Αριστοτελειο Πανεπιστημιο Θεσσαλονικης

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ ΕΞΕΙΔΙΚΕΥΣΗΣ

Ανίχνευση Λανθανουσών Παραμέτρων σε Στίχους Δημοτικών Τραγουδιών

Συγγραφέας:
Ιωάννη Ν. Μήτρο

Επιβλέπων:
Δρ. Κωνσταντίνος Κοτρόπουλος

Π.Μ.Σ Πληροφορικής - Κατεύθυνση Ψηφιακά Μέσα

ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2012-13

Με επιφύλαξη παντὸς δικαιώματος

Υπογραφή: _____

Ημερομηνία: _____

Περίληψη

Σκοπός της συγκεκριμένης εργασίας είναι η ανάδειξη και εποπτική αναπαράσταση λανθανουσών παραμέτρων σε στίχους δημοτικών τραγουδιών που συλλέχθηκαν από το διαδίκτυο και από κανάλια κοινωνικής δικτύωσης όπως το YouTube.

Αναλύθηκε το λυρικό μέρος μιας συλλογής 388 δημοτικών τραγουδιών με σκοπό την εξαγωγή όρων από κάθε κείμενο (τραγούδι), για την δημιουργία του πίνακα συχνότητας όρων-αντίστροφης συχνότητας κειμένων (tf-idf). Εν συνεχεία τα δεδομένα του πίνακα tf-idf προβλήθηκαν στον δισδιάστατο και τρισδιάστατο χώρο όπου και εφαρμόστηκε ομαδοποίηση κ-μέσων για την ανάδειξη ομάδων που μπορεί να προκύψουν βάση γεωγραφικής ή θεματικής προέλευσης των τραγουδιών.

Τέλος για την ανάδειξη λανθανουσών παραμέτρων διενεργήθηκαν ερωτήματα στο μετασχηματισμένο χώρο μειωμένης διάστασης χρησιμοποιώντας ως μεθοδολογία την LSA όπου και εφαρμόστηκε περαιτέρω εποπτική διερεύνηση των αποτελεσμάτων μέσω της ανάλυσης γραφημάτων. Συγκεκριμένα οι αλγόριθμοι γραφημάτων που χρησιμοποιήθηκαν ήταν ο Fruchterman-Reingold [Φρυςητερμαν ανδ Ρεινγολδ] και Force Atlas 2 [Θαςομψ ετ αλ.].

Στα γραφήματα που παρήχθησαν από τους παραπάνω αλγόριθμους εφαρμόστηκε ομαδοποίηση για την ανάδειξη των συχνότερα χρησιμοποιούμενων όρων σε μια ομάδα συμπεραίνοντας την θεματική περιοχή στην οποία ανήκουν τα τραγούδια καθώς και των σημαντικότερων κόμβων (τραγουδιών) που δρουν ως επίκεντρο μεταξύ των ομάδων.

Abstract

The objective of this project is to highlight and supervisory represent latent parameters in lyric folk songs collected from the web and from social activity channels such as YouTube.

The lyrical part of the collection of 388 folk songs has been analyzed in order to extract terms from each document (song) for the construction of the (tf-idf) matrix. Subsequently the tf-idf matrix data have been projected in two and three-dimensional space where k-means clustering was applied to highlight groups that may arise based on the geographical or thematic origin of the songs.

Finally for the emergence of latent parameters, queries were performed in the transformed space of reduced dimensionality using a methodology known as LSA where further supervisory analysis was applied on the obtained results using graph analysis theory. Specifically, the graph algorithms that were used were Fruchterman-Reingold [Fruchterman and Reingold] and Force Atlas 2 [Jacomy et al.].

In the graphs produced by the above algorithms clustering was applied to illustrate the most frequently used terms in a group infering this way the thematic area to which the songs belong to as well as the hubs which act as connecting links between groups.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες

IR	**I**nformation **R**etrieval	**Α**νάκτηση **Π**ληροφορίας
tf	**T**erm **F**requency	**Σ**υχνότητα **Ό**ρων
idf	**I**nverse **D**ocument **F**requency	**Α**ντίστροφη **Σ**υχνότητα **Κ**ειμένων
tf-idf	**T**erm **F**requency - **I**nverse **D**ocument **F**requency	
LSA	**L**atent **S**emantic **A**nalysis	**Α**νάλυση **Λ**ανθάνουσας **Σ**ημασιολογίας
pLSA	**P**robabilistic **L**atent **S**emantic **A**nalysis	**Π**ιθανοκρατική **Α**νάλ. **Λ**ανθάν. **Σ**ημασιολογίας
SVD	**S**ingular **V**alue **D**ecomposition	**Α**ποσύνθεση **Ι**διαζουσών **Τ**ιμών
VSM	**V**ector **S**pace **M**odel	**Μ**οντέλο **Δ**ιανυσματικού **Χ**ώρου
MDS	**M**ultidimensional **S**caling	**Π**ολυδιάσταση **Κ**λιμάκωση
PCA	**P**rincipal **C**component **A**nalysis	**Α**νάλυση **Π**ρωτευουσών **Σ**υνιστωσών
NMDS	**N**on-metric **M**ultidimensional **S**caling	**Μ**η **Π**αραμετρική **Π**ολυδιάσταση **Κ**λιμάκωση
CA	**C**orrespondence **A**anlysis	**Α**νάλυση **Α**ντιστοιχιών
DCA	**D**etrened **C**orrespondence **A**nalysis	
BCO	**B**ray **C**urtis **O**rdination	
RDA	**R**edundancy **A**nalysis	
EM	**E**xpectation **M**aximization	**Α**ναμενόμενη **Τ**ιμή-**Μ**εγιστοποίηση
WCSS	**W**ithin **C**luster **S**um of **S**quares	**Ά**θροισμα **Τ**ετραγώνων εντός **Ο**μάδων

Κεφάλαιο 1

Ανάλυση Στίχων Δημοτικών Τραγουδιών

1.1 Κίνητρο

Η ανάκτηση πληροφοριών (IR) [1] αφορά την εύρεση υλικού αδόμητης φύσης (συνήθως κείμενα) συναφούς προς ένα ερώτημα μέσα από μεγάλες συλλογές (αποθηκεύσιμες σε ηλεκτρονικούς υπολογιστές). Ζούμε σε μια εποχή όπου η πληροφορία παίζει έναν πολύ καθοριστικό παράγοντα στις ενέργειες κυβερνήσεων, επιχειρήσεων και ιδιωτών. Στο ίδιο εύρος, η τεχνολογία της πληροφορίας αυξάνεται με ραγδαίους ρυθμούς για να καλύψει τις ολοένα αυξανόμενες ανάγκες πληροφόρησης. Μέσω του διαδικτύου, παρατηρήθηκε πρωτοφανής αύξηση του αριθμού των ιστοσελίδων, αλλά και της δομής τους καθώς αναδύονται ολοένα και περισσότερες νέες τεχνολογίες. Στο μέλλον αναμένεται ακόμα μεγαλύτερη ανάπτυξη καθώς καθίσταται επιτακτική η ανάγκη το περιεχόμενο στο διαδίκτυο να καταστεί αναγνώσιμο από μηχανές, όπως έχει προταθεί και οραματιστεί στο σημασιολογικό ιστό, γεγονός το οποίο πλησιάζει πιο κοντά στην πραγματικότητα.

Σε πολλές περιοχές, η ανάκτηση εγγράφων έχει κυριολεκτικά βασιστεί στο ταίριασμα συναφών όρων που περιέχονται σε έγγραφα με τους αντίστοιχους όρους που βρίσκονται στα ερωτήματα. Ωστόσο, οι φυσικές γλώσσες θέτουν μεγαλύτερες προκλήσεις που έχουν αναπόφευκτα καταστήσει αυτές τις υπάρχουσες μεθόδους αντιστοίχισης συναφών όρων ανεπαρκείς. Αυτό προκύπτει κυρίως από το γεγονός ότι σε πολλές φυσικές γλώσσες μια λέξη μπορεί να έχει πολλαπλές σημασίες (πολυσημία) καθώς πολλές

λέξεις μιας φυσικής γλώσσας μπορούν να αποδίδουν την ίδια έννοια (συνωνυμία). Εξαιτίας λοιπόν αυτού του λόγο η λανθάνουσα σημασιολογική ανάλυση [4] υπόσχεται να ξεπεράσει τα προβλήματα της λεξιλογικής αντιστοίχισης με τη χρήση στατιστικών εννοιολογικών δεικτών, αντί μεμονωμένων λέξεων για την ανάκτηση. Όντως, η λανθάνουσα σημασιολογική ανάλυση αντιμετωπίζει αυτό το πρόβλημα με την αξιοποίηση των καθολικών σχέσεων μεταξύ των κειμένων και των όρων, με σκοπό τη δημιουργία ενός κοινού σημασιολογικού χώρου από τον οποίο μπορούν να δοθούν απαντήσεις στα ερωτήματα των χρηστών. Υποθέτει ότι υπάρχει κάποια σημασιολογική δομή που διέπει μια συλλογή κειμένων η οποία επισκιάζεται από την τυχαιότητα των λέξεων που απαντούν στα κείμενα αυτά. Οι στατιστικές τεχνικές χρησιμοποιούνται για την εκτίμηση αυτής της σημασιολογικής δομής και ακολούθως στην αντιστοίχιση σε έναν χώρο εγγύτητας με στόχο την καλύτερη ανάκτηση κειμένων στηριζόμενης στο σημασιολογικό τους περιεχόμενο σε αντίθεση με την ταύτιση των όρων αυτών καθεαυτών.

Η λανθάνουσα σημασιολογική ανάλυση φαίνεται να παρέχει ισχυρές προοπτικές εφαρμογής στις περιοχές φιλτραρίσματος ανεπιθύμητης ηλεκτρονικής αλληλογραφίας, σημασιολογικής ταξινόμησης εγγράφων, ομαδοποίηση κειμένων και ανάκτησης πληροφοριών, μεταξύ άλλων. Στην ανάκτηση πληροφοριών, τα κείμενα καθώς και οι ερωτήσεις προβάλλονται σε ένα κοινό χώρο εγγύτητας ή σημασιολογικό χώρο, όπου εφαρμόζονται μέτρα ομοιότητας για σκοπούς ανάκτησης και κατάταξης αποτελεσμάτων.

1.2 Αντικειμενικός Σκοπός της Εργασίας

Αφορά την ανάδειξη και εποπτική παρακολούθηση λανθανουσών παραμέτρων σε στίχους δημοτικών τραγουδιών χρησιμοποιώντας την ανάλυση λανθάνουσας σημασιολογίας ως μεθοδολογία ανάκτησης κειμένων μέσω ερωτημάτων.

i. Να αναδείξει θεματικές περιοχές στις οποίες κατηγοριοποιούνται τα τραγούδια καθώς και ποια είναι τα αντιπροσωπευτικότερα εξ' αυτών για μια θεματική περιοχή.

ii. Να αναλύσει τεχνικές εποπτικής αναπαράστασης συγγενών όρων και κειμένων οι οποίες προέρχονται από την θεωρία γραφημάτων και βοηθούν στην ευπαρουσίαστη εμφάνιση και ερμηνεία των αποτελεσμάτων που λαμβάνουμε στα ερωτήματα που τίθενται μέσω στην τεχνική LSA.

iii. Να χρησιμοποιήσει και να συγκρίνει αλγορίθμους ανάλυσης γραφημάτων με ευρέως γνωστά πακέτα λογισμικού ανάλυσης κοινωνικών δικτύων [νοδ, γεπ] που υπόσχονται ευελιξία και ταχύτητα.

1.3 Επισκόπηση

Κατά την προσπάθεια συλλογής περισσότερων πληροφοριών από κοινωνικά δίκτυα δημιουργήθηκαν αρχεία διαδικασιών scripts για να ελαφρύνουν τον φόρτο της όλης διαδικασίας. Κατά την εκτέλεση τους τα scripts αυτά ήταν υπεύθυνα για τη συλλογή πληροφοριών που αφορούσαν τον χρήστη, τα σχόλια που είχε υποβάλει ίσως σε κάποιο συγκεκριμένο τραγούδι, τα κανάλια στα οποία είχε κάνει εγγραφή, τον κατάλογο τραγουδιών του, τα αγαπημένα του τραγούδια καθώς και τον αριθμό των συνολικών επισκεπτών στον κανάλι του.

Τα scripts αυτά δημιουργήθηκαν σε γλώσσα προγραμματισμού εφαρμογών διαδικτύου (PHP). Επιπροσθέτως δημιουργήθηκε και ένα φιλικό περιβάλλον διεπαφής ώστε να παρέχονται στο χρήστη μέσα για την ευκολότερη επεξεργασία των δεδομένων δίνοντας του τη δυνατότητα να επιλέξει τη διατήρηση πληροφοριών που εκείνος κρίνει κρίσιμες και απαραίτητες ως προς περαιτέρω επεξεργασία. Όπως προαναφέραμε και πριν μετά την συλλογή των δεδομένων καταλήξαμε με μια συλλογή **388** τραγουδιών όπου για καθένα από αυτά αναζητήθηκαν και βρέθηκαν οι στίχοι και συγκεντρώθηκαν τα σχόλια που είχαν γίνει για κάθε τραγούδι από τους χρηστές.

Το σύνολο των δεδομένων μας δηλαδή τα **388** παραδοσιακά ελληνικά τραγούδια που προέρχονται από 10 διαφορετικές γεωγραφικές περιοχές. Μια αναλυτική παρουσίαση των περιοχών καθώς του αριθμού των τραγουδιών ανά περιοχή καταγράφονται στον παρακάτω πίνακα.

Πίνακας 1.1: Παρουσίαση τραγουδιών ανά γεωγραφική περιοχή

Μικρασιάτικα	80
Θρακιώτικα	16
Ηπειρώτικα	29
Μακεδονίτικα	20
Θεσσαλιώτικα	37
Ρουμελιώτικα	31
Μοραΐτικα	51
Ποντιακά	67
Νησιώτικα	43
Κρητικά	14
Σύνολο	**388**

Κεφάλαιο 2

Πολυδιάστατη Κλιμάκωση

2.1 Εισαγωγή

Η πολυδιάστατη κλιμάκωση (multidimensional scaling, MDS) είναι μια μέθοδος που απεικονίζει μετρήσεις ομοιότητας (ή ανομοιότητας) μεταξύ ζευγαριών αντικειμένων ως αποστάσεις μεταξύ σημείων ενός χαμηλοδιάστατου χώρου. Τα δεδομένα, για παράδειγμα, μπορεί να είναι συσχετίσεις μεταξύ διαφόρων τεστ νοημοσύνης. Η αναπαράσταση MDS δείχνει τα τεστ ως σημεία ενός επιπέδου τα οποία βρίσκονται τόσο κοντά μεταξύ τους όσο πιο θετικώς συσχετισμένα είναι. Η γραφική απεικόνιση των συσχετίσεων που παρέχεται από την MDS επιτρέπει στον αναλυτή δεδομένων κυριολεκτικά να "περιεργαστεί"τα δεδομένα και να διερευνήσει τη δομή τους εποπτικά. Αναδεικνύονται ομαλότητες που συχνά παραμένουν κρυμμένες κατά τη μελέτη πινάκων αριθμών. Μία άλλη εφαρμογή της MDS είναι η χρησιμοποίηση της για αποφάσεις ανομοιότητας. Για παράδειγμα, δοθέντων δύο αντικείμενων ενδιαφέροντος, μπορεί κανείς να εξηγήσει την αντιληπτή τους ανομοιότητα ως αποτέλεσμα μιας νοητικής αριθμητικής που μιμείται το μαθηματικό τύπο απόστασης. Σύμφωνα με αυτό το μοντέλο, το μυαλό δημιουργεί μια εντύπωση της ανομοιότητας με την πρόσθεση των αντιληπτών διαφορών των δύο αντικείμενων πάνω από τις ιδιότητές τους.

Η MDS μπορεί να χρησιμοποιηθεί για διαφορετικούς σκοπούς:

i. Ως μέθοδος που αντιπροσωπεύει δεδομένα ανομοιότητας ως αποστάσεις σε ένα χαμηλό-διάστατο χώρο προκειμένου τα στοιχεία αυτά να αναπαρασταθούν εποπτικώς.

ii. Ως μια τεχνική που επιτρέπει σε κάποιον να δοκιμάσει εάν και πώς ορισμένα κριτήρια σύμφωνα με τα οποία μπορεί κανείς να διακρίνει διαφορές μεταξύ των διαφόρων προτύπων ενδιαφέροντος αντικατοπτρίζονται σε αντίστοιχες εμπειρικές διαφορές αυτών των προτύπων.

iii. Ως μια αναλυτική προσέγγιση δεδομένων που επιτρέπει σε κάποιον να ανακαλύψει τις διαστάσεις που διέπουν τις αποφάσεις ομοιότητας/ανομοιότητας.

iv. Ως ψυχολογικό μοντέλο που εξηγεί τις αποφάσεις της ανομοιότητας σε όρους ενός κανόνα που μιμείται ένα συγκεκριμένο τύπο συνάρτησης απόστασης.

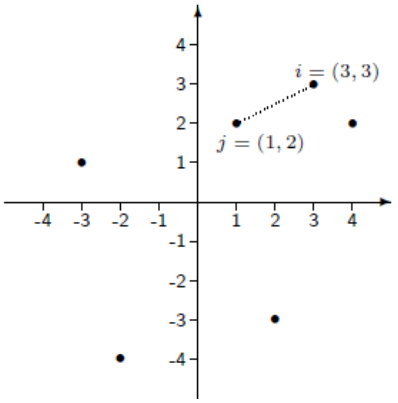
Στην ενότητα 2.2 που ακολουθεί θα εστιάσουμε στα διάφορα μοντέλα καθώς και μέτρα προσαρμογής της MDS.

2.2 Μοντέλα και Μέτρα Προσαρμογής

Τα μοντέλα MDS ορίζονται ως προς τον τρόπο με τον οποίον μεταξύ των προτύπων ενός πίνακα δεδομένων ομοιότητας, ανομοιότητας, ή εγγύτητες (proximities) p_{ij} αντιστοιχίζονται σε αποστάσεις σε χώρο m-διαστάσεων διαμόρφωσης $\mathbf{X}$. Η αντιστοίχιση δίνεται από μια συνάρτηση αναπαράστασης $f\left(p_{ij}\right)$ που καθορίζει πώς οι εγγύτητες θα πρέπει να σχετίζονται με τις αποστάσεις $d_{ij}\left(\mathbf{X}\right)$. Στην πράξη, συνήθως κάποιος δεν προσπαθεί να ικανοποιήσει αυστηρώς την f. Αντίθετα, αυτό που επιδιώκεται είναι μια διαμόρφωση (σε μια δεδομένη διάσταση) της οποίας οι αποστάσεις ικανοποιούν την f όσο το δυνατόν πιο πιστά. Ο όρος "όσο το δυνατόν πιο πιστά"ποσοτικοποιείται από ένα μέτρο τυχους ταιριάσματος (badness-of-fit) [5] ή μιας συνάρτησης απώλειας (*loss function*). Η συνάρτηση απώλειας είναι μια μαθηματική έκφραση που συγκεντρώνει τα σφάλματα εκπροσώπησης, $e_{ij} = f\left(p_{ij}\right) - d_{ij}\left(\mathbf{X}\right)$, σε όλα τα ζεύγη (i, j). Ενα κανονικοποιημένο άθροισμα των τετραγώνων αυτών των σφαλμάτων καθορίζει αυτό που ονομάζεται τάση (*Stress*), και είναι η πιο κοινή συνάρτηση απώλειας στην MDS. Το πώς θα πρέπει να αξιολογηθεί η τάση είναι ένα σημαντικό ζήτημα στην MDS.

