχε το απλό topic σύστημα να τη μοντελοποιήσει σωστά. Είναι η ταινία *Space Odyssey*, για την οποία η μοναδική εισήγηση του μοντέλου είναι η *Reservoir Dogs*, η οποία προφανώς και δεν είναι σχετική, όπως φαίνεται και από τη σειρά κατάταξής της ως 89η πιο συναφής. Στη συγκεκριμένη περίπτωση, δεν έχει γίνει καλή αναπαράσταση της ταινίας στο topic space, οπότε δεν είναι δυνατό να βρεθεί συναφή ταινία μόνο με το topic μοντέλο. Χρησιμοποιώντας όμως το προηγούμενο Mid Fusion μοντέλο παίρνουμε μία λίστα εισηγήσεων, με όχι όλες τις εισηγήσεις απαραίτητα καλές, αλλά η 1η(*Blade Runner*) και η 4η(*Alien*) εισήγηση να είναι αντίστοιχα η 1η και η 2η πιο σχετική στη βάση αλήθειας. Άρα βλέπουμε

ότι η πληροφορία από τα audio, music κομμάτια βοήθησε το σύστημά μας να προσφέρει σωστές εισηγήσεις σε μία ταινία που topic μοντέλο δεν έφερε είχε "χάσει" κατά τη μοντελοποίηση.

Από τα παραπάνω φαίνεται η μεγάλη χρησιμότητα του να είναι βασισμένο το σύστημά μας σε πολλά κανάλια πληροφορίας. Αλληλοκαλύπτουν τις ελλείψεις του καθενός και βελτιώνουν τη συνολική απόδοση του συστήματος. Επίσης, δίνουν πολύ μεγάλο βάθος και διαφορετικότητα στις προτάσεις που μπορεί να γίνουν από το εισηγητικό μας σύστημα. Σχετικά με αυτό, προσδίδουν και τη δυνατότητα εξερεύνησης των ταινιών από πολύ ευρεία σκοπιά πλέον, καθώς βλέποντας κάποιος χρήστης μία ταινία, θα μπορεί να ζητήσει εισηγήσεις, επιλέγοντας ως προς ποιο modality τον ενδιαφέρει περισσότερο. Αν δηλαδή του άρεσε πολύ η μουσική της ταινίας και η ατμόσφαιρα, θα μπορέσει να πάρει πληροφορίες από ένα σύστημα που δίνει έμφαση στα κανάλια audio,music ενώ θα του δίνεται και η δυνατότητα να περιηγηθεί και στις συναφείς ταινίες της βάσης ως προς αυτά τα κανάλια πληροφορίας , και όχι μόνο να κάνει θεματική περιήγηση ή να βλέπει εισηγήσεις μόνο βασισμένες στους διαλόγους της ταινίας.

Κεφάλαιο 5

Συμπεράσματα και Μελλοντική Εργασία

5.1 Συμπεράσματα

Η παρούσα διπλωματική εργασία αφορά τη δημιουργία ενός εισηγητικού συστήματος για ταινίες, βασισμένο κυρίως στο θεματικό περιεχόμενο των υποτίτλων και συνεπικουρούμενο από ηχητικές πληροφορίες. Δημιουργήθηκε ένα μοντέλο επεξεργασίας των ταινιών, το οποίο μέσω διαφόρων διαδικασιών είναι ικανό να αναπαραστήσει τις ταινίες στο χώρο θεμάτων και να βρει ομοιότητες μεταξύ τους. Επιπροσθέτως, είναι σε θέση να εξάγει πληροφορίες σχετικά με το είδος του μουσικού περιεχομένου και των διαφόρων ηχητικών συμβάντων που παρουσιάζονται στο ηχητικό σκέλος των ταινιών. Αυτά τα διαφορετικά κανάλια πληροφορίας συντήχθηκαν τελικώς για να μας δώσουν νέες εισηγήσεις ομοιότητας ανά ταινία, βασισμένες στην σύνθεση των διαφορετικών πηγών συνάφειας που αυτά αντιπροσωπεύουν.