2.3 Βασικά Στοιχεία των Μοντέλων Πολυδιάστατης Κλιμάκωσης

Στην ενότητα αυτή, ορίζονται και περιγράφονται τα μοντέλα MDS σε επαρκές επίπεδο για τις περισσότερες πρακτικές εφαρμογές. Ας υποθέσουμε ότι τα μέτρα ομοιότητας ή ανομοιότητας, για τα οποία χρησιμοποιούμε τον γενικό όρο εγγύτητα p_{ij} δίνονται για τα ζεύγη (i, j) n προτύπων. Ένα τέτοιο παράδειγμα εγγύτητας είναι: οι ομοιότητες εγκλημάτων, οι οποίες αξιολογούνται από τους α) συσχετισμούς των συχνοτήτων τους σε διαφορετικές πολιτείες των ΗΠΑ.



Σχήμα 2.1: Ένα καρτεσιανό επίπεδο με μερικά σημεία

β) τις συσχετίσεις μεταξύ των στάσεων, όσον αφορά συμπεριφορές πολιτικής διαμαρτυρίας γ) την άμεση αξιολόγηση της συνολικής ομοιότητας των ζευγαριών χωρών. Όλες αυτές οι περιπτώσεις αποτελούν παραδείγματα μέτρων ομοιότητας, επειδή όσο υψηλότερη είναι η συσχέτιση (ή η εκτίμηση της ομοιότητας), τόσο πιο όμοια είναι τα πρότυπα i και j. Ωστόσο, αντί να αναζητούμε τις αποφάσεις ομοιότητας, είναι εξίσου εύκολο ή ακόμα και πιο εύκολο να αναζητήσουμε τις αποφάσεις ανομοιότητας, για παράδειγμα, αναπαριστώντας μια κλίμακα εκτίμησης η οποία κυμαίνεται από 0 (δηλαδή καμία διαφορά) μέχρι το 10 (δηλαδή, μεγάλη ανομοιότητα).

2.3.1 Συντεταγμένες στο Χώρο Πολυδιάστατης Κλιμάκωσης

Η MDS επιχειρεί να αναπαραστήσει τις εγγύτητες μέσω αποστάσεων μεταξύ σημείων σ᾽ ένα m-διάστατο χώρο που ουσιαστικά είναι ο χώρος MDS. Οι αποστάσεις απόδοσης συντεταγμένων στο χώρο MDS μπορούν να υπολογισθούν μέσω ενός κανόνα μέχρι

ένα ορισμένο επίπεδο ακρίβειας και εάν ο χώρος MDS είναι το πολύ τρισδιάστατος. Ο υπολογισμός μπορεί να καταστεί εφικτός μέσω του συντονισμού του χώρου MDS. Ο πιο κοινός συντονισμός είναι ο καθορισμός πρώτα μιας σειράς από m κατευθυνόμενους άξονες που είναι κάθετοι ο ένας στον άλλο και τέμνονται σε ένα σημείο, την αρχή των αξόνων O. Αυτοί οι άξονες, που στο ισχύον περιβάλλον συχνά αποκαλούνται διαστάσεις - στη συνέχεια χωρίζονται σε διαστήματα ίσου μήκους έτσι ώστε να αναπαριστούν, σε τελικό αποτέλεσμα, ένα σύνολο κάθετων "κανόνων". Κάθε σημείο i, στη συνέχεια, περιγράφεται μοναδικά από μια m-άδα $(x_{i1}, x_{i2}, \ldots, x_{im})$, όπου x_{ia} είναι η προβολή του προτύπου x_i στη διάσταση a. Αυτή η m-άδα είναι το διάνυσμα συντεταγμένων του προτύπου i. Η αρχή των αξόνων O λαμβάνει τις συντεταγμένες $(0, 0, \ldots, 0)$. Το Σχήμα 2.1 δείχνει κάποια σημεία και τα αντίστοιχα διανύσματα συντεταγμένων σε ένα καρτεσιανό επίπεδο, δηλαδή σε ένα επίπεδο που συντονίζεται από ένα σύνολο κάθετων διαστάσεων.

2.3.2 Υπολογισμός Αποστάσεων

Λαμβάνοντας υπόψη ένα καρτεσιανό χώρο, μπορεί κανείς να υπολογίσει την απόσταση μεταξύ οποιωνδήποτε δύο σημείων, i και j. Η πιο συχνά χρησιμοποιούμενη και πιο φυσική συνάρτηση απόστασης είναι η Ευκλείδεια απόσταση. Αντιστοιχεί στο μήκος του ευθύγραμμου τμήματος μιας γραμμής που συνδέει τα σημεία i και j. Το Σχήμα 2.1 δείχνει ένα παράδειγμα. Η Ευκλείδεια απόσταση των σημείων i και j σε μια δισδιάστατη διαμόρφωση $\mathbf{X}$ υπολογίζεται από τον ακόλουθο τύπο:

$$d_{ij}(\mathbf{X}) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}. \tag{2.1}$$

Έτσι, η συνάρτηση $d_{ij}(\mathbf{X})$ είναι ίση με την τετραγωνική ρίζα του αθροίσματος των διαφορών $x_{ia} - x_{ja}$, το οποίο είναι απλά το Πυθαγόρειο θεώρημα για το το μήκος της υποτείνουσας ενός ορθογωνίου τριγώνου. Για το Σχήμα 2.1, ως εκ τούτου, η (2.1) μπορεί επίσης να διατυπωθεί ως:

$$d_{ij}(\mathbf{X}) = \left[\sum_{a=1}^{2} (x_{ia} - x_{ja})^2\right]^{1/2} \tag{2.2}$$

το οποίο μπορεί εύκολα να γενικευτεί στην m-διάστατη περίπτωση ως εξής:

$$d_{ij}\left(\mathbf{X}\right)=\left[\sum_{a=1}^{m}\left(x_{ia}-x_{ja}\right)^{2}\right]^{1/2} \tag{2.3}$$

2.3.3 Μοντέλα και Συναρτήσεις Αναπαράστασης

Η τεχνική MDS αντιστοιχεί τις εγγύτητες p_{ij} σε αποστάσεις $d_{ij}\left(\mathbf{X}\right)$ ενός χώρου MDS $\mathbf{X}$. Δηλαδή, ουσιαστικά έχουμε μια συνάρτηση αναπαράστασης για τον πίνακα δεδομένων $\mathbf{X}$:

$$f:p_{ij}\rightarrow d_{ij}\left(\mathbf{X}\right). \tag{2.4}$$

όπου η συγκεκριμένη επιλογή της f καθορίζει το μοντέλο MDS. Έτσι, ένα μοντέλο MDS είναι ένας μη-γραμμικός μετασχηματισμός της f της εγγύτητας p_{ij} στην απόσταση των προτύπων ij στο χώρο MDS λιγότερων διαστάσεων:

$$f\left(p_{ij}\right)=d_{ij}\left(\mathbf{X}\right). \tag{2.5}$$

Οι αποστάσεις $d_{ij}\left(\mathbf{X}\right)$ στην (2.4) και (2.5) είναι πάντα άγνωστες. Δηλαδή, η τεχνική MDS θα πρέπει να βρει μια διαμόρφωση $\mathbf{X}$ προκαθορισμένης διάστασης m στην οποία θα πρέπει να υπολογίζονται οι αποστάσεις. Η συνάρτηση f, από την άλλη πλευρά, μπορεί είτε να προσδιοριστεί πλήρως, ή μπορεί να περιοριστεί έτσι ώστε να προέρχεται από μια συγκεκριμένη κλάση συναρτήσεων. Ο Shepard το (1957), για παράδειγμα, είχε συλλέξει ομοιότητες p_{ij} για τις οποίες είχε προβλέψει, σε θεωρητικό επίπεδο, ότι θα πρέπει να συσχετίζονται με αποστάσεις σε έναν άγνωστο δισδιάστατο χώρο $\mathbf{X}$ μέσω της εκθετικής συνάρτησης. Δηλαδή, αυτό που ορίστηκε ως υπόθεση είναι ότι $p_{ij}=exp\left[-d_{ij}\left(\mathbf{X}\right)\right]$. Ομοίως, ο Thurstone το (1927) προέβλεψε ότι οι πιθανότητες επιλογής θα πρέπει να είναι ίσες με άγνωστες αποστάσεις μεταξύ των σημείων i και j πάνω σε μια γραμμή μετά τον μετασχηματισμό των $p_{ij}s$ με την αντίστροφη συνάρτηση κανονικής κατανομής. Στις περισσότερες εφαρμογές της MDS, υπάρχει κάποια χαλαρότητα στον προσδιορισμό της f. Για παράδειγμα, η f περιορίζεται μόνο στο να είναι "κάποια" εκθετική συνάρτηση ή "κάποια" γραμμική συνάρτηση. Οι ακριβείς παράμετροι των συναρτήσεων αυτών δεν έχουν καθοριστεί. Μια σημαντική περίπτωση είναι η *MDS* διαστήματος. Η οποία ορίζεται ως:

$$p_{ij} \to a + b \cdot p_{ij} = d_{ij}\left(\mathbf{X}\right). \tag{2.6}$$

για όλα τα σημεία (i, j). Οι παράμετροι a και b είναι ελεύθερες μεταβλητές και μπορούν να επιλεγούν έτσι ώστε να ισχύει η εξίσωση. Μια άλλη περίπτωση είναι η τακτική *MDS*, όπου η f πλέον έχει περιοριστεί στο να είναι μια μονότονη συνάρτηση που διατηρεί τη διάταξη των p_{ij}. Αυτό σημαίνει ότι οι εγγύτητες είναι μετρήσεις ή αποτελέσματα ανομοιότητας:

$$if \;\; p_{ij} < p_{kl}, \;\; then \;\; d_{ij}\left(\mathbf{X}\right) \leq d_{kl}\left(\mathbf{X}\right). \tag{2.7}$$

Εάν $p_{ij} = p_{kl}$ τότε η ανισότητα (2.7) δεν απαιτεί ιδιαίτερη σχέση των αντίστοιχων αποστάσεων. Αυτό είναι γνωστό ως *πρωταρχική προσέγγιση* σε συνδεδεμένες εγγύτητες (tied proximities), όπου οι δεσμοί μπορεί να είναι "ασταθείς"στις αντίστοιχες αποστάσεις. Η *δευτερεύουσα προσέγγιση* όσον αφορά τις συνδέσεις απαιτεί ότι αν $p_{ij} = p_{kl}$, τότε επίσης $d_{ij} = d_{kl}$. Η *πρωταρχική προσέγγιση* είναι η προεπιλογή στα περισσότερα προγράμματα ordinal MDS. Μια μικρή τροποποίηση της εξίσωσης (2.7) αντικαθιστά τη σχέση $\leq$ με $<$. Η πρώτη σχέση καθορίζει μια *ασθενικώς μονότονη* συνάρτηση f, η δεύτερη μια *αυστηρή μονότονη* συνάρτηση. Τις περισσότερες φορές, η τακτική MDS χρησιμοποιείται σε συνδυασμό με μια ασθενικώς μονότονη συνάρτηση. Πώς πρέπει να επιλέξει κανείς μια συγκεκριμένη συνάρτηση αναπαράστασης;

Εάν δεν υπάρχει συγκεκριμένη συνάρτηση f η οποία μπορεί να προέλθει μέσω θεωρητικού συλλογισμού, τότε συνήθως κάποιος προσφεύγει συχνά στον περιορισμό της f σε μια συγκεκριμένη κατηγορία συναρτήσεων με βάση το βαθμό/επίπεδο κλιμάκωσης της εγγύτητας. Κάτω από αυτές τις υποθέσεις, δεν υπάρχει λόγος να επιμένει κάποιος στο ότι αυτά τα διαστήματα να αναπαρασταθούν πιστά ως αποστάσεις στον MDS χώρο. Επιπλέον, μια ασθενικώς μονότονη συνάρτηση κλιμάκωσης καθιστά ευκολότερη την κατά προσέγγιση αναπαράσταση των βασικών πληροφοριών σε ένα χώρο MDS χαμηλής διάστασης. Αντιστρόφως, ξεκινώντας από ένα μοντέλο MDS, μπορεί κανείς να επιλέξει μια συνάρτηση αναπαράστασης g στην υπόθεση παλινδρόμησης g: $d_{ij}\left(\mathbf{X}\right) \to p_{ij}$. Η συγκεκριμένη υπόθεση πρέπει να ελεγχθεί στα δεδομένα. Κάποιος μπορεί να επιλέξει οποιαδήποτε g: εάν οδηγεί σε ένα μοντέλο με το οποίο είναι εμπειρικά ικανοποιημένος και με την προϋπόθεση ότι το μοντέλο δεν ισχύει για τυπικούς λόγους μόνο.

2.3.4 Σφάλματα, Συναρτήσεις Απώλειας και Τάση

Τα μοντέλα MDS απαιτούν κάθε τιμή εγγύτητας να αντιστοιχηθεί επ᾽ ακριβώς σε αποστάσεις. Αυτό αφήνει απ᾽ έξω κάθε έννοια σφάλματος/λάθους. Αλλά οι εμπειρικές τιμές εγγύτητας περιέχουν πάντοτε θόρυβο λόγω της ανακρίβειας των μετρήσεων, της αναξιοπιστίας, των σφαλμάτων δειγματοληψίας, και ούτω καθεξής. Ως εκ τούτου, κάποιος δεν θα πρέπει να επιμείνει, ότι στην πράξη, $f(p_{ij}) = d_{ij}(\mathbf{X})$, αλλά μάλλον ότι $f(p_{ij}) \approx d_{ij}(\mathbf{X})$. Δεδομένου ότι οι εγγύτητες περιέχουν σφάλματα, αυτό επιτρέπει τη δημιουργία καλύτερων και ανθεκτικότερων, ισχυρότερων, πιο αξιόπιστων και ουσιαστικότερων αναπαραστάσεων από ό,τι οι αντίστοιχες τέλειες μη-γραμμικές απεικονίσεις, διότι μπορούν να εξομαλύνουν το θόρυβο. Εάν το σφάλμα της αναπαράστασης των δεδομένων είναι "πάρα πολύ μεγάλο", μπορεί κανείς να απορρίψει ή να τροποποιήσει τη θεωρία, αλλά, προφανώς, πρέπει πρώτα να γνωρίζει πόσο καλά η θεωρία αντιπροσωπεύει τα δεδομένα. Οποιαδήποτε αναπαράσταση η οποία είναι αρκετά ακριβής, ώστε να ελέγξει την εγκυρότητα της θεωρίας είναι αρκετά ορθή. Περαιτέρω επιχειρήματα μπορούν να δημιουργηθούν για την εγκατάλειψη της απαίτησης της ισότητας $f(p_{ij}) = d_{ij}(\mathbf{X})$. Αυτοματοποιημένες διαδικασίες για την εξεύρεση μιας αναπαράστασης MDS συνήθως ξεκινούν με κάποια αρχική διαμόρφωση και βελτιώνουν αυτή τη διαμόρφωση μετακινώντας τριγύρω τα δεδομένα που αναπαριστώνται ως σημεία με μικρά ("επαναληπτικά") βήματα για την προσέγγιση της ιδανική σχέσης μοντέλου $f(p_{ij}) = d_{ij}(\mathbf{X})$ όλο και πιο στενά. Εφ᾽ όσον η εκπροσώπηση δεν είναι τέλεια, κάποιος έχει μόνο την προσέγγιση ότι $f(p_{ij}) \approx d_{ij}(\mathbf{X})$, όπου $\approx$ σημαίνει "ίσα εκτός από κάποια μικρή απόκλιση".

2.3.4.1 Η Συνάρτηση Τάσης/Stress Function

Για να καθορίσουμε έννοιες, όπως "σχεδόν", "παρά λίγο", και ούτω καθεξής, πιο συγκεκριμένα, θα επιστρατεύσουμε τη συχνά χρησιμοποιούμενη στατιστική έννοια του σφάλματος/λάθους. Ένα (τετραγωνικό) σφάλμα αναπαράστασης ορίζεται ως εξής.

$$e_{ij}^2 = \left[f(p_{ij}) - d_{ij}(\mathbf{X})\right]^2. \tag{2.8}$$

Αθροίζοντας το e_{ij}^2 ανάμεσα σε όλα τα ζεύγη (i, j) αποδίδει μια μέτρηση σφάλματος προσαρμογής για ολόκληρη την αναπαράσταση MDS, την επονομαζόμενη *ακατέργαστη τάση/raw stress*.

$$\sigma_r = \sigma_r(\mathbf{X}) = \sum_{(i,j)} \left[f(p_{ij}) - d_{ij}(\mathbf{X}) \right]^2. \tag{2.9}$$

Η τιμή της *ακατέργαστης τάσης* από μόνη της δεν είναι πολύ κατατοπιστική. Μια μεγάλη τιμή δεν σημαίνει κατ᾽ ανάγκη κακή εφαρμογή του μοντέλου. Για παράδειγμα, ας υποθέσουμε ότι οι ανομοιότητες σε ένα σύνολο δεδομένων εκφράζουν οδικές αποστάσεις μεταξύ πόλεων σε χιλιόμετρα. Η ανάλυση MDS σε αυτά τα δεδομένα επιστρέφει την ακόλουθη τιμή του μοντέλου προσαρμογής $\sigma_r(\mathbf{X}_1) = .043$. Εκτελώντας ξανά την ανάλυση με ανομοιότητες που εκφράζονται σε μέτρα αυτή τη φορά παίρνουμε την ίδια λύση, αλλά σε μια κλίμακα που είναι 1000 φορές μεγαλύτερη από την προηγούμενη, οπότε κανείς λαμβάνει $\sigma_r(\mathbf{X}_2) = 43000$. Αυτό δεν σημαίνει ότι το μοντέλο $\mathbf{X}_2$ έχει χειρότερη προσαρμογή στα δεδομένα από ό, τι το $\mathbf{X}_2$, αντανακλά απλώς τη διαφορετική βαθμονόμηση των διαφορών. Για να αποφευχθεί αυτή η κλίμακα εξάρτησης, το σ_r μπορεί, για παράδειγμα, να κανονικοποιηθεί ως εξής.

$$\sigma_1^2 = \sigma_1^2(\mathbf{X}) = \frac{\sigma_r(\mathbf{X})}{\sum d_{ij}^2(\mathbf{X})} = \frac{\sum \left[f(p_{ij}) - d_{ij}(\mathbf{X}) \right]^2}{\sum d_{ij}^2(\mathbf{X})}. \tag{2.10}$$

Υπολογίζοντας την τετραγωνική ρίζα του σ_1^2 αποδίδει μια τιμή γνωστή ως *Τάση-1* (Kruskal, 1964) εξίσωση (2.11). Ο λόγος για τη χρήση του σ_1 αντί για το σ_1^2 είναι ότι το σ_1^2 είναι σχεδόν πάντα πολύ μικρό στην πράξη, οπότε οι σ_1 τιμές είναι ευκολότερο να καταστούν διακριτές. Ορίζοντας το πιο ρητά, έχουμε:

$$\text{Τάση-1} = \sigma_1 = \sqrt{\frac{\sum \left[f(p_{ij}) - d_{ij}(\mathbf{X}) \right]^2}{\sum d_{ij}^2(\mathbf{X})}}. \tag{2.11}$$

Τα αθροίσματα εκτείνονται πάνω από όλα τα σημεία p_{ij} για τα οποία υπάρχουν παρατηρήσεις. Ελλείποντα δεδομένα παραλείπονται. Στην τυπική περίπτωση συμμετρικής εγγύτητας, όπου $p_{ij} = p_{ji}$ (για όλα τα i, j), αρκεί να αθροίσουμε μόνο πάνω στο ήμισυ των ζευγαριών των δεδομένων αποστάσεων. Προφανώς, $\sigma_1 = 0$ μόνο αν $d_{ij}(\mathbf{X}) = f(p_{ij})$. Η ελαχιστοποίηση της *Τάσης-1* απαιτεί πάντα την εύρεση ενός βέλτιστου $\mathbf{X}$ δοσμένης μιας διάστασης m.