Πιο συγκεκριμένα τα αποτελέσματα στα οποία καταλήξαμε είναι τα ακόλουθα:

- Αποδείξαμε ότι μία topic modeling προσέγγιση για τους υπότιτλους των ταινιών μας δίνει καλύτερα αποτελέσματα σχετικά με την ομοιότητα μεταξύ διαφόρων ταινιών, απ'ότι μία απλή tf-idf προσέγγιση. Αυτό φαίνεται τόσο στη βελτίωση των διαφόρων μετρικών που χρησιμοποιήσαμε, όσο και στις πιο ποιοτικές εισηγήσεις που παρείχε το θεματικό μοντέλο. Το τελευταίο γίνεται φανερό ειδικά σε ταινίες που δεν χρησιμοποιούν το ίδιο ακριβώς λεξιλόγιο, αλλά είναι θεματικά κοντά και το topic model μπορεί και βρίσκει αυτή τη συνάφεια, εν αντιθέσει με το tf-idf μοντέλο.

- Η παρουσία δεδομένων ομοιότητας και από άλλα κανάλια πληροφορίας, όπως ο ήχος ή η μουσική, οδηγεί σε βελτίωση των αποτελεσμάτων. Αυτό συμβαίνει σε κάποιες early fusion προσεγγίσεις και πιο συγκεκριμένα τα καλύτερα αποτελέσματα προκύπτουν για την topic,audio σύντηξη. Παρατηρείται, σύμφωνα και με το πίνακα αποτελεσμάτων, βελτίωση ως προς το απλό topic μοντέλο. Αντίστοιχα, βλέπουμε και βελτιστοποίηση για κάποια σχήματα mid fusion των πινάκων ομοιότητας. Βλέπουμε αύξηση των διαφόρων μετρικών με διαφορετικά βάρη σύντηξης, με συνολικά καλύτερο όμως απ'όλα τα δοκιμασμένα μοντέλα σύντηξης το mid fusion μοντέλο και με τα τρία modalities, audio, music, topic, με βαθμό συμμετοχής $0.3, 0.2, 0.5$ αντίστοιχα.

- Υπάρχουν επίσης σαφή παραδείγματα ταινιών, που το απλό topic μοντέλο δεν επιστρέφει καλές εισηγήσεις, ενώ με τη σύντηξη αυτού με μοντέλα που βασίζονται στον ήχο της ταινίας, ανανεώνονται οι εισηγήσεις σε πολύ βελτιωμένες. Αυτό μας δείχνει ότι μπορεί να υπάρχει ομοιότητα μεταξύ των ταινιών που δεν εκφράζεται στο χώρο των θεμάτων, αλλά αναπαρίσταται στα ηχητικά σκέλη των ταινιών, ωθώντας μας σε διάφορα σχήματα σύντηξης των καναλιών πληροφορίας για τη παραγωγή καλύτερων εισηγήσεων.

Πέρα όμως από τα παραπάνω μετρήσιμα αποτελέσματα, πιστεύουμε ότι η καινοτόμα ιδέα που αναπτύχθηκε στα πλαίσια αυτής της διπλωματικής, είναι ο τρόπος αναπαράστασης των ταινιών στο θεματικό χώρο με βάση το περιεχόμενο των υποτίτλων τους. Αυτό μας δίνει τη δυνατότητα για θεματική περιήγηση των ταινιών και επιλογή νέων για παρακολούθηση, όχι με γνώμονα κάποια μεταδεδομένα όπως στις συνήθεις εφαρμογές, αλλά μέσω πλοήγησής μας στα θέματα από τα οποία απαρτίζεται η κάθε ταινία. Έτσι μας προσφέρονται νέοι τρόποι για αναπαράσταση της ομοιότητας μεταξύ των ταινιών, νέες κατευθύνσεις για εύρεση ομοιότητας κατ'ευθείαν από το περιεχόμενο των ταινιών και ακολούθως διαφορετικών εισηγήσεων από τις συνηθισμένες.

5.2 Μελλοντική Εργασία

Όπως αναφέρθηκε, η παρούσα εργασία αποτελεί μία πρωτότυπη ιδέα στο χώρο των εισηγητικών συστημάτων για ταινίες. Εντούτοις, τα περιθώρια διερεύνησης νέων μεθόδων με σκοπό τη βελτίωση όσων παρουσιάστηκαν εδώ είναι μεγάλα. Αρχικά, μία πρώτη προσέγγιση θα ήταν να αυξήσουμε τη βάση δεδομένων μας και με άλλες ταινίες, αυξάνοντας έτσι τα δεδομένα εκπαίδευσης του μοντέλου και οδηγώντας σε πιο συνεκτικά θέματα, μεγαλύτερη θεματική κάλυψη