Επιπλέον, αν η f καθορίζεται μόνο μέχρι ορισμένες ελεύθερες παραμέτρους, τότε θα πρέπει επίσης να βρεθούν οι βέλτιστες τιμές για αυτές τις παραμέτρους. Αυτό το

πρόβλημα συνήθως επιλύεται μέσω της παλινδρόμησης των p_{ij} επάνω στις υπολογιζόμενες αποστάσεις $\mathbf{X}$. Στο μοντέλο της (*διαστήματος MDS*), κάποιος χρησιμοποιεί γραμμική παλινδρόμηση, ενώ στη *τακτική MDS* τη μονότονη παλινδρόμηση.

Η παλινδρόμηση επιστρέφει μετασχηματισμένες εγγύτητες, $f(p_{ij})$ που είναι "κατά προσέγγιση αποστάσεις"ή ασυμφωνίες(disparities), όπως αλλιώς αναφέρονται στην βιβλιογραφία της MDS [5].

Κεφάλαιο 3

Ανάλυση Λανθάνουσας Σημασιολογίας

3.1 Συνοπτική Περιγραφή της μεθόδου

Η ανάλυση λανθάνουσας σημασιολογίας (Latent Semantic Analysis, LSA) [1], κοινώς γνωστή και ως Ευρετηρίαση Λανθάνουσας Σημασιολογίας στο πλαίσιο της ανάκτησης πληροφοριών [6] είναι μια στατιστική τεχνική για την εξαγωγή σημασιολογικών σχέσεων από τα συμφραζόμενα των λέξεων σε αποσπάσματα ομιλίας ή κειμένων. Βασίζεται στην εφαρμογή της ανάλυσης ιδιαζουσών τιμών (Singular Value Decomposition, SVD), σε έναν όρων-κειμένων πίνακα $W^{t\times d} \in \mathbb{R}^{t\times d}$ όπου t = αριθμός του λεξιλογίου και d = πλήθος των κειμένων. Ο πίνακας W μπορούσε να αντικατασταθεί από μια σειρά πινάκων, οι οποίοι αποτελούνται από πρωτογενές κείμενο το οποίο έχει αναλυθεί σε λήμματα και έχει χωριστεί σε αποσπάσματα, όπως προτάσεις ή παραγράφους. Η LSA παρέχει έναν τρόπο παρατήρησης της συνολική σχέσης μεταξύ των όρων στην συλλογή του συνόλου των κειμένων, που επιτρέπει στις σημασιολογικές δομές εντός της συλλογής των κειμένων να αποκαλυφθούν.

Η εφαρμογή της LSA στην ανάκτηση πληροφοριών έχει ως κίνητρο τις προκλήσεις που ανέκυψαν κατά την επεξεργασία της φυσικής γλώσσας, όπου μια λέξη μπορεί να έχει πολλαπλές σημασίες (πολυσημία, polysemy) ή πολλές λέξεις να αναφέρονται στην ίδια έννοια (συνωνυμία, synonymy). Παρουσιάζοντας έτσι ασάφειες στην έκφραση των εννοιών. Για παράδειγμα ορισμένες εμπειρικές μελέτες δείχνουν ότι η πιθανότητα δύο άτομα να επιλέξουν την ίδια λέξη κλειδί για ένα οικείο αντικείμενο είναι μικρότερη από 15% [7]. Λόγω αυτών των προκλήσεων οι απλές τεχνικές αναζήτησης

μέσω λέξεων-κλειδιών αποδείχθηκαν ανεπαρκείς στην εύρεση κειμένων συναφών με ερωτήματα χρηστών. Η τεχνική LSA [8] επιτρέπει την ανάκτηση με βάση το σημασιολογικό περιεχόμενο, αντί των απλοϊκών τεχνικών που απλώς ταιριάζουν λέξεις μεταξύ των ερωτημάτων και κειμένων. Χρησιμοποιώντας τεχνικές μείωσης διαστάσεων που τείνουν να τιθασεύουν τα πολύπλοκα φαινόμενα της φυσικής γλώσσας και την εγγενώς καθολική προσέγγιση της σημασιολογίας που συμπληρώνει την τοπική βελτιστοποίηση που πραγματοποιείται από πιο συμβατικές τεχνικές, φαίνεται ότι η LSA είναι μια πολύ καλύτερη προσέγγιση στην ανάκτηση πληροφοριών.

Η LSA επίσης εμφανίζεται ιδιαίτερα ελκυστική λόγω της απεικόνισης των διακριτών οντοτήτων (όροι ή κείμενα) σε ένα συνεχή παραμετρικό χώρο, όπου μπορούν να εφαρμοστούν αποδοτικοί αλγόριθμοι μηχανικής μάθησης. Η LSA μπορεί να εφαρμοστεί σε πολλούς τομείς, εφ᾽ όσον υπάρχει ένα σύνολο των αναγνωρίσιμων μεμονωμένων μονάδων (δηλαδή όρων) και μια συλλογή για τις μονάδες αυτές. Στην ανάκτηση πληροφοριών, χρησιμοποιείται ένα σύνολο από επιμέρους όρους οι οποίοι περιέχονται σ᾽ ένα σύνολο κειμένων. Η LSA υποθέτει ότι υπάρχουν κρυμμένες λανθάνουσες σημασιολογικές δομές που οφείλονται στην τυχαιότητα των λέξεων στα κείμενα και οι οποίες θα μπορούσαν να αποκαλυφθούν με την εφαρμογή ανάλυσης ιδιαζουσών τιμών. Η διαδικασία περιλαμβάνει την ανάλυση της συλλογής των κειμένων ως προς την εξαγωγή λημμάτων που ενδέχεται να χρησιμοποιηθούν στα ερωτήματα των χρηστών και την κατασκευή ενός πίνακα διαστάσεων $W^{t \times d}$.

Έπειτα εφαρμόζουμε SVD ώστε να αποκαλύψουμε τις σημασιολογικές δομές χρησιμοποιώντας τον πίνακα $W^{t \times d}$. Η τεχνική SVD προβαίνει σε μείωση των διαστάσεων του W ο οποίος συνήθως είναι πολύ αραιός, συνεπώς βελτιστοποιώντας την αποθήκευση, καθώς και απομακρύνοντας το θορύβου από τα δεδομένα. Τα ερωτήματα των χρηστών απαντιούνται σ᾽ αυτό το μειωμένο χώρο, όπου η LSA είναι σε θέση μέσω της συσχέτισης των λέξεων και κειμένων να συμπεράνει τη σχέση μεταξύ κειμένων και λέξεων [9].

3.2 Μέτρηση απόδοσης συστημάτων ανάκτησης πληροφοριών (IR)

Η απόδοση ενός συστήματος ανάκτησης πληροφοριών περιγράφεται βασιζόμενη κυρίως σε δύο μέτρα, συγκεκριμένα την *Ανάκληση (Recall) και Ακρίβεια (Precision)*:

- *Η Ανάκληση:* μας δίνει το ποσοστό των σχετικών κειμένων στη συλλογή που επιστρέφει το σύστημα ανάκτησης [1]. Ως εκ τούτου, μετρά την ικανότητα του συστήματος ανάκτησης να εντοπίσει όλα τα σχετικά αντικείμενα/έγγραφα. Μια ανάκληση πληροφορίας 100% μπορεί να επιτευχθεί μόνο εάν όλα τα σχετικά έγγραφα ανακτηθούν επιτυχώς:

$$\text{Ανάκληση} := \frac{|\ \{\text{σχετικά κείμενα}\} \cap \{\text{ανακληθέντα κείμενα}\}\ |}{|\ \{\text{σχετικά κείμενα}\}\ |} \tag{3.1}$$

- *Η Ακρίβεια:* μας επιστρέφει το ποσοστό των ανακληθέντων αποτελεσμάτων που είναι σχετικά ως προς το ερώτημα που τέθηκε [1]. Μετρά ως εκ τούτου την ακρίβεια της ανάκτησης του συστήματος πληροφοριών:

$$\text{Ακρίβεια} := \frac{|\ \{\text{σχετικά κείμενα}\} \cap \{\text{ανακληθέντα κείμενα}\}\ |}{|\ \{\text{ανακληθέντα κείμενα}\}\ |} \tag{3.2}$$

3.3 Μειονεκτήματα της τεχνικής ανάκτησης πληροφοριών term-matching

Τα δεδομένα προτού μετασχηματιστούν στον πίνακα tf-idf βρίσκονται στην πιο ακατέργαστη μορφή τους και χρειάζεται να μετατραπούν σε μια μορφή ώστε η ανάκτηση χρήσιμων πληροφοριών να είναι εφικτή (δηλ. η μετατροπή αφορά την αναπαράσταση του κειμένου των στίχων των τραγουδιών σε μορφή επεξεργάσιμη από τον Η/Υ). Η αρχαιότερη τεχνική που χρησιμοποιείται στην ανάκτηση πληροφοριών είναι η ανάκτηση μέσω ταιριάσματος συναφών όρων *"term-matching retrieval"* η οποία επιτρέπει να αναπαραστήσουμε τη συλλογή κειμένων ως πίνακα αποτελούμενο από όρους και κείμενα με απώτερο σκοπό να θέτουμε ερωτήματα χρησιμοποιώντας ένα σύνολο από διαφορετικούς όρους έτσι ώστε να ανακτώνται σχετικά κείμενα προς τα ερωτήματα που τίθενται.

Πίνακας 3.1: Πίνακας όρων-κειμένων

$$W = \begin{pmatrix} & \text{Κείμενο}^1 & \text{Κείμενο}^2 & \text{Κείμενο}^3 & \dots & \text{Κείμενο}^n \\ \text{Όρος}^1 & 5 & 0 & 6 & \dots & 1 \\ \text{Όρος}^2 & 54 & 21 & 0 & \dots & 21 \\ \text{Όρος}^3 & 0 & 5 & 21 & \dots & 5 \\ \vdots & \vdots & \vdots & \vdots & & \\ \text{Όρος}^m & 25 & 0 & 51 & \dots & 32 \end{pmatrix}$$

Παρόλη την απλότητα της τεχνικής εντοπίζονται ορισμένα ελαττώματα. Τα συστήματα ανάκτησης πληροφοριών που βασίζονται σε τεχνικές συναφών όρων προσπαθούν να απαντήσουν σε ερωτήματα υποθέτοντας ότι οι όροι ενός ερωτήματος ταιριάζουν με τους όρους ενός κειμένου. Αυτό σε κάποιο βαθμό, ικανοποιεί τις ανάγκες του χρήστη. Παρόλα αυτά, πολλά κείμενα τα οποία είναι σχετικά ως προς τις ανάγκες του χρήστη δεν ανακτώνται ενώ άσχετα κείμενα ανακτώνται. Οι ακόλουθοι λόγοι οι οποίοι προκύπτουν την πολυσημία και συνωνυμία εξηγούν αυτή την φτωχή απόδοση:

i Τα ευρετήρια δεν περιέχουν ένα συνδυασμό όλων των όρων που οι χρήστες θα μπορούσαν να χρησιμοποιήσουν για να αναζητήσουν κείμενα, αλλά ένα υποσύνολο αυτών των όρων.

ii Τεχνικές οι οποίες έχουν χρησιμοποιηθεί, ώστε να λύσουν το πρόβλημα της πολυσημίας, όπως το λεξικό συνωνύμων και συγγενικών όρων (thesaurus) μπορεί να παρουσιάσουν περισσότερα προβλήματα. Για παράδειγμα, λέξεις που προστίθενται μπορεί να έχουν διαφορετικό νόημα από το επιδιωκόμενο, ως εκ τούτο προκαλώντας περισσότερη υποβάθμιση στην μέτρηση της ακρίβειας.

iii Δεν υπάρχει ικανοποιητική ή επαρκής αυτόματη μεθοδολογία για την αντιμετώπιση της πολυσημίας. Προσεγγίσεις ώστε να χρησιμοποιηθούν ελεγχόμενα λεξικά είναι εξαιρετικά δαπανηρές και λιγότερο αποτελεσματικές.

iv Στο μοντέλο σάκου λέξεων (bag of words): Κάθε τύπος λέξης [9] αντιμετωπίζεται ανεξάρτητα από οποιοδήποτε άλλο τύπο. Κατά συνέπεια, ταιριάζοντας τους όρους εκείνους που σχεδόν πάντα εμφανίζονται μαζί παράγει ισοδύναμο αποτέλεσμα με το ταίριασμα δύο όρων που σπάνια βρίσκονται στο ίδιο κείμενο. Η διαδικασία αυτή αποτυγχάνει να λάβει υπόψη της τον πλεονασμό, ο οποίος μπορεί να οδηγήσει σε στρέβλωση των αποτελεσμάτων.

3.4 Διαδικασία Ανάλυσης LSA

3.4.1 Στάδιο Προ Επεξεργασίας

Τα κείμενα ως γνωστόν περιέχουν όρους μερικοί από τους οποίους είναι πολύ συχνοί, ενώ άλλοι είναι πολύ σπάνιοι με τρόπο τέτοιο, ώστε η εμφάνισή τους να εξαρτάται από το θέμα του κειμένου. Λόγω αυτών των κοινών όρων, τα κείμενα αναλύονται σε αυτό το στάδιο ώστε να εξαχθούν μόνο οι λέξεις-κλειδιά για να χρησιμοποιηθούν στην κατασκευή του πίνακα όρων-κειμένων. Το συγκεκριμένο στάδιο περιλαμβάνει διεργασίες που περιγράφονται στις ενότητες 3.4.2, 3.4.3:

3.4.2 Εξάλειψη κοινών λέξεων

Αυτό το στάδιο περιλαμβάνει την εξάλειψη κοινών λέξεων (stop words) που έχουν μικρότερη διακριτική ικανότητα στη συλλογή των κειμένων. Μια στρατηγική για την εξάλειψη μπορεί να περιλαμβάνει την εξαγωγή όλων των όρων που εμφανίζονται σε ολόκληρη την συλλογή των εγγράφων και εν συνεχεία την απομάκρυνση εκείνων με υψηλή συχνότητα εμφάνισης σε κάθε έγγραφο. Για παράδειγμα, στην Ελληνική γλώσσα, οι ακόλουθες λέξεις μπορεί να έχουν υψηλή συχνότητα εμφάνισης σε μια συλλογή κειμένων και ως εκ τούτου θα πρέπει να απορρίπτονται κατά την κατασκευή του πίνακα όρων-κειμένων:

$\{$*είναι, από, αυτό, για, έχει, στο, θα, με, ήταν, μαζί, όπως, όπου, κατά, όταν...*$\}$

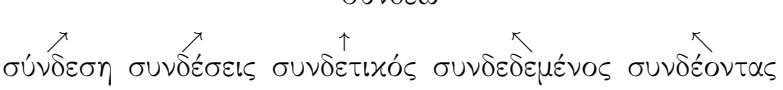
Η απόρριψη των κοινών λέξεων μπορεί να επηρεάσει, ωστόσο, αρνητικά την απόδοση ενός συστήματος ανάκτησης. Αυτό οφείλεται στο γεγονός ότι ορισμένες κοινές λέξεις μπορούν να συμβάλλουν στο νόημα του κειμένου. Για παράδειγμα, μερικοί στίχοι τραγουδιών μπορεί να αποτελούνται μόνο από κοινές λέξεις. π.χ. Να ζει κανείς ή να μην ζει, ας είναι και δεν θέλω να είμαι, μεταξύ άλλων. Η εξάλειψη των κοινών λέξεων οδηγεί στον αποκλεισμό των κειμένων από τη διαδικασία ανάκτησης.

3.4.3 Εύρεση Θέματος και Λημματογράφηση

Στη φυσική γλώσσα, έχουμε λέξεις οι οποίες έχουν διαφορετικές μορφές επιφανείας π.χ. μορφολογικές παραλλαγές αλλά φέρουν το ίδιο νόημα. Για παράδειγμα, στην Ελληνική γλώσσα, μπορούμε να έχουμε στο ίδιο κείμενο λέξεις όπως: *διοργάνωση,*

διοργανώνοντας και διοργανώνει οι οποίες θα αύξαναν άσκοπα τις διαστάσεις του πίνακα όρων-κειμένων. Θα ήταν λογικό να αποθηκεύονταν απλά η λέξη *διοργάνωση* αντί και οι τρεις λέξεις. Αυτό επιτυγχάνεται με την τεχνική εύρεσης θέματος και τη λημματογράφηση.

Η εύρεση θέματος (stemming) [1] συνήθως αναφέρεται σε μια πρώιμη διαδικασία που αφαιρεί τα προσφύματα (προσθήματα ή επιθήματα) από τις λέξεις ελπίζοντας στον επιτυχή εντοπισμό του θέματος. Η λημματογράφηση [1] αναφέρεται στη χρήση λεξιλογίου και μορφολογικής ανάλυσης των λέξεων, στη συμπίεση όλων των μορφών επιφανείας χρησιμοποιώντας μορφολογική ανάλυση σε ένα λήμμα. π.χ. ονομαστική ενικού αριθμού ουσιαστικών. Υπάρχουν πολλοί αλγόριθμοι εύρεσης θέματος και λημματογράφησης και η επιλογή εξαρτάται κυρίως από το είδος της εφαρμογής. Ο πιο κοινός και εμπειρικά ενεργός stemming αλγόριθμος για την Ελληνική γλώσσα είναι ο ***Greek Stemmer***, ο οποίος περιγράφεται εν συντομία στο [10]. Ένα παράδειγμα της εύρεσης θέματος εμφανίζεται παρακάτω:



Σχήμα 3.1: stemming process

Οι στρατηγικές εύρεσης θέματος και λημματογράφησης μπορεί να επηρεάσουν τις επιδόσεις ενός σύστημα ανάκτησης πληροφοριών. Για παράδειγμα εάν χρησιμοποιούσαμε τον αλγόριθμο Greek Stemmer, οι ακόλουθες λέξεις θα μπορούσαν όλες να πηγάζουν από το ίδιο πρόθεμα (stem) ***λειτ...*** επηρεάζοντας το ερώτημα του χρήστη για λέξεις όπως ***λειτουργικό σύστημα, λειτουργική οδοντιατρική και λειτουργική έρευνα***:
{*λειτουργώ, λειτουργώντας, λειτουργεί, λειτουργία, λειτουργική, λειτουργοί, λειτουγρικός...*}

3.4.4 Συντελεστές

Η κατασκευή του πίνακα όρων-κειμένων $W^{m\times n}$ ακολουθεί μετά την εξαγωγή του λεξικού από το σύνολο των κειμένων. Τα στοιχεία του πίνακα $W_{i,j}$ **αντιπροσωπεύουν το βάρος του** i^{th} **όρου στο** j^{th} **έγγραφο**. Υπάρχουν πολλές προσεγγίσεις που μπορούν να χρησιμοποιηθούν ώστε να σταθμίσουμε τους όρους σε μια συλλογή. Οι ακόλουθες προσεγγίσεις είναι οι πιο κοινές:

a) **Συχνότητα Όρων** $tf_{i,j}$

Αποδίδονται βάρη στους όρους με βάση την συχνότητα εμφάνισης ενός όρου i σε ένα κείμενο j. Το κρίσιμο πρόβλημα της συγκεκριμένης προσέγγισης είναι ότι όλοι οι όροι θεωρούνται ίσης σημασίας στη διαδικασία απάντησης ερωτημάτων,

ακόμη χειρότερα όταν παρατηρείται μεγάλη διακύμανση μεταξύ κοινών όρων και σπάνια εμφανιζόμενων όρων. Για παράδειγμα ας εξετάσουμε το κείμενο που εμφανίζεται παρακάτω στην Αγγλική γλώσσα:

"ParseTreeView defines a view that is used to display the abstract syntax tree which is resulted by parsing the text content of an editor."

Στο κείμενο ο όρος ***"view"*** και ***"ParseTreeView"*** θα έχουν το ίδιο βάρος με βάση τη συχνότητα εμφάνισης τους, παρόλο που ο όρος ***"ParseTreeView"*** θα άξιζε μεγαλύτερο βάρος. Το επόμενο πρόβλημα προκύπτει από τα διαφορετικά μεγέθη των κειμένων. Σ᾽ αυτή την περίπτωση ενδέχεται, κάποιοι όροι να λάβουν υψηλότερη συχνότητα σε κείμενα μεγάλου μεγέθους/μήκους, σε αντίθεση με μικρότερα κείμενα. Για να αποφευχθεί η προκατάληψη προς τα κείμενα μεγάλου μήκους, όπου οι περισσότεροι όροι μπορεί να εμφανίσουν αδικαιολόγητα υψηλή συχνότητα εμφάνισης, είναι επιθυμητό η μέτρηση της συχνότητας ενός όρου $tf_{i,j}$ να κανονικοποιείται με έναν παράγοντα λ_j που εξαρτάται από το κείμενο j.

$$W_{i,j} = \frac{tf_{i,j}}{\lambda_j} \tag{3.3}$$

Ανάλογα με την εφαρμογή, μπορούν να χρησιμοποιηθούν διαφορετικοί τρόποι για τον υπολογισμό του παράγοντα λ_j.

Παραδείγματα υπολογισμού του λ_j:

i.
$$\lambda_j = \|tf_{i,j}\|_1 := \sum_i \mid tf_{i,j} \mid \quad \text{όπου } \|\cdot\|_1 \text{ είναι η } \ell_1 \text{ νόρμα} \tag{3.4}$$

ii.
$$\lambda_j = \|tf_{i,j}\|_2 := \sqrt{\sum_i f_{i,j}^2} \text{ όπου } \|\cdot\|_2 \text{ είναι η } \ell_2 \text{ νόρμα} \tag{3.5}$$

b) **Συντελεστές συχνότητας όρων αντίστροφης συχνότητας κειμένων**

Το συγκεκριμένο σύστημα στάθμισης βαρών συνδυάζει την ιδέα της συχνότητας εμφάνισης όρων $tf_{i,j}$ και της αντίστροφης συχνότητας κειμένων (inverse document frequency, idf_i). Η αντίστροφη συχνότητα εγγράφου προσπαθεί να καταλήξει σε έναν καλύτερο μηχανισμό διάκρισης για την απάντηση των ερωτημάτων, λαμβάνοντας υπόψη τις στατιστικές σε επίπεδο κειμένου. Η κύρια ιδέα είναι να εκχωρηθούν σε όρους που εμφανίζονται σε πολλαπλά κείμενα μικρότερο βάρος επίδρασης σε ένα ερώτημα από τους όρους εκείνους που εμφανίζονται σε

λιγότερα κείμενα. Η αντίστροφη συχνότητα κειμένου ενός όρου i στη συλλογή N κειμένων δίνεται από την ακόλουθη σχέση:

$$idf_i = log\left(\frac{N}{df_i}\right) \tag{3.6}$$

Όπου Idf_i είναι ο συνολικός αριθμός κειμένων σε όλη τη συλλογή που περιέχουν τον όρο i. Για λόγους απεικόνισης χρησιμοποιούμε το παράδειγμα που δίνεται στο [1], το οποίο είναι από τη συλλογή Reuters-RCVI 806.791 κειμένων.

$$\begin{bmatrix} i \text{ στός όρος} & df_i & idf_i \\ Car & 18,165 & 1.65 \\ Auto & 6,723 & 2.08 \\ Insurance & 19,241 & 1.62 \\ Best & 25,235 & 1.5 \end{bmatrix}$$

Σχήμα 3.2: Παράδειγμα αντίστροφης συχνότητας κειμένων για τέσσερις όρους από τη συλλογή Reuters [1].

Από το Σχήμα 3.2 μπορούμε να διακρίνουμε καθαρά πως όροι με υψηλή συχνότητα εμφάνισης σε πολλαπλά κείμενα, ουσιαστικά λαμβάνουν μικρότερη στάθμιση ως προς την αντίστροφη συχνότητα εμφάνισης κειμένων idf_i. Για τον υπολογισμό των βαρών του i οστού όρου, συνδυάζουμε πλέον την ιδέα της συχνότητας κειμένων με την συχνότητα εμφάνισης ενός όρου:

$$tf - idf_{i,j} = tf_{i,j} \quad \times \quad idf_i. \tag{3.7}$$

Η παραπάνω εξίσωση σημαίνει πως η τεχνική $tf - idf_{i,j}$ [4] αναθέτει σε ένα συγκεκριμένο όρο i του κειμένου j βάρος που είναι:

- υψηλότερο, όταν ο όρος i εμφανίζεται πολλές φορές σε ένα μικρό αριθμό κειμένων
- μικρότερο, όταν ο όρος i εμφανίζεται λιγότερες φορές σε ένα κείμενο ή εμφανίζεται συχνά σε πολλά κείμενα
- ακόμα μικρότερο, όταν ο όρος συναντιέται σχεδόν σε όλα τα κείμενα

Η συγκεκριμένη προσέγγιση καθορισμού και ανάθεσης βαρών είναι καλύτερη από τη συχνότητα όρων που συζητήθηκε νωρίτερα, αφού συνδυάζει τόσο σε τοπικό

όσο και σε καθολικό επίπεδο στατιστικά στοιχεία σε μια συλλογή κειμένων ώστε να υπολογιστεί το βάρος του όρου i οστού όρου στο j οστό κείμενο.

3.4.5 Μοντέλο Αντιστοίχισης Διανυσματικού Χώρου

Όταν πλέον, καθοριστεί και επιλεχθεί ένα σύστημα στάθμισης, τα κείμενα και οι όροι τους αντιστοιχίζονται σε ένα μοντέλο διανυσματικού χώρου (VSM). Το VSM είναι ένα αλγεβρικό μοντέλο που χρησιμοποιείται για την αντιπροσώπευση των κειμένων ως διανύσματα. Το μοντέλο διανυσματικού χώρου περιλαμβάνει κάθε κείμενο που αντιστοιχίζεται σε ένα διάνυσμα cj. Το cj αντιστοιχεί στη j οστή στήλη του πίνακα όρων κειμένων. Η σχέση αυτή παρουσιάζεται παρακάτω:

$$W := [c_1|c_2|\dots|c_n] = (r_1, r_2, \dots, r_m)^T \in \mathbb{R}^{m\times n} \tag{3.8}$$

Στην 3.8 $r_1, r_2, \dots, r_m$ αναφέρονται στις γραμμές του πίνακα $W^{m\times n}$ οι οποίες συγκροτούνται από τα βάρη των όρων για κάθε κείμενο.

3.4.6 Μείωση Διαστάσεων

Η μείωση διάστασης είναι μια στρατηγική με στόχο την εξασφάλιση οικονομικότερης αντιπροσώπευσης καθώς και καλύτερης σημασιολογικής αναπαράσταση των δεδομένων. Στην LSA, ο πίνακας όρων-κειμένων είναι πολύ αραιός και μεγάλος σε μέγεθος, κάτι το οποίο έχει δραματική επίδραση στη μνήμη καθώς απαιτεί μεγάλο αποθηκευτικό χώρο. Χρησιμοποιείται η αποσύνθεση ιδιαζουσών τιμών (SVD) για την μείωση της διάστασης του πίνακα και για να αναδείξει λανθάνουσα σημασιολογική δομή.

3.4.6.1 Αποσύνθεση Ιδιαζουσών Τιμών

Η αποσύνθεση ιδιαζουσών τιμών είναι μια τεχνική που σχετίζεται στενά με την αποσύνθεση ιδιοδιανυσμάτων (eigenvector decomposition) και την ανάλυση παραγόντων. Είναι μια σημαντική παραγοντοποίηση ενός "ορθογώνιου"πραγματικού ή μιγαδικού πίνακα καθώς χρησιμοποιείται στη Γραμμική Άλγεβρα με αρκετές εφαρμογές στην Επεξεργασία Σήματος και τη Στατιστική. Για έναν πίνακα $A^{m\times n}$ γραμμοβαθμού r, υπάρχει μια παραγοντοποίηση της μορφής [11]:

$$\begin{array}{ccccc} A = & U & & S & V^T \\ & \nearrow & & \uparrow & \nwarrow \\ & [m\times m] & & [m\times n] & [n\times n] \end{array} \tag{3.9}$$

Όπου:

- $U^{m\times m}$ είναι ένας πίνακας διαστάσεων $m \times m$ του οποίου οι στήλες είναι ορθογώνια ιδιοδιανύσματα του πίνακα AA^T.
- $S^{m\times n}$ είναι ένας πίνακας διαστάσεων $m \times n$ του οποίου οι r τιμές επί της διαγωνίου με $r \leq \min(m, n)$ αποτελούν τις ιδιάζουσες τιμές του $A^{m\times n}$ και είναι μη μηδενικές.
- V^T είναι ο ανάστροφος του πίνακα $V^{n\times n}$, του οποίου οι στήλες είναι τα ιδιοδιανύσματα του πίνακα A^TA.

Παράδειγμα:

Δοσμένου ενός πίνακα:

$$A = \begin{pmatrix} 1.000 & 0.000 & 0.000 & 0.000 & 2.000 \\ 0.000 & 0.000 & 3.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 4.000 & 0.000 & 0.000 & 0.000 \end{pmatrix}$$

η ανάλυση ιδαζουσών τιμών οδηγεί στους πίνακες:

$$U = \left(\begin{array}{c|c|c|c} 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & -1.000 \\ 1.000 & 0.000 & 0.000 & 0.000 \end{array}\right) \times S = \left(\begin{array}{ccc|cc} 4.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 3.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 2.236 & 0.000 & 0.000 \\ \hline 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \end{array}\right)$$

$$\times \; V^T = \left(\begin{array}{ccccc} 0.000 & 1.000 & 0.000 & 0.000 & 0.000 \\ \hline 0.000 & 0.000 & 1.000 & 0.000 & 0.000 \\ \hline 0.447 & 0.000 & 0.000 & 0.000 & 0.894 \\ \hline 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ -0.894 & 0.000 & 0.000 & 0.000 & 0.447 \end{array}\right) \tag{3.10}$$

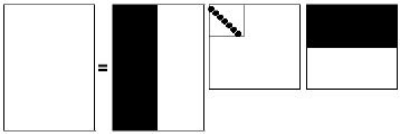
οπότε η μείωση διαστάσεων περιλαμβάνει στη συνέχεια περικοπή των τριών πινάκων που λαμβάνονται από την SVD. Βασικά οι πρώτες $r = 3$ στήλες του $U^{m\times n}$, οι πρώτες $r = 3$ γραμμές του V^T και ο κύριος $r \times r$ υπο πίνακας του $S^{m\times n}$ διατηρούνται. Η

βασική αρχή είναι να καταλήξουμε με έναν $m \times n$ πίνακα A_r γραμμοβαθμού (rank) το πολύ r, έτσι ώστε να ελαχιστοποιείται η Frobenius νόρμα της διαφοράς του πίνακα $X = A - A_r$ που μετρά την απόκλιση μεταξύ του πίνακα A_r και A και ορίζεται ως εξής:

$$\|X\|_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} X_{ij}^2}. \tag{3.11}$$

Ο στόχος σε αυτή την περίπτωση είναι να βρούμε έναν πίνακα A_r που να ελαχιστοποιεί τη διαφορά αυτή, ενώ παράλληλα περιορίζει το πολύ σε r το γραμμοβαθμό του A_r. Αυτή η διαδικασία της εύρεσης ενός r τέτοιου ώστε ο πίνακας A_r να καταλήξει με χαμηλότερο γραμμοβαθμό από εκείνον του αρχικού πίνακα A ονομάζεται προσέγγιση χαμηλού γραμμοβαθμού (low rank approximation). Δεν υπήρξε κάποια κοινώς αποδεκτή στρατηγική μέχρι σήμερα για την απόφαση εύρεσης του βέλτιστου r για να χρησιμοποιηθεί στην διατήρηση των διαστάσεων. Συνήθως γίνεται μια εμπειρική επιλογή του r και αξιολογείται η επιλογή αυτή εμμέσως από τις επιδόσεις της ανάκτησης. Αν το r είναι πολύ μεγάλο, μπορεί να παραμείνει θόρυβος στον διανυσματικό χώρο, ενώ αντίστοιχα μια πολύ χαμηλή του r μπορεί να οδηγήσει σε εξάλειψη σημαντικών πληροφοριών. Η διαδικασία μείωσης διαστάσεων εμφανίζεται παρακάτω:

$$\begin{array}{rccc} A_r = & U_r & S_r & V_r^T. \\ & \downarrow & \downarrow & \downarrow \end{array} \tag{3.12}$$



Σχήμα 3.3: Μείωση διαστάσεων μέσω SVD [1].

Η μείωση των διαστάσεων θα δώσει μια νέα αναπαράσταση για τους όρους και τα κείμενα της συλλογής. Ενώ το VSM που συζητήθηκε παραπάνω είναι ικανό να αντιμετωπίσει ομοιόμορφα τα ερωτήματα και τα κείμενα, η μείωση διαστάσεων είναι αυτή που ουσιαστικά αποκαλύπτει τη σημασιολογική δομή που δεν αποκαλύπτονταν στο VSM. Επίσης η μείωση διαστάσεων βοηθά στην αποκάλυψη περισσότερων σχέσεων μεταξύ όρων και κειμένων, γεγονός το οποίο συμβάλει στην υψηλότερη επίδοση ανάκτησης σε σύγκριση με το VSM.

3.4.7 Ανάκτηση Κειμένων

Στο χώρο μειωμένου γραμμοβαθμού, οι όροι που εμφανίζονται σε παρόμοια κείμενα είναι κοντινοί, ακόμη και αν δεν μπορούν να συνυπάρχουν στο ίδιο κείμενο. Οι όροι του ερωτήματος επίσης αντιστοιχίζονται στον ίδιο χώρο, όπου μετρικές ομοιότητας μπορούν να χρησιμοποιηθούν για να μετρήσουν τις αποστάσεις μεταξύ όρων του ερωτήματος και των κειμένων. Χάρη σε αυτή τη σύγκριση κειμένων και των όρων των ερωτημάτων στον ίδιο χώρο, είναι δυνατή η ανάκτηση πληροφοριών από την LSA.