του χώρου ταινιών και καλύτερες δυνατότητες γενίκευσης. Παράλληλα, θα μπορούσαμε να εφαρμόσουμε και μία online μορφή της LDA[43], έτσι ώστε να είναι δυνατή η άμεση ενημέρωση του μοντέλου μας με νέες ταινίες για να γίνει πιο εύκολα επεκτάσιμο το σύστημά μας. Άλλη διαφοροποίηση μπορεί να επέλθει και πάνω στο είδος του μοντέλου, βασιζόμενοι σε ιδέες που αναφέραμε και προηγουμένως(**3.2.3**), όπως Correlated Topic Models που επιτρέπουν τη συσχέτιση μεταξύ των θεμάτων και οδηγούν σε πιο συγκεντρωμένες αναπαραστάσεις ή Hierarchical Dirichlet Processes που δεν χρειάζονται να οριστεί a priori ένας αριθμός θεμάτων, αλλά εντοπίζεται ο βέλτιστος κατά το στάδιο της εκπαίδευσης.

Όσον αφορά τώρα το είδος των δεδομένων, η πιο εμφανής, αλλά αρκετά δύσκολη, επέκταση τους συστήματος είναι η ανάλυση της πληροφορίας και από το οπτικό περιεχόμενο της ταινίας. Στόχος μας είναι μέσα από την εξαγωγή και επεξεργασία low-level χαρακτηριστικών των ταινιών να μπορέσουμε να οδηγηθούμε σε πιο υψηλού επιπέδου χαρακτηριστικά, που θα προσφέρουν νέες αναπαραστάσεις των ταινιών. Δηλαδή, θα μπορούν να μοντελοποιούν καλλιτεχνικές ιδιαιτερότητες για κάθε ταινία, όπως το ιδιαίτερο σκηνοθετικό ύφος κ.α. Προφανώς μία τέτοια προσέγγιση θα έδινε πολύ περισσότερο βάθος στο σύστημά μας, καθώς τα οπτικά χαρακτηριστικά μίας ταινίας είναι πάρα πολύ πλούσια σε πληροφορία σχετικά με τη μορφή της ταινίας και σίγουρα θα μας οδηγήσει σε καλύτερες εισηγήσεις.

Επίσης, θα μπορούσαμε να κατασκευάσουμε ένα υβριδικό μοντέλο, με το ένα σκέλος το ήδη υπάρχον σύστημα και το άλλο σκέλος ένα collaborative filtering σύστημα, το οποίο θα περιέχει τις γνώμες χρηστών και βαθμολογίες προτιμήσεων αυτών. Έτσι, συμπεριλαμβάνοντας τον ανθρώπινο παράγοντα, σίγουρα θα πετύχουμε και πιο πρακτικές εισηγήσεις, που θα ακολουθούν τις προτιμήσεις των χρηστών. Πέρα όμως από αυτό, μία πολύ σημαντική συνεισφορά ενός τέτοιου μοντέλου θα είναι στο πλαίσιο της ανακάλυψης γνώσης, δηλαδή τη συσχέτιση μεταξύ του περιεχομένου των ταινιών και τις προτιμήσεις των χρηστών. Αυτό οδηγεί σε εύρεση διασυνδέσεων, οι οποίες δεν είναι εμφανείς από πριν, πράγμα που δίνει ώθηση σε προσωποποιημένες εισηγήσεις για κάθε χρήστη και ομαδοποίηση των χρηστών ανάλογα με τις προτιμήσεις τους ως προς το περιεχόμενο. Τέλος, στο πλαίσιο του υβριδικού μοντέλου πάλι, θα μπορούσαμε αντί να συνδυάσουμε το υπάρχον με ένα collaborative filtering, να το συνδυάσουμε και με ένα content based μοντέλο, όπου ως περιεχόμενο εννοούμε τα μεταδεδομένα που αναφέρονται στην ταινία. Για παράδειγμα, θα μπορούσε κάθε ταινία να συνδυάζεται και με ένα πλήθος από ετικέτες σχετικά με σκηνοθέτη, είδος κ.α. ή να παρέχεται και η περίληψη της

ταινίας στο σύστημά μας και με κάποια ανάλυση πάνω σε αυτές τις περιλήψεις, όπως η LDA μέθοδος ή ο tf-idf μετασχηματισμός, να βρίσκουμε την ομοιότητα μεταξύ των ταινιών και ως προς αυτές τις πληροφορίες.