3.4.7.1 Αντιστοίχιση Ερωτημάτων στον r-διανυσματικό χώρο

Στο σημείο αυτό χρειαζόμαστε ένα μηχανισμό αντιστοίχισης ερωτημάτων στο διανυσματικό χώρο γραμμοβαθμού καθώς και μια στρατηγική βαθμολόγησης για την ταξινόμηση των κειμένων. Η ιδέα αφορά στην μετατροπή του ερωτήματος σε μια διανυσματική αναπαράσταση ώστε να είναι δυνατή η σύγκριση μεταξύ των κειμένων και των ερωτημάτων. Ένα διάνυσμα ερωτήματος q (query vector) αντιπροσωπεύεται στον διανυσματικό χώρο γραμμοβαθμού r μέσω του ακόλουθου διανύσματος q_r:

$$q_r = S_r^{-1} \; U_r^T \; q. \tag{3.13}$$

3.4.7.2 Μετρικές Ομοιότητας

Μόλις ένα ερώτημα αντιστοιχηθεί στο διανυσματικό χώρο γραμμοβαθμού r μπορεί πλέον να συγκριθεί με τα αντίστοιχα κείμενα στον ίδιο διανυσματικό χώρο χρησιμοποιώντας μία μετρική ομοιότητας. Στο VSM, για να υπολογίσουμε την απόσταση ενός κειμένου d απ᾽ ένα οποιοδήποτε ερώτημα q χρησιμοποιούμε το συνημίτονο της γωνίας των δύο διανυσμάτων. Η μέθοδος αυτή μπορεί επίσης κάλλιστα να υπολογίσει την απόσταση μεταξύ δύο οποιονδήποτε κειμένων ή όρων. Η μέτρηση της απόστασης από ένα κείμενο ως προς ένα ερώτημα περιλαμβάνει τα ακόλουθα:

Ας υποθέσουμε ότι συμβολίζουμε με $\vec{V}(d)$ το διάνυσμα που αντιπροσωπεύει το κείμενο d, χρησιμοποιώντας κάθε όρο του λεξικού (δηλαδή όλες τις λέξεις-κλειδιά που χρησιμοποιούνται για την κατασκευή του πίνακα όρων-κειμένων). Ας υποθέσουμε επίσης ότι αναπαριστούμε το διάνυσμα ερωτήματος με $\vec{V}(q)$. Στη συνέχεια, το συνημίτονο της γωνίας μεταξύ των $\vec{V}\,(d)$ και $\vec{V}\,(q)$ χρησιμοποιείται για να διαβαθμίσει τη συνάφεια του ερωτήματος q με το κείμενο d.

$$score(q,\ \ d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{\mid \vec{V}(q) \mid \cdot \mid \vec{V}(d) \mid}. \tag{3.14}$$

Στην 3.14 ο αριθμητής αντιπροσωπεύει το εσωτερικό γινόμενο των δύο διανυσμάτων $\vec{V}(q)$ και $\vec{V}(d)$ που θα ήταν αρκετά επαρκές για να μετρήσει τη συνάφεια του κειμένου d και του ερωτήματος q που ως διανύσματα είχαν το ίδιο μέτρο. Ωστόσο, δεδομένου ότι αυτό δεν ισχύει στην προκειμένη περίπτωση, ο παρονομαστής εισάγεται ως τρόπος κανονικοποίησης μήκους. Αυτό ονομάζεται Ευκλείδεια κανονικοποίηση, δεδομένου ότι πρόκειται για τον υπολογισμό του Ευκλείδειου μήκους για κάθε διάνυσμα που λαμβάνεται υπόψη. Για την κατάταξη των άλλων κειμένων βάσει του ερωτήματος, υπολογίζεται η βαθμολογία για κάθε κείμενο. Με βάση το σύνολο των αποτελεσμάτων, το έγγραφο με την υψηλότερη βαθμολογία κατατάσσεται στην πρώτη θέση. Ένα τέτοιο παράδειγμα εμφανίζεται παρακάτω:

$$\begin{bmatrix} \text{Διάνυσμα κειμένου} & score(q,\ d) & rank \\ V(\vec{d}_1) & 0.8882 & 1 \\ V(\vec{d}_2) & 0.3338 & 3 \\ V(\vec{d}_3) & 0.5983 & 2 \end{bmatrix}$$

Σχήμα 3.4: Παράδειγμα 3 κειμένων, $V(\vec{d}_1)$, $V(\vec{d}_2)$, $V(\vec{d}_3)$ διατεταγμένα ως προς το συνημίτονο της μεταξύ τους γωνίας

Κεφάλαιο 4

Ανάλυση Γραφημάτων

4.1 Εισαγωγή

Τα ελεύθερα πακέτα λογισμικού ανάλυσης δικτύων και γράφων NodeXL [νοδ] και Gephi [γεπ] που χρησιμοποιήθηκαν για την σχεδίαση των γραφημάτων στην Ενότητα 5 παρέχουν μια πληθώρα επιλογών αλγορίθμων εποπτικής αναπαράστασης γραφημάτων. Στο Κεφάλαιο αυτό θα εστιάσουμε σε δύο από αυτούς, στον αλγόριθμο ***Fruchterman-Reingold*** [Φρυχτερμαν ανδ Ρεινγολδ] και τον αλγόριθμο ***Force Atlas 2*** [Θαχομψ ετ αλ.]. Τα γραφήματα χρησιμεύουν ως μαθηματικά μοντέλα για την επιτυχή εποπτική παρακολούθηση δεδομένων σε πραγματικό χρόνο. Ορισμένα προβλήματα από διάφορα επιστημονικά πεδία όπως φυσική, χημεία, τηλεπικοινωνίες, γενετική, επιστήμη υπολογιστών, ψυχολογία, κοινωνιολογία και γλωσσολογία μπορούν να διατυπωθούν ως προβλήματα γραφημάτων. Επίσης πολλοί κλάδοι των Μαθηματικών, όπως η θεωρία ομάδων, θεωρία πινάκων, πιθανότητες και τοπολογία, έχουν στενή σχέση με την θεωρία Γραφημάτων. Το διάσημο πρόβλημα της γέφυρας του Konigsberg αποτέλεσε την έμπνευση για την ανάπτυξη της θεωρίας Γραφημάτων κατά Euler. Η αρκετά δύσκολη και προκλητική θωρία Hamiltonian γραφημάτων από το παιχνίδι "*Around the World*" του Sir William Hamilton. Η θεωρία των ακυκλικών γραφημάτων αναπτύχθηκε για την επίλυση των προβλημάτων των ηλεκτρικών δικτύων. Το γνωστό πρόβλημα των 4 χωμάτων αποτέλεσε την απαρχή της βάσης για την ανάπτυξη της συνδυαστικής τοπολογίας. Προβλήματα γραμμικού προγραμματισμού καθώς και Επιχειρησιακής Έρευνας (π.χ. θαλάσσια κυκλοφορικά προβλήματα) μπορούν να αντιμετωπιστούν με την θεωρία ροών σε δίκτυα. Το πρόβλημα της μαθήτριας του Kirkman καθώς και τα προβλήματα προγραμματισμού μπορούν να επιλυθούν μέσω των γνωστών τεχνικών χρωματισμού γραφημάτων. Πολλά περισσότερα τέτοια προβλήματα μπορούν να προστεθούν στον κατάλογο αυτό. [12]

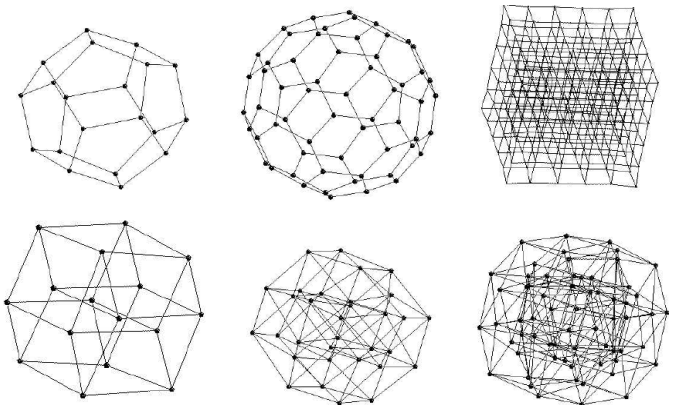
4.2 Αλγόριθμος Fruchterman-Reingold

Οι μέθοδοι δύναμης force directed είναι από τις πιο ευέλικτες μεθόδους για τον υπολογισμό διατάξεων απλών μη κατευθυνόμενων γραφημάτων, είναι γνωστοί και ως ενσωματωτές ελατηρίων (spring embedders). Οι αλγόριθμοι αυτοί υπολογίζουν τη διαμόρφωση (layout) ενός γραφήματος χρησιμοποιώντας μόνο πληροφορίες που περιέχονται μέσα στη δομή του ίδιου του γραφήματος, αντί να στηρίζονται σε εξειδικευμένη γνώση (domain-specific knowledge). Γραφήματα που σχεδιάζονται μέσω αυτών των αλγορίθμων τείνουν να είναι αισθητικά ευχάριστοι, παρουσιάζουν συμμετρίες, και τείνουν να παράγουν διατάξεις ελεύθερης διέλευσης (crossing-free) για επίπεδα γραφήματα.

Ο αλγόριθμος των Fruchterman και Reingold [Φρυςητερμαν ανδ Ρεινγολδ] βασίζεται σε δυνάμεις ελατηρίων, παρόμοιες με εκείνες του νόμου του Hooke. Σ' αυτές τις μεθόδους, υπάρχουν απωθητικές δυνάμεις μεταξύ όλων κόμβων, αλλά επίσης και ελκτικές δυνάμεις μεταξύ γειτονικών κόμβων. Εναλλακτικά, οι δυνάμεις μεταξύ των κόμβων μπορούν να υπολογιστούν βάσει των θεωρητικών αποστάσεων των γραφημάτων, που καθορίζονται από το μήκος των μικρότερων μονοπατιών μεταξύ των κόμβων. Σε γενικές γραμμές, οι μέθοδοι δύναμης ορίζουν μια αντικειμενική συνάρτηση που αντιστοιχίζει κάθε διαμόρφωση του γραφήματος σε έναν θετικό πραγματικό αριθμό που αντιπροσωπεύει την ενέργεια της διαμόρφωσης. Αυτή η συνάρτηση ορίζεται κατά τέτοιο, τρόπο ώστε οι χαμηλές ενέργειες να αντιστοιχούν σε διαμορφώσεις στις οποίες γειτονικοί κόμβοι είναι κοντά ως προς προκαθορισμένη απόσταση, και στην οποία οι μη γειτονικοί κόμβοι είναι καλά διαχωρισμένοι μεταξύ τους. Στη συνέχεια μια διαμόρφωση για ένα γράφημα υπολογίζεται με την εύρεση ενός συχνά τοπικού ελαχίστου της αντικειμενικής συνάρτησης. (βλ. Σχήμα 4.1). Η χρηστικότητα της μεθόδου της δύναμης περιορίζεται σε μικρά γραφήματα και τα αποτελέσματα είναι φτωχά όσον αφορά γραφήματα με μερικές εκατοντάδες κορυφές. Υπάρχουν πολλοί λόγοι για τους οποίους οι παραδοσιακοί αλγόριθμοι δύναμης δεν αποδίδουν καλά σε μεγάλα γραφήματα. Ένα από τα κύρια εμπόδια για την επεκτασιμότητα αυτών των προσεγγίσεων είναι το γεγονός ότι το φυσικό μοντέλο τυπικά έχει πολλά τοπικά ελάχιστα.

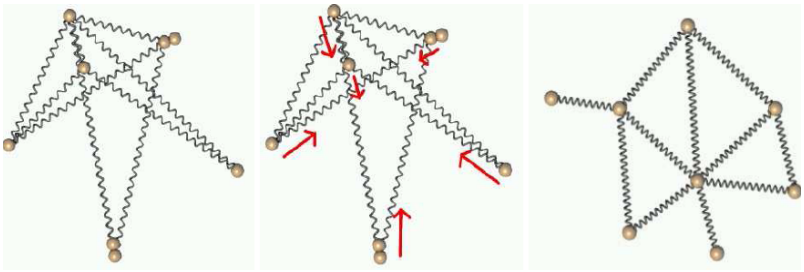
Ακόμη και με τη βοήθεια εξελιγμένων μηχανισμών για την αποφυγή τοπικών ελαχίστων οι βασικοί αλγόριθμοι δύναμης δεν είναι σε θέση να παράγουν σταθερά καλές διαμορφώσεις για μεγάλα γραφήματα. Οι βαρυκεντρικές μέθοδοι, επίσης, δεν αποδίδουν καλά σε μεγάλα γραφήματα, κυρίως λόγω των προβλημάτων ανάλυσης. Για μεγάλα γραφήματα, ο διαχωρισμός της ελάχιστης κορυφής τείνει να είναι πολύ μικρός,

με αποτέλεσμα την παραγωγή δυσανάγνωστων γραφημάτων.



Σχήμα 4.1: Παραδείγματα διαμορφώσεων από τον αλγόριθμο δύναμης. Πρώτη σειρά: μικροί γράφοι: δωδεκάεδρο (20 κορυφές), C60 bucky ball (60 κορυφές), 3D cube mesh (210 κορυφές). Δεύτερη σειρά: Κύβοι σε 4,5,6 διαστάσεις [2].

Τα τέλη του 1990 εμφανίστηκαν πολλές τεχνικές για την επέκταση της λειτουργίας των μεθόδων δύναμης σε γραφήματα με δεκάδες χιλιάδες, ακόμη και εκατοντάδες χιλιάδες κορυφές. Ένα κοινό στοιχείο σε αυτές τις προσεγγίσεις είναι η πολυεπίπεδη τεχνική διαμόρφωσης. Το γράφημα αντιπροσωπεύεται από μια σειρά από προοδευτικά απλούστερες δομές οι οποίες ορίζονται με αντίστροφη σειρά: από την πιο απλή μέχρι την πιο σύνθετη. Οι κλασικοί αλγόριθμοι δύναμης περιορίζονται στον υπολογισμό της διαμόρφωσης ενός γραφήματος μέσω της Ευκλείδειας Γεωμετρίας, τυπικά στον $\mathbb{R}^2$, $\mathbb{R}^3$ και πιο πρόσφατα $\mathbb{R}^n$ για μεγαλύτερες τιμές του n. Υπάρχουν, ωστόσο, περιπτώσεις όπου Ευκλείδεια Γεωμετρία μπορεί να μην είναι η καλύτερη επιλογή. Ορισμένα γραφήματα όπως γνωρίζουμε έχουν μία δομή η οποία θα μπορούσε να υλοποιηθεί καλύτερα σε μια διαφορετική Γεωμετρία, όπως στην επιφάνεια μιας σφαίρας ή ενός τοροειδούς. Ειδικότερα, τα δεδομένα 3D πλεγμάτων μπορούν να παραμετροποιηθούν στη σφαίρα για την αντιστοίχιση υφής. Επιπλέον, έχει επίσης παρατηρηθεί ότι ορισμένες μη-Ευκλείδειες Γεωμετρίες, συγκεκριμένα η Υπερβολική Γεωμετρία, έχουν ιδιότητες οι οποίες είναι ιδιαίτερα καλά προσαρμοσμένες για τη διαμόρφωση και την εποπτική αναπαράσταση γραφημάτων μεγάλων τάξεων [13, 14]. Πλέον οι Ευκλείδειοι αλγόριθμοι δύναμης έχουν επεκταθεί σε χώρους Riemann [15].



Σχήμα 4.2: Παράδειγμα ενός γενικού ενσωματωτή ελατηρίου (spring embedder). Ξεκινώντας από τυχαίες θέσεις,το γράφημα αντιμετωπίζεται ως ένα σύστημα ελατηρίων και αναζητείται μια σταθερή διαμόρφωση [3].

4.2.1 Συστήματα Ελατηρίων & Ηλεκτρικές Δυνάμεις

Ο αλγόριθμος του Eades [16] στοχεύει σε γραφήματα με 30 κορυφές το πολύ και χρησιμοποιεί ένα μηχανικό μοντέλο για την παραγωγή "αισθητικά ευχάριστων"2D διαμορφώσεων. Ο αλγόριθμος συνοπτικά συνοψίζονται ως εξής:

Για την ενσωμάτωση ενός γραφήματος αντικαθιστούμε τις κορυφές με δαχτυλίδια χάλυβα καθώς και κάθε ακμή με ελατήρια για να σχηματίσουμε ένα μηχανικό σύστημα (βλ. Σχήμα 4.2). Οι κορυφές τοποθετούνται σε κάποια αρχική διάταξη και αφήνονται, έτσι ώστε οι δυνάμεις των ελατηρίων στα δαχτυλίδια να μετακινήσουν το σύστημα σε μια ελάχιστη ενεργειακή κατάσταση. Έγιναν δύο πρακτικές προσαρμογές σε αυτή την ιδέα: ***Πρώτον***, χρησιμοποιήθηκαν ελατήρια λογαριθμικής δύναμης. Δηλαδή, η δύναμη που ασκείται από ένα ελατήριο είναι: $c_1 \cdot \log\left(\frac{d}{c_2}\right)$, όπου d είναι το μήκος του ελατηρίου, και c_1 και c_2 είναι σταθερές. Η εμπειρία δείχνει ότι τα (γραμμικά) ελατήρια του νόμου του Hooke είναι πολύ ισχυρά, όταν οι κορυφές είναι πολύ μακριά. Η λογαριθμική δύναμη λύνει αυτό το πρόβλημα. Σημειώστε ότι τα ελατήρια δεν ασκούν καμία δύναμη ισχύος, όταν $d = c_2$. ***Δεύτερον***, ορίστηκαν μη γειτονικές κορυφές να απωθούνται. Για την επίτευξη αυτού του στόχου είναι κατάλληλος ένας αντίστροφος νόμος τετραγωνικής ισχύος, $\frac{c_3}{\sqrt{d}}$, όπου c_3 είναι μια σταθερά και d είναι η απόσταση μεταξύ των κορυφών.

Algorithm 1: SPRING

Input: Graph G
Output: Straight-line drawing of G
Initialize Positions: place vertices of G in random locations;
for $i = 1$ **to** M **do**
    calculate the force acting on each vertex;
    move the vertex $c_4 *$ (force on vertex);
draw a filled circle for each vertex;
draw a straight-line segment for each edge;

Σχήμα 4.3: Περιγραφή αλγορίθμου ελατηρίου (spring algorithm)

Αυτή η περιγραφή συμπυκνώνει την ουσία των αλγορίθμων ελατηρίων καθώς και τη φυσική απλότητα τους. Οι στόχοι πίσω από τις "αισθητικά ευχάριστες"διαμορφώσεις είχαν αρχικά συλληφθεί από τα δύο ακόλουθα κριτήρια: "όλα τα μήκη των ακμών θα έπρεπε να είναι ίδια", και "η διάταξη θα πρέπει να εμφανίζει όσο το δυνατόν περισσότερη συμμετρία". Ο αλγόριθμος των Fruchterman και Reingold είχε ως προσθήκη την "ίση κατανομή κορυφών"στα προηγούμενα δύο κριτήρια. Αντιμετωπίζει επίσης τις κορυφές στο γράφημα ως "ατομικά σωματίδια ή ουράνια σώματα, ασκώντας ελκτικές

και απωθητικές δυνάμεις μεταξύ τους". Οι ελκτικές f_a και απωθητικές δυνάμεις f_r ορίζονται ως:

$$f_a(d) = \frac{d^2}{k}, \qquad f_r(d) = \frac{-k^2}{d}. \tag{4.1}$$

σε όρους της απόστασης d ανάμεσα σε δύο κορυφές, και της βέλτιστης απόστασης ανάμεσα σε k κορυφές που ορίζεται ως:

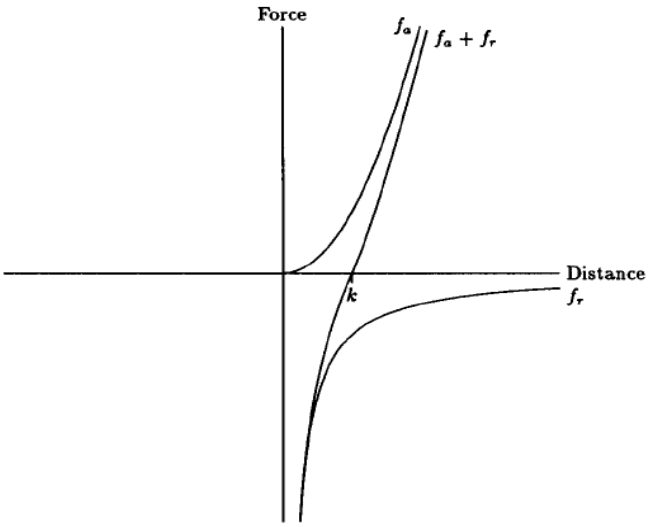
$$k = C\sqrt{\frac{area}{number\ \ of\ \ vertices}} \tag{4.2}$$

Ο αλγόριθμος, όπως προαναφέραμε, υπολογίζει ελκτικές δυνάμεις μεταξύ γειτονικών κορυφών και απωθητικές δυνάμεις μεταξύ όλων των ζευγών των κορυφών. Επιπλέον ο αλγόριθμος Fruchterman - Reingold προσθέτει την έννοια της "θερμοκρασίας" η οποία θα μπορούσε να χρησιμοποιηθεί ως εξής: "η θερμοκρασία θα μπορούσε να λάβει μια αρχική τιμή (δηλαδή το ένα δέκατο του πλάτους του πλαισίου) και να φθίνει στο 0 με έναν αντίστροφα γραμμικό τρόπο". Η θερμοκρασία ελέγχει τη μετατόπιση των κορυφών έτσι ώστε καθώς η διαμόρφωση γίνεται καλύτερη, οι προσαρμογές να γίνονται μικρότερες. Η χρήση της θερμοκρασίας εδώ είναι μια ειδική περίπτωση μιας γενικής τεχνικής που ονομάζεται προσομοιωμένη ανόπτηση (simulated annealing). Ο ψευδοκώδικας του αλγόριθμου Fruchterman - Reingold παρατίθεται αναλόγως στο Σχήμα 4.4.

Algorithm 2: Fruchterman-Reingold

```
area ← W * L ;                                   /* frame: width W and length L */
initialize G = (V, E) ;                          /* place vertices at random */
k ← sqrt(area/|V|) ;                             /* compute optimal pairwise distance */
function f_r(x) = k^2/x ;                        /* compute repulsive force */
for i = 1 to iterations do
    foreach v ∈ V do
        v.disp := 0; ;                           /* initialize displacement vector */
        for u ∈ V do
            if (u ≠ v) then
                Δ ← v.pos − u.pos ;              /* distance between u and v */
                v.disp ← v.disp + (Δ/|Δ|) * f_r(|Δ|) ;   /* displacement */
    function f_a(x) = x^2/k ;                    /* compute attractive force */
    foreach e ∈ E do
        Δ ← e.v.pos − e.u.pos ;                  /* e is ordered vertex pair .v and .u */
        e.v.disp ← e.v.disp − (Δ/|Δ|) * f_a(|Δ|);
        e.u.disp ← e.u.disp + (Δ/|Δ|) * f_a(|Δ|);
    foreach v ∈ V do
                /* limit max displacement to frame; use temp. t to scale */
        v.pos ← v.pos + (v.disp/|v.disp|) * min(v.disp, t);
        v.pos.x ← min(W/2, max(−W/2, v.pos.x));
        v.pos.y ← min(L/2, max(−L/2, v.pos.y));
    t ← cool(t) ;                                /* reduce temperature for next iteration */
```

Σχήμα 4.4: Περιγραφή αλγόριθμου Fruchterman - Reingold [3].



Σχήμα 4.5: Μεταβολή των σημείων ως προς την απόσταση

Το Σχήμα 4.5 απεικονίζει τις ελκτικές f_a και απωθητικές f_r δυνάμεις καθώς και το άθροισμα τους ως προς τις αποστάσεις όπως ορίστηκε στην (4.2). Το σημείο όπου το άθροισμα των ελκτικών και απωθητικών δυνάμεων τέμνει το άξονα, x είναι εκεί όπου οι δυο δυνάμεις ακυρώνονται μεταξύ τους, και αυτό είναι το k, η ιδανική απόσταση ανάμεσα στις κορυφές.

Σε κάθε επανάληψη ο βασικός αλγόριθμος υπολογίζει $O(|E|)$ ελκτικές δυνάμεις και $O(|V|^2)$ αποκρουστικές δυνάμεις, όπου $|E|$ δηλώνει τον αριθμό των ακμών και $|V|$ δηλώνει τον αριθμό των κορυφών. Για να μειωθεί η τετραγωνική πολυπλοκότητα των απωθητικών δυνάμεων, οι Fruchterman και Reingold προτείνουν τη χρήση μιας παραλλαγής του βασικού αλγόριθμου τους, όπου οι απωθητικές δυνάμεις μεταξύ απομακρυσμένων κορυφών δεν λαμβάνονται υπόψη. Για αραιά γραφήματα με ομοιόμορφη κατανομή των κορυφών, αυτή η μέθοδος επιτρέπει $O(|V|)$ χρόνο για υπολογισμό των απωθητικών δυνάμεων.

4.3 Αλγόριθμος Force Atlas 2

4.3.1 Ανατομία του Force Atlas 2

Ο Force Atlas 2 είναι ένας αλγόριθμος διαμόρφωσης κατευθυνόμενης ισχύος [Θαχομψ ετ αλ.]: Προσομοιώνει ένα φυσικό σύστημα. Οι κόμβοι απωθούνται μεταξύ τους σαν μαγνήτες, ενώ οι ακμές προσελκύουν τους κόμβους που συνδέουν σαν ελατήρια. Αυτές οι δυνάμεις δημιουργούν μια κίνηση που συγκλίνει σε μια κατάσταση

ισορροπίας. Αυτή η τελική διαμόρφωση αναμένεται να βοηθήσει στην ερμηνεία των δεδομένων. Υπάρχει τουλάχιστον ένα θέμα με αυτή τη στρατηγική: Όλα τα γραφήματα δεν συγκλίνουν πάντα στην ίδια τελική διαμόρφωση.

Το αποτέλεσμα εξαρτάται από τις δυνάμεις που εφαρμόζονται, αλλά επίσης και από την αρχική κατάσταση και ακόμη και τις προσεγγίσεις του αλγορίθμου. Η διαδικασία δεν είναι ντετερμινιστική, και οι συντεταγμένες κάθε σημείου δεν αντανακλούν καμία συγκεκριμένη μεταβλητή. Το αποτέλεσμα είναι πολύ διαφορετικό από μια καρτεσιανή προβολή. Δεν μπορεί να διαβαστεί η θέση ενός κόμβου. Θα πρέπει να συγκριθεί η θέση του με τους άλλους κόμβους. Παρά το θέμα αυτό, η τεχνική έχει ένα μοναδικό πλεονέκτημα: ενισχύει σε μεγάλο βαθμό την εποπτική ερμηνεία του γραφήματος. Η αρχή λειτουργίας του συνίσταται στο να μετατρέψει τις διαρθρωτικές προσεγγίσεις σε οπτικές, διευκολύνοντας την ανάλυση γραφημάτων και ειδικότερα την ανάλυση κοινωνικών δικτύων. Ο αλγόριθμος διαφοροποιείται από τις άλλες διαμορφώσεις κατευθυνόμενης ισχύς, λόγω δύο κύριων χαρακτηριστικών:

1) Η ταυτότητα ενός αλγορίθμου διάταξης κατευθυνόμενης ισχύος βρίσκεται πρώτα απ΄όλα στο "ενεργειακό μοντέλο"του, δηλαδή στον τρόπο που υπολογίζει απωθητικές και ελκτικές δυνάμεις. Υπάρχουν διαφορετικοί τύποι και ο καθένας παράγει διαφορετικά σχήματα για το ίδιο γράφημα. Το ενεργειακό μοντέλο του Force Atlas 2 βρίσκεται μεταξύ του Noack's LinLog [17] και του Fruchterman - Rheingold [Φρυχτερμαν ανδ Ρεινγολδ], με μια ουσιώδη διαφορά: Η απώθηση εξαρτάται από τον βαθμό των κόμβων (δηλαδή τον αριθμό των συνδέσεων ενός κόμβου προς του υπόλοιπους).

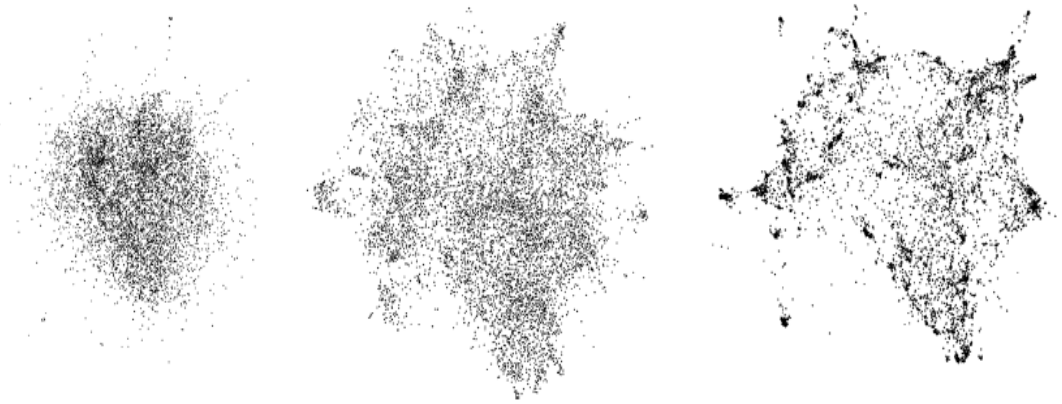
2) Σε έναν αλγόριθμο κατευθυνόμενης ισχύος, η διαμόρφωση εφαρμόζεται βήμα προς βήμα στο γράφημα. Αυτά τα βήματα προσεγγίζουν τις κινήσεις που οδηγούν στην ισορροπία. Σε κάθε στάδιο υπολογίζονται οι δυνάμεις, και οι κόμβοι μετατοπίζονται ως αποτέλεσμα αυτών των δυνάμεων. Η τεχνική αυτή υπαινίσσεται ένα συμβιβασμό μεταξύ ταχύτητας σε σχέση με την ακρίβεια. Προκειμένου να επιτευχθεί η ισορροπία πιο γρήγορα μπορεί να επιταχυνθεί η κίνηση των κόμβων, αλλά αυτό δημιουργεί και ανακρίβειες, όπως κόμβους που ταλαντεύονται και αδυνατούν να καταλήξουν σε μια ισορροπημένη θέση. Ο Force Atlas 2 εκτιμά συνεχώς την κατάσταση προς μεγιστοποίηση της ταχύτητας και την ελαχιστοποίηση της ταλάντευσης.

Παρά τις δύο αυτές καινοτομίες και άλλες που παρουσιάζονται ακολούθως, ο Force Atlas 2 επιμένει στον παραδοσιακό σχεδιασμό διατάξεων κατευθυνόμενης ισχύος. Εκτελείτε συνεχώς και ομοιόμορφα μέχρι ο χρήστης να τον σταματήσει. Σε αντίθεση με άλλους πρόσφατους αλγόριθμους, σέβεται την αίσθηση ενός φυσικού συστήματος, εμφυσώντας ζωή στο δίκτυο, όπως ένα σύστημα Φυσικής σε ένα βιντεοπαιχνίδι. Με

αυτό τον τρόπο οι χρήστες μπορούν να κατανοήσουν ένα δίκτυο απλά παίζοντας με αυτό.

4.3.2 Ενεργειακό Μοντέλο

Κάθε αλγόριθμος κατευθυνόμενης ισχύος βασίζεται σε μια συγκεκριμένη εξίσωση για την ελκτική δύναμη και μια συγκεκριμένη εξίσωση για την απωθητική δύναμη. Η διαμόρφωση ηλεκτρικών ελατηρίων (spring-electric) [16] είναι μια προσομοίωση εμπνευσμένη από την πραγματική ζωή. Χρησιμοποιεί τον τύπο απώθησης των ηλεκτρικά φορτισμένων σωματιδίων $\left(F_r = \frac{k}{d^2}\right)$ και τον τύπο έλξης των ελατηρίων $\left(F_a = \frac{-k}{d}\right)$ που αφορούν την απόσταση d μεταξύ δύο κόμβων. Όπως και στα φυσικά συστήματα, οι δυνάμεις αυτές είναι ανάλογες προς την απόσταση μεταξύ των αλληλεπιδρουσών οντοτήτων. Οντότητες που βρίσκονται κοντά ελκύουν λιγότερο και απωθούν περισσότερο από οντότητες που βρίσκονται μακρύτερα και αντιστρόφως. Η δύναμη μπορεί να είναι συνάρτηση της απόστασης, γραμμική, εκθετική ή λογαριθμική. Το μοντέλο ελατηρίων, για παράδειγμα, αναπαράγει επακριβώς τις φυσικές δυνάμεις από τις οποίες είναι εμπνευσμένο, καθιερώνοντας έτσι μια γραμμική αναλογία μεταξύ απόστασης και δύναμης όπως η έλξη ελατηρίου καθώς και μια τετραγωνική αναλογικότητα μεταξύ της απόστασης και της δύναμης, όπως η ηλεκτρομαγνητική απώθηση. Το ενεργειακό μοντέλο του Force Atlas 2 έχει μια ενδιάμεση θέση μεταξύ του Noack's LinLog [17] και του Fruchterman-Rheingold [Φρυςητερμαν ανδ Ρεινγολδ]. Ο Noack [18] αναφέρει ότι οι "αποστάσεις εξαρτώνται λιγότερο από τις πυκνότητες για ένα μεγάλο $a-r$ όπου a δηλώνει έλξη και r δηλώνει απώθηση, και είναι λιγότερο εξαρτώμενες από μήκη διαδρομής για μικρές τιμές του a". Αυτό σημαίνει ότι η πυκνότητα κόμβων (όπου οι κόμβοι συγκεντρώνονται) υποδηλώνει πυκνά συνδεδεμένα υπο-γραφήματα όταν το $a-r$ λαμβάνει χαμηλή τιμή, δηλαδή όταν η ελκτική δύναμη εξαρτάται λιγότερο από την απόσταση, ενώ αντίστροφα η δύναμη απώθησης εξαρτάται περισσότερο από αυτή. Ο αλγόριθμος Force Atlas 2 είναι σαφώς καλύτερος από τον αλγόριθμο Fruchterman-Rheingold και χειρότερος από τον LinLog για να αναδείξει ομάδες σε έναν γράφημα.



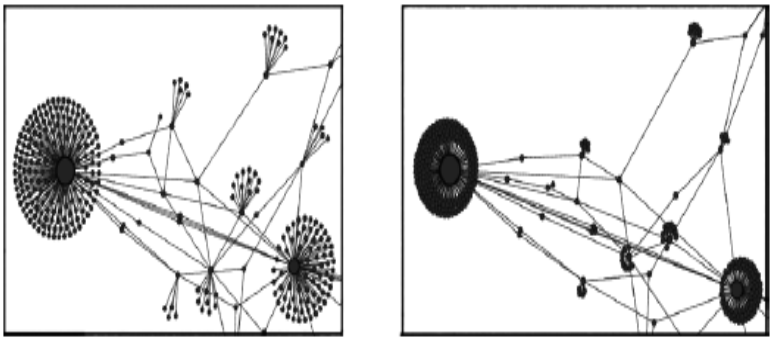
Σχήμα 4.6: Διατάξεις για Fruchterman-Reingold ($a-r=3$), ForceAtlas2 ($a-r=2$), LinLog λειτουργία του ForceAtlas2 ($a-r=1$) [Θαςομψ ετ αλ.].

4.3.3 Κλασική Δύναμη Έλξης

Η δύναμη έλξης F_a μεταξύ δύο συνδεδεμένων κόμβων n_1 και n_2 δεν έχει τίποτα το αξιοσημείωτο. Εξαρτάται γραμμικά ως προς την απόσταση $d(n_1, n_2)$. Παρατηρείστε ότι στον συγκεκριμένο τύπο δεν υπάρχει κάποια σταθερά "c" η οποία να προσαρμόζει την δύναμη όπως στο Fruchterman-Reingold.

4.3.4 Απώθηση Ανάλογα με τον Βαθμό Διασύνδεσης Κόμβων

Ο ForceAtlas2 σχεδιάστηκε ώστε να ερμηνεύει τα γραφήματα στο διαδίκτυο και τα κοινωνικά δίκτυα. Ένα κοινό χαρακτηριστικό αυτών των δικτύων είναι η παρουσία πολλών κόμβων που έχουν μόνο ένα γείτονα. Αυτό οφείλεται στην κατανομή νόμου δύναμης των βαθμών διασύνδεσης των κόμβων που χαρακτηρίζει πολλά πραγματικά δεδομένα. Το δάσος των κόμβων που έχουν μόνο ένα γείτονα (leaves) που περιβάλλει τους λίγους, αλλά ιδιαίτερα συνδεδεμένους κόμβους είναι μία από τις κύριες πηγές της οπτικής σύγχυσης. Για τη βελτίωση της αναγνωσιμότητας αυτών των γραφημάτων, προτείνεται η τροποποιήσει του τύπου της δύναμης απώθησης.



Σχήμα 4.7: Διάταξη Fruchterman-Reingold στα αριστερά (κανονικός τύπος απώθησης) και ForceAtlas2 στα δεξιά (τύπος απώθησης κατά βαθμό διασύνδεσης κόμβων). Παρόλο που η συνολική εικόνα παραμένει αμετάβλητη, κακώς συνδεδεμένοι κόμβοι βρίσκονται πιο κοντά σε αυτούς με υψηλή συνδεσιμότητα [Θαξομψ ετ αλ.].

Η ιδέα βασίζεται στην τοποθέτηση κακώς συνδεδεμένων κόμβων κοντά σε κόμβους με πολύ υψηλή συνδεσιμότητα. Η λύση που προτείνει ο ForceAtlas2 είναι μια μετατροπή της απωθητικής δύναμης έτσι ώστε κόμβοι με χαμηλή ή κακή συνδεσιμότητα και κόμβοι με υψηλή συνδεσιμότητα να απωθούνται λιγότερο. Κατά συνέπεια, θα καταλήξουν να βρίσκονται πιο κοντά στην ισορροπημένη κατάσταση του γραφήματος (βλ. Σχήμα 4.7). Η απωθητική δύναμη F_r είναι ανάλογη με την παραγωγή των διασυνδέσεων συν ένα των δύο κόμβους. Ο συντελεστής k_r είναι μια σταθερά που ορίζεται από το χρήστη:

$$F_r\left(n_1, n_2\right) = k_r \frac{\left(deg\left(n_1\right)+1\right)\left(deg\left(n_2\right)+1\right)}{d\left(n_1, n_2\right)}. \tag{4.3}$$

Ο συγκεκριμένος τύπος μοιάζει πολύ με αυτόν που προτάθηκε από τον Noack [17] εκτός από το γεγονός ότι εκείνος χρησιμοποιεί το βαθμό και όχι τον προσαυξημένο κατά μια μονάδα βαθμό κόμβων. Στον τύπο του αλγόριθμου ForceAtlas2, το +1 είναι σημαντικό, καθώς εξασφαλίζει ότι ακόμη και κόμβοι με μηδενικό βαθμό εξακολουθούν να έχουν κάποια δύναμη απώθησης. Αυτή η βαθμονόμηση της δύναμης απώθησης (που αποτέλεσε την ταυτότητα του ForceAtlas2 από την αρχή) είναι τουλάχιστον εξίσου σημαντική όσο και το μοντέλο έλξης απώθησης που εφαρμόστηκε στον ForceAtlas2 και ίσως περισσότερο δεδομένου ότι συμβάλει εντυπωσιακά στη βελτίωση της εποπτείας των γραφημάτων.