Επιπροσθέτως, στο πλαίσιο της εργασίας μας δεν δώσαμε ιδιαίτερη έμφαση στο τρόπο σύντηξης των καναλιών πληροφορίας. Υπάρχουν πολλοί διαφορετικοί τρόποι για να γίνει η σύντηξη των καναλιών και εμείς επιλέξαμε δύο γρήγορα υλοποιήσιμους τρόπους. Θα μπορούσε να διερευνηθεί η ιδέα κάθε κανάλι πληροφορίας να εκφράζει για κάθε query ένα confidence level, δηλαδή ένα βαθμό βεβαιότητας για το πόσο σωστές πιστεύει ότι είναι οι εισηγήσεις του και αναλόγως με αυτό το βαθμό να αποφασίζαμε ποιες ταινίες να προτείνουμε τελικά στο χρήστη. Ομοίως, θα μπορούσε κατά περιπτώσεις ένα κανάλι πληροφορίας να τεθεί εκτός λειτουργίας, για παράδειγμα λόγω μη καλής μοντελοποίησης της ταινίας στο συγκεκριμένο χώρο, οπότε οι εισηγήσεις να προέκυπταν από τα υπόλοιπα. Μια διαφορετική προσέγγιση, θα ήταν να γίνει πραγματικά early fusion , όπως αυτή ορίζεται και στο (**3.5**), για τα διάφορα χαρακτηριστικά των καναλιών πληροφορίας. Δηλαδή, να φτιάχναμε ένα διάνυσμα το οποίο θα αποτελείται από όλα τα διαφορετικά χαρακτηριστικά των διαφορετικών αναπαραστάσεων, εδώ δηλαδή τις topic αναπαραστάσεις μαζί με τα ποσοστά των κλάσεων από music & audio events από το ηχητικό κομμάτι, και αυτό το διάνυσμα κάθε ταινίας να προωθείται σε έναν metaclassifier, ο οποίος θα παράγει κάποια ταξινόμηση αυτών, για παράδειγμα σε είδη. Στη συνέχεια έχοντας τις αναπαραστάσεις των ταινιών σε αυτό το fused feature space, μπορούμε να βρούμε πάλι τις ομοιότητες μεταξύ τους σύμφωνα με τα αποτελέσματα του metaclassifier για κάθε ταινία.

Να τονιστεί επίσης, ότι άμεσος στόχος μετά την εργασία αυτή, είναι η υλοποίηση ενός topic browser-recommender system για της ταινίες. Δηλαδή, θα δίνεται η δυνατότητα στο χρήστη όταν θα του παρουσιάζονται οι εισηγήσεις των ταινιών, να βλέπει ποια θέματα ενεργοποιούνται περισσότερο για κάθε ταινία που του προτείνεται και να παρατηρεί τα θέματα αυτά, αλλά και να μπορεί να πλοηγηθεί σε όλες τις ταινίες που έχουν σχέση με το εκάστοτε θέμα. Επομένως, θα προσφέρεται τόσο η αναζήτηση για ταινίες μέσω των θεμάτων, όσο και η εισήγηση ταινιών για κάθε ταινία με παράλληλη παρουσίαση των θεμάτων που οδήγησαν σε αυτές τις εισηγήσεις, καθώς και των υπολοίπων ταινιών που σχετίζονται έντονα με τα συγκεκριμένα θέματα. Έτσι θα προσφέρεται στο χρήστη η δυνατότητα πλοήγησης στο θεματικό χώρο που προέκυψε από όλες τις ταινίες και η εξερεύνηση της ομοιότητας των ταινιών, με ενδιάμεσο βήμα τα θέματα που τις αποτελούν.

Ολοκληρώνοντας, ως τελικό στόχο θα μπορούσε να φανταστεί κανείς ένα εισηγητικό σύστημα για ταινίες που περιέχει πληροφορίες από όλες τις δυνατές κατευθύνσεις. Δηλαδή, και user ratings και μεταδεδομένα για κάθε ταινία και πληροφορία από το περιεχόμενο αυτό καθ'εαυτό, από το κείμενο, τον ήχο και την εικόνα της ταινίας. Αυτό θα μας επιτρέψει την αναπαράσταση κάθε ταινίας ως μία οντολογία που περιέχει πλήθος πληροφορίες, δίνοντάς μας τη δυνατότητα να προσφέρουμε εισηγήσεις είτε βασισμένοι εξ'ολοκλήρου σε όλες αυτές τις διαφορετικές πηγές πληροφορίας είτε εν γνώση του χρήστη να προτιμηθεί μία από αυτές. Αν για παράδειγμα, ο χρήστης δει μία ταινία και του αρέσει η σκηνοθεσία θα είναι σε θέση να επιλέγει εισηγήσεις με βάση το σκηνοθέτη, αν του αρέσει η μουσική της ταινίας. τότε με βάση μόνο το είδος της μουσικής κ.ο.κ. Αυτό θα μας οδηγήσει σε μία πολύ διαφορετική κατανόηση των ταινιών και θα μας επιτρέψει να περιηγηθούμε στο χώρο των ταινιών από διαφορετικές οπτικές γωνίες, ανακαλύπτοντας ομοιότητες που δεν μπορούμε να εκφράσουμε στη παρούσα φάση, με τα συμβατικά συστήματα εισηγήσεων.

Βιβλιογραφία

[1] *Making Personalised Flight Recommendations using Implicit Feedback.* Διδακτορική Διατριβή , 2004.

[2] G. Adomavicius και A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 2005.

[3] Alekh Agarwal, Olivier Chapelle, Miroslav Dudík και John Langford. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15(1):1111–1133, 2014.

[4] Davide Albanese, Roberto Visintainer, Stefano Merler, Samantha Riccadonna, Giuseppe Jurman και Cesare Furlanello. mlpy: Machine learning python, 2012.

[5] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik και Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379, 2010.

[6] Pradeep K. Atrey, Mohan S. Kankanhalli και John B. Oommen. Goal-oriented optimal subset selection of correlated multimedia streams. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1), 2007.

[7] N. Babaguchi, Y. Kawai και T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. *Multimedia, IEEE Transactions on*, 4(1):68–75, 2002.

[8] Marko Balabanovic και Yoav Shoham. Fab: Content-based, collaborative recommendation. τόμος 40, σελίδες 66–72, 1997.

[9] James Bergstra και Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

[10] Suvir Bhargav. Efficient features for movie recommendation systems. Διπλωματική εργασία Master, KTH, Communication Theory, 2014.

[11] D. Blei και J. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1:17–35., 2007.

[12] D. M. Blei, T. L. Griffiths, M. I. Jordan και J. B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. Στο *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[13] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[14] David M. Blei και John D. Lafferty. Correlated topic models. Στο *NIPS*, σελίδες 147–154, 2005.

[15] D.M. Blei, A.Y. Ng και M.I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[16] Jordan L. Boyd-Graber, David M. Mimno και David Newman. Care and feeding of topic models. Στο *Handbook of Mixed Membership Models and Their Applications*Edoardo M. Airoldi, David M. Blei, Elena A. Erosheva και Stephen E. Fienberg, επιμελητές, σελίδες 225–254. Chapman and Hall/CRC, 2014.

[17] Robin Burke. Hybrid web recommender systems. Στο *The Adaptive Web*Peter Brusilovsky, Alfred Kobsa και Wolfgang Nejdl, επιμελητές, τόμος 4321 στο *Lecture Notes in Computer Science*. Springer, 2007.

[18] George Casella και Edward I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[19] George Casella και Christian P. Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

[20] Jonathan Chang, Jordan L Boyd-Graber, Sean Gerrish, Chong Wang και David M Blei. Reading tea leaves: How humans interpret topic models. Στο *Neural Information Processing Systems*, τόμος 22, 2009.

[21] Corinna Cortes και Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

[22] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas και Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[23] James M. Dickey. Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, 78(383):628–637, 1983.

[24] Verena Dorner και Michael Scholz. Predicting and economically exploiting utility thresholds with utility-based recommendation systems. Στο *ECIS*, σελίδα 30, 2013.

[25] Kai Bo Duan και S. Sathiya Keerthi. Which is the best multiclass svm method? an empirical study. Στο *Proceedings of the 6th International Conference on Multiple Classifier Systems*, MCS'05, σελίδες 278–285, Berlin, Heidelberg, 2005. Springer-Verlag.

[26] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas και Yannis S. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568, 2013.

[27] Eyrun A Eyjolfsdottir, Gaurangi Tilak και Nan Li. Moviegen: A movie recommendation system. *UC Santa Barbara: Technical Report*, 2010.

[28] Ronald Fagin, Ravi Kumar και D. Sivakumar. Comparing top k lists. Στο *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, σελίδες 28–36, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

[29] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[30] S. Geman και D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Recognition*, 6:721–741, 1984.