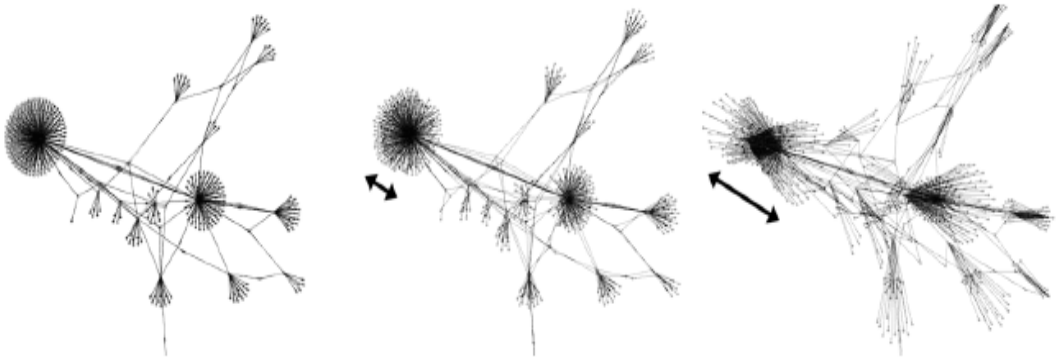
4.3.5 Αυτόματη Προσέγγιση Ταχύτητας έναντι Ακρίβειας

4.3.5.1 Το Πρόβλημα της Ταχύτητας

Όταν χρησιμοποιείται μια διαμόρφωση γραφήματος που είναι βασισμένη στην ισχύ, τότε συνήθως οι χρήστες βρίσκονται αντιμέτωποι με την ανταλλαγή ταχύτητας/ακρίβειας. Σε έναν κλασικό αλγόριθμο κατευθυνόμενης ισχύος, αύξηση στην ταχύτητα επιφέρει μείωση της ακρίβειας. Η επίδραση της προσέγγισης είναι ότι ορισμένοι κόμβοι (μερικές φορές το σύνολο του γραφήματος) αδυνατεί να βρει μια σταθερή κατάσταση οπότε αρχίζει να ταλαντεύεται γύρω από τη θέση ισορροπίας τους (βλέπε Σχήμα 4.8). Θεωρητικά, η χωρική ταχύτητα θα πρέπει να επιλεγεί ανάλογα με το μέγεθος και τις ιδιότητες του γραφήματος, σύμφωνα με το διαθέσιμο χρόνο και τον χωρικό στόχο. Σε πρακτικό επίπεδο, είναι εφικτό να βρεθεί η σωστή ταχύτητα, αυξάνοντας την μέχρι τη στιγμή όπου θα αρχίσουν να εμφανίζονται προβλήματα ακρίβειας. Δυστυχώς, οι περισσότεροι χρήστες δεν χρησιμοποιούν αυτήν την τεχνική, κυρίως διότι αλλοιώνονται τα αποτελέσματα από την απώλεια της ακρίβειας. Ο αλγόριθμος ForceAtlas2 αναπτύχθηκε ώστε να εκτελεί αυτή την τεχνική αυτόματα χωρίς να απαιτείται καμία παρέμβαση από το χρήστη.

Με τον αλγόριθμο Fruchterman-Reingold οι χρήστες πρέπει μόνο να επιλέξουν πόση ταλάντωση είναι διατεθειμένοι να ανεχθούν καθώς η ταχύτητα υπολογίζεται συνεχώς ώστε να πληρεί αυτή την απαίτηση.

Στο ελεύθερο διαθέσιμο λογισμικό ανάλυσης γράφων Gephi [γεπ], υπάρχουν τρεις προεπιλεγμένες τιμές για την "Ανοχή (ταχύτητα)"/"Tolerance (speed)"παράμετρος: 0,1 με 5000 κόμβους, 1 έως 50000 κόμβων και 10 για μεγαλύτερο πλήθος κόμβων.



Σχήμα 4.8: Διάταξη Fruchterman-Reingold σε ταχύτητες 100, 500 και 2500 (εφαρμογή σε δύο διαδοχικά στάδια). Η ταλάντωση των κόμβων αυξάνεται με την ταχύτητα [Θαςομψ ετ αλ.].

4.3.5.2 Προσαρμογή της Τοπικής Ταχύτητας

Προκειμένου να ελαχιστοποιηθεί η ταλάντευση $swg\,(n)$ κάθε κόμβου, δηλαδή η απόκλιση μεταξύ της δύναμης που εφαρμόζεται σε έναν κόμβο σε ένα δεδομένο βήμα και της δύναμης που εφαρμόζεται στον ίδιο κόμβο στο προηγούμενο βήμα. Διαισθητικά, όσο περισσότερο ο κόμβος καλείται να αλλάξει κατεύθυνση, τόσο περισσότερο ταλαντεύεται. Έστω $F_t\,(n)$ η τελική δύναμη που εφαρμόζεται στο κόμβο n στο βήμα t. Τότε:

$$swg_{(t)}\,(n) = |F_{(t)}\,(n) - F_{(t-1)}\,(n)\,|. \tag{4.4}$$

Για έναν κόμβο που κινείται προς τη θέση ισορροπίας, $swg_{(t)}\,(n)$ θα παραμείνει κοντά στο μηδέν. Το αντίθετο ισχύει για έναν κόμβο που δεν συγκλίνει. Θα έχει υψηλές τιμές ταλάντευσης και η κίνηση του θα πρέπει να επιβραδυνθεί για να βοηθηθεί να βρεί τη θέση ισορροπίας του. Η ταχύτητα $s\,(n)$ ενός κόμβου n καθορίζει πόση μετατόπιση $D\,(n)$, θα πρέπει να προκαλείται από την προκύπτουσα δύναμη $F\,(n) : D\,(n) = s\,(n)\,F\,(n)$. Η συνισταμένη δύναμη είναι το άθροισμα όλων των δυνάμεων που εφαρμόζεται σε κάθε κόμβο (έλξη, απώθηση και βαρύτητα: $F = F_a + F_r + F_g$). Ενώ στους περισσότερους αλγορίθμους κατευθυνόμενης ισχύος η ταχύτητα συνήθως είναι γενική και καθορίζεται από το χρήστη, στον αλγόριθμο ForceAtlas2 η ταχύτητα είναι διαφορετική για κάθε κόμβο, και υπολογίζεται ως ακολούθως:

$$s\,(n) = \frac{k_s s\,(G)}{1 + s\,(G)\,\sqrt{swg\,(n)}}. \tag{4.5}$$

όπου $s\,(G)$ είναι η συνολική ταχύτητα του γραφήματος, k_s είναι μια σταθερά ορισμένη

ίση με 0.1 στο Gephi. Όσο περισσότερο ταλαντεύεται ένας κόμβος, τόσο περισσότερο επιβραδύνεται. Αν δεν υπάρχει ταλάντευση, ο κόμβος κινείται με την καθολική ταχύτητα του γράφου. Ως προστασία, ο ForceAtlas2 υλοποιεί έναν επιπρόσθετο περιορισμό που εμποδίζει την τοπική ταχύτητα να είναι πάρα πολύ υψηλή, ακόμη και στην περίπτωση των πολύ υψηλών καθολικών ταχυτήτων:

$$s\left(n\right) < \frac{k_{smax}}{\mid F\left(n\right) \mid}. \tag{4.6}$$

4.3.5.3 Προσαρμογή της Καθολικής Ταχύτητας

Σε κάθε βήμα, υπολογίζονται δύο καθολικές τιμές και χρησιμοποιούνται για να καθοριστεί η καθολική/συνολική ταχύτητα. Η συνολική ταλάντευση $swg\,(G)$ και η συνολική αποτελεσματική πρόσφυση. Η συνολική ταλάντευση $swg\,(G)$ αντιπροσωπεύει την ποσότητα της ασταθούς κίνησης που βρίσκεται παρούσα στη συνολική κίνηση του γράφου. Πρόκειται για το άθροισμα των τοπικών τιμών ταλάντευσης, σταθμισμένο με το βαθμό διασύνδεσης του κάθε κόμβου προς τους άλλους, όπως ισχύει στον απωθητικό τύπο δύναμης $(degree + 1)$.

$$swg\,(G) = \sum_{n} (deg\,(n) + 1)\, swg\,(n)\,. \tag{4.7}$$

Η αποτελεσματική έλξη (effective traction) $tra\,(n)$ ενός κόμβου είναι η ποσότητα "χρήσιμης" δύναμης που εφαρμόζεται στο εν λόγω κόμβο. Αντιπροσωπεύει το αντίθετο της ταλάντευσης, επομένως τη σύγκλιση της κίνησης. Ορίζεται ως ο μέσος όρος δυνάμεων σε διαφορετικές χρονικές στιγμές:

$$tra_{(t)}\,(n) = \frac{\mid F_{(t)}\,(n) + F_{(t-1)}\,(n) \mid}{2}. \tag{4.8}$$

Αν ένας κόμβος συνεχίζει στη πορεία του, τότε $tra\,(n) = F\,(n)$. Αν επιστρέψει πίσω στην προηγούμενη θέση του (μια τέλεια ταλάντευση), τότε $tra\,(n) = 0$. Η καθολική αποτελεσματική έλξη (global effective traction) $tra\,(G)$ είναι το σταθμισμένο άθροισμα των αποτελεσματικών κινήσεων των κόμβων:

$$tra\,(G) = \sum_{n} (deg\,(n) + 1)\, tra\,(n)\,. \tag{4.9}$$

Η καθολική ταχύτητα $s\,(G)$ διατηρεί την καθολική ταλάντευση $swg\,(G)$ κάτω από μια ορισμένη αναλογία τ της καθολικής αποτελεσματικής έλξης $tra\,(G)$ και ορίζεται ως εξής:

$$s\,(G) = \tau \frac{tra\,(G)}{swg\,(G)}. \tag{4.10}$$

Ο λόγος τ αντιπροσωπεύει την ανοχή της ταλάντευσης και καθορίζεται από το χρήστη. Η υπερβολική αύξηση της συνολικής ταχύτητας θα μπορούσε να έχει αρνητικές επιπτώσεις. Αυτός είναι και ο λόγος που η αύξηση της συνολικής ταχύτητας $s_{(t)}\,(G)$ έχει μειωθεί στο 50% του προηγούμενου βήματος $s_{(t-1)}\,(G)$.

Κεφάλαιο 5

Πειράματα

5.1 Δημιουργία tf-idf matrix

Μετά το πέρας του σταδίου της προ επεξεργασίας των δεδομένων που αναφέρθηκε στο κεφάλαιο 1, επόμενο μας βήμα είναι η μετατροπή του term document matrix $W^{t\times d}$ σε ένα κανονικοποιημένο πίνακα όπου κάθε εγγραφή θα δηλώνει συχνότητα εμφάνισης του όρου t στο έγγραφο d. Ο λόγος για την κατασκευή ενός τέτοιου πίνακα οφείλεται στην δυνατότητα που μας παρέχει να θέτουμε ερωτήματα προς περαιτέρω διερεύνηση και να εξετάζουμε τις λαμβανόμενες απαντήσεις που αφορούν τα δεδομένα. Απώτερος σκοπός της συγκεκριμένης τεχνικής είναι ο υπολογισμός της βαθμολογίας ανάμεσα σε έναν όρο t του ερωτήματος που θέτουμε και σε ένα έγγραφο d, βασιζόμενη στο βάρος του t στο d.

Term frequency and weighting

Η απλούστερη μέθοδος όπως περιγράφηκε σύντομα στο κεφάλαιο 3 είναι μέσω ανάθεσης βαρών που ισούται με τον αριθμό εμφάνισης της συχνότητας μιας λέξης t σε ένα έγγραφο d. Το συγκεκριμένο σύστημα στάθμισης ονομάζεται term frequency και δηλώνεται ως $tf_{t,d}$. Για ένα έγγραφο d, το σύνολο των βαρών προσδιοριζόμενο από τα παραπάνω tf βάρη, μπορεί να θεωρηθεί ως μια ποσοτική σύνοψη/περίληψη του συγκεκριμένου εγγράφου. Σε αυτή την άποψη ενός εγγράφου, που είναι γνωστή στη βιβλιογραφία ως το μοντέλο του σάκου των λέξεων (bag of words model), η ακριβής διάταξη των όρων σε ένα έγγραφο δεν λαμβάνεται υπόψη, αλλά ο αριθμός εμφάνισης κάθε όρου είναι ουσιαστικός. Το μοντέλο αυτό θα διατηρήσει μόνο τις πληροφορίες σχετικά με τον αριθμό των εμφανίσεων κάθε όρου. Έτσι, το έγγραφο ***"Η Μαίρη είναι πιο γρήγορη από τον Γιάννη"*** είναι, κατά την άποψη αυτή, όμοιο με το έγγραφο ***"Ο Γιάννης είναι πιο γρήγορος από την Μαίρη"***. Παρ΄όλα αυτά,

εξάγεται το συμπέρασμα ότι δύο έγγραφα με παρόμοιες αναπαραστάσεις μοντέλων bag of words είναι παρόμοια σε περιεχόμενο.

Inverse document frequency

Η συχνότητα εμφάνισης όρων που αναφέρθηκε ανωτέρω πάσχει από ένα κρίσιμο πρόβλημα: όλοι οι όροι θεωρούνται εξίσου σημαντικοί όταν πρόκειται για την αξιολόγηση συνάφειας σε ένα ερώτημα. Στην πραγματικότητα, ορισμένοι όροι έχουν μικρή ή καμία διακριτική ισχύ στον καθορισμό συνάφειας. Για παράδειγμα, μία συλλογή εγγράφων σχετικά με την αυτοκινητοβιομηχανία είναι πιθανό να έχουν τον όρο αυτοκίνητο σχεδόν σε κάθε έγγραφο. Για το λόγο αυτό, έχει θεσπιστεί ένας μηχανισμός εξασθένισης της επίδρασης των όρων που εμφανίζονται πάρα πολύ συχνά στη συλλογή ώστε να έχει νόημα ως προς τον προσδιορισμό συνάφειας. Μια ιδέα θα ήταν είναι να μειωθούν τα βάρη των όρων με υψηλή συγκέντρωση συχνότητας, που ορίζεται να είναι ο συνολικός αριθμός των εμφανίσεων ενός όρου στη συλλογή. Η ιδέα θα ήταν να μειωθεί το tf βάρος ενός όρου από έναν παράγοντα που αυξάνεται με τη συχνότητα της συλλογής του. Αντί αυτού, είναι καλύτερη ιδέα να χρησιμοποιηθεί για το σκοπό αυτό η συχνότητα του εγγράφου df_t, που ορίζεται ως ο αριθμός των εγγράφων της συλλογής που περιέχουν έναν όρο t. Αυτό οφείλεται στο γεγονός ότι στην προσπάθεια διάκρισης μεταξύ εγγράφων για σκοπούς βαθμολόγησης, είναι καλύτερο να χρησιμοποιηθεί μια στατιστική κλίμακας εγγράφου (όπως ο αριθμός των εγγράφων που περιέχουν έναν όρο) από το να χρησιμοποιηθεί μια στατιστική κλίμακας συλλογής για τον όρο. Ο λόγος προτίμησης της στατιστικής κλίμακας εγγράφου df από την στατιστική κλίμακας cf είναι ότι συμπεριφέρονται μάλλον διαφορετικά. Ειδικότερα, σε διάφορα παραδείγματα οι cf τιμές για διάφορους όρους είναι περίπου ίσα, αλλά οι df τιμές τους διαφέρουν σημαντικά. Κάτι τέτοιο μπορεί να αλλοιώσει τα αποτελέσματα ενός ενδεχόμενου ερωτήματος.

Συνδυάζοντας τους ανωτέρω ορισμούς του term frequency και inverse document frequency παράγουμε ένα σύνθετο σύστημα ανάθεσης βαρών για κάθε όρο σε κάθε έγγραφο το οποίο ονομάζεται tf-idf. Παρακάτω παραθέτουμε ένα υποσύνολο του σταθμισμένου tf-idf πίνακα διαστάσεων 3815×388 των **388** τραγουδιών της συλλογής.

	Αλατσατιανή	Ανάθεμα σε Σμύρνη	Η χήρα	Ο μάγκας	Τι κρίμα
ΝΤΟΥΣΕΜ	0	0.1756	0	0.0878	0
ΜΩΡ	0	0	0.1698	0	0
ΚΟΡ	0	0	0	0	0.0351
ΧΑΜΗΛΩΣ	0	0.0283	0	0	0
ΑΓΑΠ	0.0526	0	0	0	0
ΑΓΓΕΛ	0	0.1245	0	0	0
ΒΑΡΚΑΔ	0	0	0	0.1354	0
ΓΑΡΥΦΑΛ	0.0261	0	0	0	0
ΠΑΤΡΕΥΤ	0	0	0.1485	0	0
ΔΑΝΕΙΖ	0	0	0	0	0.1854

Πίνακας 5.1: Σταθμισμένος αραιός tf-idf πίνακας $W^{t\times d}$ όπου $t = 3815$ (όροι του λεξιλογίου) & $d = 388$ (τραγούδια της συλλογής)

5.2 Μείωση Διαστάσεων

Η πολυδιάστατη κλιμάκωση, *"Multidimensional scaling (MDS)"* είναι ένα μέσο για την απεικόνιση του επιπέδου της ομοιότητας των μεμονωμένων περιπτώσεων ενός σύνολου δεδομένων. Αναφέρεται σε ένα σύνολο σχετικών τεχνικών συντονισμού (ordination techniques) που χρησιμοποιούνται στην απεικόνιση της πληροφορίας, κυρίως για την απεικόνιση πληροφοριών που περιέχονται σε έναν πίνακα αποστάσεων. Ένας αλγόριθμος MDS στοχεύει στην τοποθέτηση κάθε αντικειμένου στον N-διάστατο χώρο έτσι ώστε οι μεταξύ των αντικειμένων αποστάσεις να διατηρούνται όσο το δυνατόν καλύτερα. Στη συνέχεια σε κάθε αντικείμενο εκχωρούνται συντεταγμένες σε κάθε μία από τις N διαστάσεις. Ο αριθμός των διαστάσεων μιας MDS απεικόνισης N μπορεί να υπερβαίνει τις 2 και έχει οριστεί εκ των προτέρων. Επιλέγοντας $N = 2$ βελτιστοποιεί τις θέσεις του αντικειμένου για ένα δισδιάστατο scatterplot. Στην πολυμεταβλητή ανάλυση, η τεχνική συντονισμού (ordination technique) είναι μια συμπληρωματική μέθοδος ομαδοποίησης (clustering) δεδομένων, και χρησιμοποιείται κυρίως στην διερευνητική ανάλυση των δεδομένων (και όχι σε έλεγχο υποθέσεων).

Η τεχνική συντονισμού διατάσσει τα αντικείμενα που χαρακτηρίζονται από τιμές σε πολλαπλές μεταβλητές (δηλ., πολυμεταβλητά αντικείμενα), έτσι ώστε παρόμοια αντικείμενα να βρίσκονται κοντά μεταξύ τους και ανόμοια αντικείμενα να είναι μακρύτερα το ένα από το άλλο. Αυτές οι σχέσεις μεταξύ των αντικειμένων, σε κάθε έναν από τους πολλαπλούς άξονες (ένας άξονας για κάθε μεταβλητή), χαρακτηρίζονται στη συνέχεια αριθμητικά και/ή γραφικά. Πολλές τεχνικές συντονισμού (ordination technique) υφίστανται, συμπεριλαμβανομένων της ανάλυσης κυρίων συνιστωσών (PCA), της μημετρικής πολυδιάστατης κλιμάκωσης (NMDS), της ανάλυσης ανταπόκρισης (CA) και

των παραγώγων της (detrended CA (DCA), canonical CA (CCA)), Bray-Curtis ordination, και ανάλυσης πλεονάσματος (RDA), μεταξύ άλλων.