[31] Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou και Sergios Theodoridis. Violence content classification using audio features. Στο *Proceedings of the 4th Helenic Conference on Advances in Artificial Intelligence*, SETN'06, σελίδες 502–507, Berlin, Heidelberg, 2006. Springer-Verlag.

[32] Theodoros Giannakopoulos, Alexandros Makris, Dimitrios Kosmopoulos, Stavros Perantonis και Sergios Theodoridis. Audio-visual fusion for detecting violent scenes in videos. Στο *Artificial Intelligence: Theories, Models and Applications*, σελίδες 91–100. Springer Berlin Heidelberg, 2010.

[33] W.R. Gilks, S. Richardson και D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995.

[34] Jennifer Golbeck. Generating predictive movie recommendations from trust in social networks. Στο *iTrust*Ketil Stølen, William H. Winsborough, Fabio Martinelli και Fabio Massacci, επιμελητές, τόμος 3986 στο *Lecture Notes in Computer Science*, σελίδες 93–104. Springer, 2006.

[35] David Goldberg, David Nichols, Brian M. Oki και Douglas Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

[36] Michael D. Gordon και Manfred Kochen. Recall-precision trade-off: A derivation. *JASIS*, 40(3):145–151, 1989.

[37] Derek Greene, Derek O'Callaghan και Pádraig Cunningham. How many topics? stability analysis for topic models. *CoRR*, 2014.

[38] T. L. Griffiths και M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Συππλ. 1):5228–5235, 2004.

[39] Tom Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. Τεχνική Αναφορά υπ. αρίθμ., 2002.

[40] G. Heinrich. Parameter estimation for text analysis. *Web: http://www. arbylon. net/publications/text-est. pdf*, 2005.

[41] J. L. Herlocker, J. A. Konstan και J. T. Riedl. Explaining collaborative filtering recommendations. Στο *Computer Supported Cooperative Work*, σελίδες 241–250, 2000.

[42] Will Hill, Larry Stead, Mark Rosenstein και George Furnas. Recommending and evaluating choices in a virtual community of use. Στο *Conference on Human Factors in Computing Systems*, σελίδες 194–201, Denver, 1995. ACM.

[43] Matthew D. Hoffman, David M. Blei και Francis R. Bach. Online learning for latent dirichlet allocation. Στο *NIPS*John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel και Aron Culotta, επιμελητές, σελίδες 856–864. Curran Associates, Inc., 2010.

[44] Thomas Hofmann. Probabilistic latent semantic indexing. σελίδες 50–57, 1999.

[45] Diane J. Hu. Latent dirichlet allocation for text, images, and music, 2009.

[46] Pengfei Hu, Wenju Liu, Wei Jiang και Zhanlei Yang. Latent topic model based on gaussian-lda for audio retrieval. Στο *Pattern Recognition*Cheng Lin Liu, Changshui Zhang και Liang Wang, επιμελητές, τόμος 321 στο *Communications in Computer and Information Science*, σελίδες 556–563. Springer Berlin Heidelberg, 2012.

[47] Yuening Hu, Jordan L. Boyd-Graber, Brianna Satinoff και Alison Smith. Interactive topic modeling. *Machine Learning*, 95(3):423–469, 2014.

[48] G. Iyengar, H. J. Nock και C. Neti. Audio-visual synchrony for detection of monologues in video archives. Στο *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 2*, ICME '03, σελίδες 329–332, Washington, DC, USA, 2003. IEEE Computer Society.

[49] Xin Jin, Yanzan Zhou και Bamshad Mobasher. A maximum entropy web recommendation system: Combining collaborative and content features. Στο *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, σελίδες 612–617, New York, NY, USA, 2005. ACM.

[50] Joseph A. Konstan, Bradley N. Miller και *et al*. Grouplens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM*, 40(3):77–87, 1997.

[51] P Koutras, A Zlatintsi, E Iosif, A Katsamanis, P Maragos και A Potamianos. Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization. *acc. ICIP-2015*, χ.χ.

[52] Ravi Kumar και Sergei Vassilvitskii. Generalized distances between rankings. Στο *WWW*Michael Rappa, Paul Jones, Juliana Freire και Soumen Chakrabarti, επιμελητές, σελίδες 571–580. ACM, 2010.

[53] Taras Lehinevych, Nikolaos Kokkinis-Ntrenis, Giorgos Siantikos, Theodoros Giannakopoulos, Stasinos Konstantopoulos και others. Discovering similarities for content-based recommendation and browsing in multimedia collections. Στο *Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on*, σελίδες 237–243. IEEE, 2014.

[54] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707, 1966.

[55] Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Publishing Company, Incorporated, 2008.

[56] W. Li, D. Blei και A. McCallum. Nonparametric Bayes pachinko allocation. Στο *Proc. 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

[57] Andreas Lommatzsch, Benjamin Kille και Sahin Albayrak. A framework for learning and analyzing hybrid recommenders based on heterogeneous semantic data. Στο *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, σελίδες 137–140. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013.

[58] Edward Loper και Steven Bird. Nltk: The natural language toolkit. Στο *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, σελίδες 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[59] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4):309–317, 1957.

[60] David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002.

[61] Tariq Mahmood και Francesco Ricci. Improving recommender systems with adaptive conversational strategies. Στο *Hypertext*Ciro Cattuto, Giancarlo Ruffo και Filippo Menczer, επιμελητές, σελίδες 73–82. ACM, 2009.

[62] Arun Mahmoudi, Mohammad Reza. A new approach in designing the transportation path of urban buses using gis (a case study of district no. 10 of tehran). *Indian Streams Research Journal*, 3(6), 2013.

[63] Christopher D. Manning, Prabhakar Raghavan και Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[64] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. 2002.

[65] Tao Mei, Bo Yang, Xian Sheng Hua και Shipeng Li. Contextual video recommendation by multimodal relevance and user feedback. *ACM Trans. Inf. Syst.*, 29(2):10:1–10:24, 2011.

[66] S. Middleton, N. Shadbolt και D. De Roure. Ontological User Profiling in Recommender Systems. Στο *ACM Transactions on Information Systems*, τόμος 22, σελίδες 54–88, 2004.

[67] Thomas Minka και John Lafferty. Expectation-propagation for the generative aspect model. Στο *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, σελίδες 352–359, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[68] Thomas P. Minka. Estimating a dirichlet distribution. Τεχνική Αναφορά υπ. αρίθμ., 2000.

[69] Roberto Mirizzi, Tommaso Di Noia, Azzurra Ragone, Vito Claudio Ostuni και Eugenio Di Sciascio. Movie recommendation with dbpedia. Στο *IIR*Giambattista Amati, Claudio Carpineto και Giovanni Semeraro, επιμελητές, τόμος 835 στο *CEUR Workshop Proceedings*, σελίδες 101–112. CEUR-WS.org, 2012.

[70] Jeho Nam, Masoud Alghoniemy και Ahmed H. Tewfik. Audio-visual content-based violent scene characterization. Στο *ICIP (1)*, σελίδες 353–357, 1998.

[71] Radford M. Neal. Probabilistic inference using markov chain monte carlo methods. Τεχνική Αναφορά υπ. αρίθμ., Department of Computer Science, University of Toronto, 1993.

[72] R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, σελίδες 249–265, 2000.

[73] Ilya Nemenman, F. Shafee και William Bialek. Entropy and inference, revisited. Στο *NIPS*Thomas G. Dietterich, Suzanna Becker και Zoubin Ghahramani, επιμελητές, σελίδες 471–478. MIT Press, 2001.

[74] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García και Rahul Sukthankar. Violence detection in video using computer vision techniques. Στο *Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns - Volume Part II*, CAIP'11, σελίδες 332–339, Berlin, Heidelberg, 2011. Springer-Verlag.

[75] Jianjun Ni, Xiaoping Ma, Lizhong Xu και Jianying Wang. An image recognition method based on multiple bp neural networks fusion. Στο *Information Acquisition, 2004. Proceedings. International Conference on*, σελίδες 323–326, 2004.

[76] Michael Pazzani και Daniel Billsus. Content-based recommendation systems. Στο *The Adaptive Web*Peter Brusilovsky, Alfred Kobsa και Wolfgang Nejdl, επιμελητές, τόμος 4321 στο *Lecture Notes in Computer Science*. Springer, Berlin / Heidelberg, 2007.

[77] J. K. Pritchard, M. Stephens και P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–955, 2000.

[78] Misha Rabinovich και Yogesh Girdhar. Gaining insight into films via topic modeling & visualization. *Parsons Journal for Information Mapping*, 7(1), 2015.

[79] Radim Řehůřek και Petr Sojka. Software Framework for Topic Modelling with Large Corpora. Στο *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, σελίδες 45–50, Valletta, Malta, 2010. ELRA.

[80] Joseph Reisinger, Austin Waters, Bryan Silverthorn και Raymond J. Mooney. Spherical topic models. Στο *ICML*Johannes Fürnkranz και Thorsten Joachims, επιμελητές, σελίδες 903–910. Omnipress, 2010.

[81] Reede Ren, Hemant Misra και JoemonM. Jose. Semantic based adaptive movie summarisation. Στο *Advances in Multimedia Modeling*Susanne Boll, Qi Tian, Lei Zhang, Zili Zhang και Yi PingPhoebe Chen, επιμελητές, τόμος 5916 στο *Lecture Notes in Computer Science*, σελίδες 389–399. Springer Berlin Heidelberg, 2010.

[82] P. Resnick και H.R. Varian. Recommender Systems. *Communications of the ACM*, 40(3):56–58, 1997.

[83] Berthier Ribeiro-Neto και Ricardo Baeza-Yates. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[84] Baeza Yates Ricardo και Ribeiro Neto Berthier. *Modern Information Retrieval*. 1999.

[85] Francesco Ricci, Lior Rokach, Bracha Shapira και Paul B. Kantor. *Recommender systems handbook*. Springer, New York; London, 2011.

[86] G. Salton και M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[87] J. Ben Schafer, Joseph A. Konstan και John Riedl. *E-Commerce Recommendation Applications*, τόμος 5, σελίδες 115–153. Kluwer Academic Publishers, 2001.

[88] Ervin Sejdić, Igor Djurović και Jin Jiang. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digit. Signal Process.*, 19(1):153–183, 2009.

[89] Upendra Shardanand και Patti Maes. Social information filtering: Algorithms for automating "word of mouth". 1:210–217, 1995.

[90] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman και W. T. Freeman. Discovering object categories in image collections. Στο *Proceedings of the International Conference on Computer Vision*, 2005.

[91] Cees G. M. Snoek, Marcel Worring και Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. Στο *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, σελίδες 399–402, New York, NY, USA, 2005. ACM.

[92] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval:document retrieval systems. σελίδες 132–142. Taylor Graham Publishing, London, UK, UK, 1988.

[93] M. Steyvers και T. Griffiths. Probabilistic topic models. *Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum*, 2005.

[94] Jian Tang, Zhaoshi Meng, XuanLong Nguyen, Qiaozhu Mei και Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. Στο *ICML*, τόμος 32 στο *JMLR Proceedings*, σελίδες 190–198. JMLR.org, 2014.

[95] Y. W. Teh, D. Newman και M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. Στο *Advances in Neural Information Processing Systems*, τόμος 19, 2007.

[96] Y.W. Teh, M.I. Jordan, M.J. Beal και D.M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[97] Sergios Theodoridis και Konstantinos Koutroumbas. *Pattern Recognition, Fourth Edition*. Academic Press, 4τηη έκδοση, 2008.

[98] C. J.van Rijsbergen. *Information retrieval*. Butterworths, London, 2η έκδοση, 1979.

[99] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[100] Jesse Vig, Shilad Sen και John Riedl. The tag genome: Encoding community knowledge to support novel interaction. *TiiS*, 2(3):13, 2012.

[101] Hanna M. Wallach, David M. Mimno και Andrew McCallum. Rethinking lda: Why priors matter. Στο *NIPS*Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams και Aron Culotta, επιμελητές. Curran Associates, Inc., 2009.

[102] Chong Wang, John William Paisley και David M. Blei. Online variational inference for the hierarchical dirichlet process. Στο *AISTATS*Geoffrey J. Gordon, David B. Dunson και Miroslav Dudík, επιμελητές, τόμος 15 στο *JMLR Proceedings*, σελίδες 752–760. JMLR.org, 2011.

[103] Ge Wang. Hall david l, mcmullen sonya ah: Mathematical techniques in multisensor data fusion. 2nd ed. *BioMedical Engineering OnLine*, 4(1), 2005.

[104] Yi Wang. Distributed gibbs sampling of latent topic models: The gritty details. Τεχνική Αναφορά υπ. αρίθμ., 2008.

[105] William Webber, Alistair Moffat και Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20, 2010.

[106] Rong Yan. *Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval*. Διδακτορική Διατριβή , Pittsburgh, PA, USA, 2006.

[107] Xuchang Zou, Raffaella Settimi, Jane Clel και Chuan Duan. Thresholding strategy in requirements trace retrieval. Στο *CTI Research Symposium*, σελίδες 100–103, 2004.