Συνήθως στην (MDS) τα δεδομένα προς ανάλυση είναι μια συλλογή από I αντικείμενα (χρώματα, πρόσωπα, μετοχές,...) στην προκειμένη περίπτωση τραγούδια, πάνω στα οποία ορίζεται μια *συνάρτηση αποστάσεων*, $\delta_{i,j} :=$ απόσταση ανάμεσα στο i^{th} και j^{th} αντικείμενο. Αυτές οι αποστάσεις συνήθως είναι οι εγγραφές ενός πίνακα ανομοιότητας (dissimilarity matrix) Δ.

$$\Delta := \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,I} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,I} \\ \vdots & \vdots & & \vdots \\ \delta_{I,1} & \delta_{I,2} & \cdots & \delta_{I,I} \end{pmatrix}$$

Πίνακας 5.2: πίνακας ανομοιότητας (dissimilarity matrix)

Στόχος του MDS είναι, δεδομένου ενός πίνακα Δ, να βρει I διανύσματα $x_1, \ldots, x_I \in \mathbb{R}^{\mathbb{N}}$ τέτοια ώστε $\|x_i - x_j\| \approx \delta_{i,j} \ \ \forall i, j \in I$, όπου $\|\cdot\|$ είναι διανυσματική νόρμα. Στο κλασικό MDS, η παραπάνω νόρμα δηλώνει Ευκλείδεια απόσταση, αλλά, σε μια ευρύτερη έννοια μπορεί να είναι μια μετρική ή αυθαίρετη συνάρτηση απόστασης. Με άλλα λόγια, η MDS επιχειρεί να βρει μια ενσωμάτωση από τα I αντικείμενα στον $\mathbb{R}^{\mathbb{N}}$ χώρο έτσι ώστε να διατηρούνται οι αποστάσεις. Αν η διάσταση του N επιλέγεται να είναι 2 ή 3, μπορούμε να παραστήσουμε γραφικά τα διανύσματα x_i για να ληφθεί μια απεικόνιση των ομοιοτήτων μεταξύ των I αντικειμένων. Να σημειώσουμε ότι τα διανύσματα x_i δεν είναι μοναδικά: Μέσω της Ευκλείδειας απόστασης, μπορούν αυθαίρετα να μετατοπιστούν, περιστραφούν, και αντικατοπτριστούν, δεδομένου ότι αυτοί οι μετασχηματισμοί δεν αλλάζουν τις αποστάσεις κατά ζεύγη $\|x_i - x_j\|$.

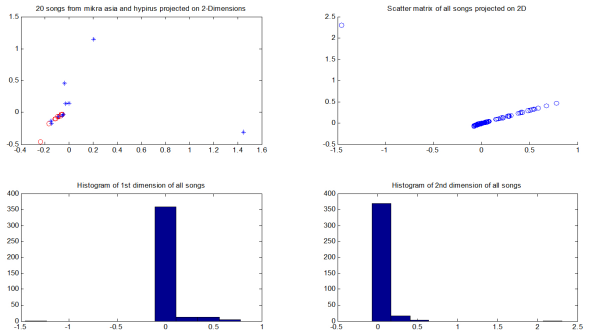
Υπάρχουν διάφορες προσεγγίσεις για τον προσδιορισμό των διανυσμάτων x_i. Συνήθως, η MDS διατυπώνεται ως ένα πρόβλημα βελτιστοποίησης, όπου $(x_1, \ldots, x_I)$ δηλώνονται ως ελαχιστοποιητής κάποιας συνάρτησης κόστους, για παράδειγμα:

$$\min_{x_1,\ldots,x_I} \sum_{i<j} (\|x_i - x_j\| - \delta_{i,j})^2 \tag{5.1}$$

Μια λύση θα μπορούσε να βρεθεί μέσω τεχνικών αριθμητικής βελτιστοποίησης. Για ορισμένες ειδικά επιλεγμένες συναρτήσεις κόστους, οι ελαχιστοποιητές μπορούν να διατυπωθούν αναλυτικά μέσω αποσύνθεσης ίδιο ανάλυσης πινάκων. Χρησιμοποιώντας το λογισμικό πακέτο [δρτ] σε περιβάλλον MATLAB προσπαθήσαμε μέσω της τεχνικής MDS να μειώσουμε τις διαστάσεις του tf-idf term document matrix $(W^{3815\times388})^T$ στις

2 με απώτερο σκοπό την σχηματική απεικόνιση των δεδομένων και την επιβεβαίωση τυχόν δομών ή ομάδων που μπορεί να προϋπήρχαν.

Για το σκοπό αυτό στην αρχή επιλέξαμε ένα πολύ μικρό σύνολο τραγουδιών από την συλλογή, 20 στον αριθμό, όπου τα 10 από αυτά προέρχονταν από την γεωγραφική περιοχή της Μικράς Ασίας και τα υπόλοιπα 10 από την Ήπειρο. Μειώνοντας τις διαστάσεις και απεικονίζοντας τα δεδομένα λαμβάνουμε μια σφαιρική άποψη για το εάν τα δεδομένα μας τελικά είναι διαχωρίσιμα ή όχι στον μειωμένο αυτό χώρο. Παρακάτω παραθέτουμε την σχηματική απεικόνιση των 20 τραγουδιών στον δισδιάστατο χώρο.



Σχήμα 5.1: Προβολή των 20 τραγουδιών στον δισδιάστατο χώρο μέσω της τεχνικής MDS

Όπως παρατηρούμε στο παραπάνω σχήμα στην αριστερή επάνω γωνία εμφανίζονται τα 20 τραγούδια προβαλλόμενα στον δισδιάστατο χώρο, όπου κάποια από αυτά είναι διαχωρίσιμα. Στην δεξιά επάνω γωνία εμφανίζεται ο term document matrix $(W^{3815\times388})^T$ προβαλλόμενος στις 2 διαστάσεις που περιλαμβάνει όλα τα τραγούδια της συλλογής. Κάτω αριστερά εμφανίζεται το ιστόγραμμα της πρώτης διάστασης όλων των τραγουδιών ενώ κάτω δεξιά εμφανίζεται το ιστόγραμμα της δεύτερης διάστασης όλων των τραγουδιών. Εξοπλισμένοι πλέον με την σχηματική αναπαράσταση ότι τα δεδομένα μας είναι διαχωρίσιμα προχωρούμε πλέον στο επόμενο μας πείραμα που αφορά την ομαδοποίηση και προβολή των δεδομένων στις 3 διαστάσεις. Για την επίτευξη του στόχου μας χρησιμοποιούμε έναν πολύ απλό και πλέον διαδεδομένο αλγόριθμο τον k-means.

5.3 Ομαδοποίηση K-μέσων

Η ομαδοποίηση κατά k-means είναι μια μέθοδος διανυσματική κβάντωσης που προήλθε αρχικά από την επεξεργασία σήματος, που είναι αρκετά δημοφιλής για την ανάλυση

ομαδοποίησης (cluster analysis) στον τομέα της εξόρυξης δεδομένων. Η ομαδοποίηση κατά k-means στοχεύει στην κατάτμηση n παρατηρήσεων σε k συστάδες όπου κάθε παρατήρηση ανήκει στην ομάδα με την πλησιέστερη μέση τιμή, χρησιμεύοντας ως πρωτότυπο της ομάδας/(cluster). Αυτό οδηγεί σε κατάτμηση του χώρου δεδομένων σε κελιά Voronoi. Το πρόβλημα είναι υπολογιστικά δύσκολο (NP-hard). Ωστόσο, υπάρχουν αποδοτικοί ευρετικοί (heuristic) αλγόριθμοι που χρησιμοποιούνται συνήθως και συγκλίνουν γρήγορα σε ένα τοπικό βέλτιστο. Αυτοί οι αλγόριθμοι είναι συνήθως παρόμοιοι με τον αλγόριθμο (EM) για μείγματα γκαουσιανών κατανομών μέσω του καθορισμού μιας επαναληπτικής προσέγγισης χρησιμοποιούμενη και από τους δύο αλγορίθμους. Επιπλέον, και οι δύο αλγόριθμοι χρησιμοποιούν κέντρα διασποράς cluster centers για να μοντελοποιήσουν τα δεδομένα. Ωστόσο, Η ομαδοποίηση κατά k-means τείνει να βρίσκει ομάδες συγκρίσιμων χωρικών εκτάσεων, ενώ ο μηχανισμός expectation-maximization επιτρέπει στις ομάδες να έχουν διαφορετικά σχήματα.

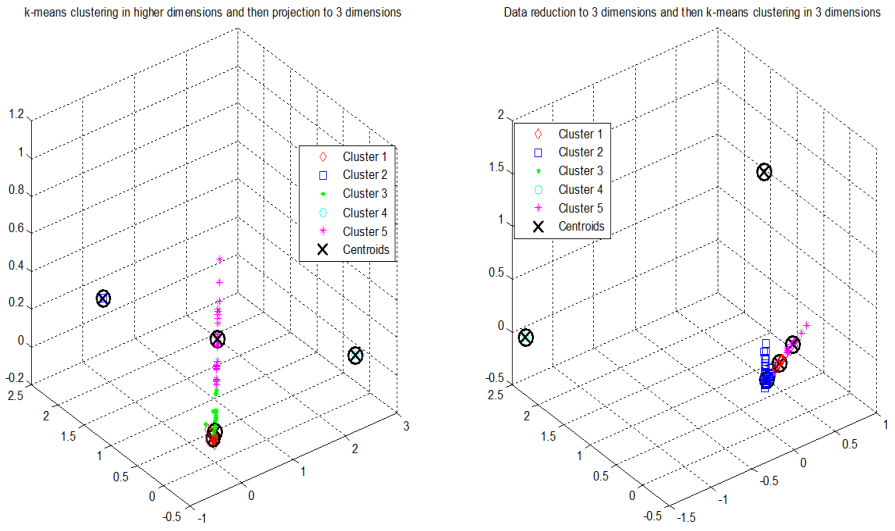
Περιγραφή:

Λαμβάνοντας υπόψη ένα σύνολο παρατηρήσεων $(x_1, x_2, \ldots, xn)$, όπου κάθε παρατήρηση είναι ένα d-διάστατο πραγματικό διάνυσμα, η ομαδοποίηση κατά k-means στοχεύει στην κατάτμηση των n παρατηρήσεων σε k σύνολα $(k \leq n)$ $\mathbf{S} = \{S1, S2, ..., Sk\}$, έτσι ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων κατά-ομάδα "within-cluster sum of squares"(WCSS):

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \tag{5.2}$$

όπου μ_i είναι ο μέσος των σημείων στο σύνολο S_i. Παρακάτω απεικονίζεται σχηματικά η διαδικασία μείωσης διαστάσεων των δεδομένων στις 3 διαστάσεις και η ομαδοποίηση τους μέσω του k-means.

Να σημειώσουμε πως η ομαδοποίηση έγινε για $k = 5$, και προσπαθήσαμε να δείξουμε τυχόν διαφορές ανάμεσα στην ομαδοποίηση στον χώρο υψηλών διαστάσεων και έπειτα προβολή των δεδομένων στις 3 διαστάσεις, έναντι στην προβολή πρώτα των δεδομένων στον τρισδιάστατο χώρο και έπειτα ομαδοποίησης τους. Όπως παρατηρείτε και στις δύο περιπτώσεις υπάρχουν clusters τα οποία είναι αρκετά διάσπαρτα μεταξύ τους καθώς και άλλα που βρίσκονται πολύ κοντά μεταξύ τους. Εξετάζοντας τα γραφήματα καταλήγουμε να διερωτηθώμεθα για το εάν τα clustres που δημιουργούνται έχουν κάποια εννοιολογική σημασία ή γεωγραφική σημασία όσον αφορά τα τραγούδια της συλλογής. Για παράδειγμα κάποια από τα ερωτήματα που τίθενται είναι:



Σχήμα 5.2: Αριστερό διάγραμμα απεικονίζει την ομαδοποίηση στον χώρο υψηλών διαστάσεων και έπειτα προβολή στις 3 διαστάσεις, ενώ το δεξί διάγραμμα απεικονίζει την προβολή των δεδομένων στις 3 διαστάσεις και έπειτα την ομαδοποίηση

i. **Q1**: Άραγε μήπως τα clusters που δημιουργήθηκαν περιέχουν τραγούδια από την ίδια γεωγραφική περιοχή;

ii. **Q2**: Ποιοι είναι οι πιο πολυσύχναστοι όροι σε κάθε cluster που θα μπορούσαν κάλλιστα να περιγράφουν τα τραγούδια που βρίσκονται στο cluster αυτό καθαυτό;

Προτού όμως δούμε την γεωγραφική περιοχή των τραγουδιών αναφορικά με το πρώτο ερώτημα καλό θα ήταν να κοιτάξουμε πόσα τραγούδια περιέχονται σε κάθε cluster και ποια είναι αυτά.

Γνωρίζοντας πλέον πόσα στον αριθμό τραγούδια υπάρχουν σε κάθε cluster, ίσως έτσι να μπορούμε να προσδιορίσουμε και την γεωγραφική περιοχή που το καθένα από αυτά αντιστοιχεί.

$Cluster_1$	$Cluster_2$	$Cluster_3$
Έλα έλα λέγωσε	Ινάντια στην τύχη μου	Αποφάσισα να γίνω
		Άτταρης
		Εγώ είμαι προσφυγάκι
		Θα σπάσω κούπες
		Μεμέτης
		Ο μεμέτης
		Σωκιανή
		Τζιβαέρ
		Της τριανταφυλλιάς τα φύλλα
		Αητέντς επαραπέτανεν
		Απ αδά και πέρα θέλω
		Αφήνω Γειά
		Καραβίτικος
		Ποταμίτικος
		Εξορία - (Αρκαδία)
		Περβόλια μου με τάνθη σου

Πίνακας 5.3: Παρουσίαση τραγουδιών που ομαδοποιήθηκαν στα clusters 1-3

$Cluster_4$	$Cluster_5$	**Συνολ. Τραγ**
Άσπρο μου τριανταφυλλάκι	Αλατσατιανή	$cluster_1 : 1$
Έσπασες τα πιάτα	Ανάθεμά σε Σμύρνη	$cluster_2 : 1$
Η Έλλη	Από ξένο τόπο	$cluster_3 : 16$
Κόνιαλης	Από τα γλυκά σου μάτια	$cluster_4 : 13$
Μανές της καληνυχτιάς	Άστραψεν η ανατολή	$cluster_5 : 357$
Όσο βαρούν τα σίδερα	Αυτά τα μάτια	
Σαν πεθάνω στο καράβι	Αχ μελαχρινό μου	
Χριστόδουλος αγάπησε	Γιαρούμπι	
Ο Γιωργαλάκης	Δε σε θέλω πια	
Διαμάντι δαχτυλίδι	Δεν είναι αυγή να σηκωθώ	
Άσπρο τριαντάφυλλο	Εγγλεζίτσα	
Το Τσάμπασιν	Είναι καρδιές όπου γελούν	
Καραλάζο	Ερηνιώ Κατερινιώ	
	Έρχομαι κι εσύ κοιμάσαι	
	Εφτά βδομάδες έκανα	
	⋮	

Πίνακας 5.4: Παρουσίαση τραγουδιών που ομοαδοποίηθηκαν στα clusters 4-5

	$cluster_1$	$cluster_2$	$cluster_3$	$cluster_4$	$cluster_5$
Μικρασιάτικα			56.25%(9/16)	53.85%(7/13)	17.53%(64/357)
Θρακιώτικα					4.48%(16/357)
Ηπειρώτικα					8.12%(29/357)
Μακεδονίτικα				7.69%(1/13)	5.6%(20/357)
Ρουμελιώτικα				7.69%(1/13)	8.4%(30/357)
Θεσσαλιώτικα				7.69(1/13)	10.07%(36/357)
Μωραΐτικα				7.69%(1/13)	13.73%(49/357)
Ποντιακά	100%		18.75%(3/16)	15.38%(2/13)	17.09%(61/357)
Νησιώτικα			18.75%(3/16)		11.2%(40/357)
Κρητικά		100%	6.25%(1/16)		3.36%(12/357)
Συνολ. Τραγ.	1	1	16	13	357

Πίνακας 5.5: Κατανομή τραγουδιών ανά γεωγραφική περιοχή μέσα σε κάθε cluster

Εφόσον γνωρίζουμε πλέον την γεωγραφική κατανομή των τραγουδιών σε κάθε cluster, επόμενος μας στόχος είναι η εύρεση των πιο πολυσύχναστων όρων σε κάθε cluster ώστε να δώσουμε απάντηση στο ερώτημα **Q3** που θέσαμε νωρίτερα. Η διαδικασία περιλαμβάνει **δύο στάδια** όπου σε **πρώτο στάδιο** δημιουργήθηκε μια μέθοδος σε MATLAB για την εύρεση των αντίστοιχων όρων σε κάθε cluster καθώς και την κατάταξη τους σε φθίνουσα σειρά ανάλογα με την συχνότητα εμφάνισης. Σε **δεύτερο στάδιο** υλοποιήθηκε μια διαδικασία αντιστοίχισης κάθε όρου στις πραγματικές λέξεις προέλευσης που ονομάστηκε *inverse stemming process*. Τα αποτελέσματα που λήφθηκαν από τις παραπάνω διεργασίες απεικονίζονται σχηματικά στους παρακάτω πίνακες όπου επιλεκτικά παρουσιάζουμε τους *top* $k = 5$ όρους και το αντίστοιχο λεξιλόγιο προέλευσης σε κάθε cluster.

$Cluster_1$	ΚΑΡΔ(7)	ΕΜΟΡΦ(6)	ΚΟΡΤΣ(6)	ΑΓΑΠ(6)	ΦΑΡΜΑΚ(3)
$Cluster_2$	ΙΝΑΝΤ(2)	ΑΓΙ(1)	ΑΝΘΡΩΠ(1)	ΕΝΑΝΤΙ(1)	ΠΙΕΝ(1)
$Cluster_3$	ΑΜ(191)	ΠΑΡ(10)	ΚΑΡΔ(9)	ΜΑΥΡ(8)	ΜΑΤ(6)
$Cluster_4$	ΑΜ(96)	ΘΕΛ(15)	ΚΑΝ(11)	ΓΙΟΥΡΓΑΛΑΚ(11)	ΚΑΨ(10)
$Cluster_5$	ΑΓΑΠ(170)	ΜΩΡ(154)	ΜΑΝ(153)	ΑΙΝΤ(119)	ΚΑΛ(90)

Πίνακας 5.6: Εμφάνιση πιο πολυσύχναστων όρων σε κάθε cluster καθώς και τον αντίστοιχο αριθμό εμφάνισης του κάθε όρου που δηλώνεται μέσα σε παρένθεση μετά από κάθε όρο

Stem Terms				
ΚΑΡΔ	ΚΑΡΔΙΑ	ΚΑΡΔΙΑΝ	ΚΑΡΔΙΑΜ	
ΕΜΟΡΦ	ΕΜΟΡΦΟ	ΕΜΟΡΦΟΝ	ΕΜΟΡΦΗ	ΕΜΟΡΦΙΝΕΝ
ΚΟΡΤΣ	ΚΟΡΤΣΟΠΟΝ	ΚΟΡΤΣΟΥΔΙ	ΚΟΡΤΣΑΡΙ	
ΑΓΑΠ	ΑΓΑΠΑ	ΑΓΑΠΑΓΕ	ΑΓΑΠΑΕΙ	ΑΓΑΠΑΣ
ΦΑΡΜΑΚ	ΦΑΡΜΑΚΙ	ΦΑΜΑΚΟ	ΦΑΡΜΑΚΟΥ	ΦΑΡΜΑΚΟΥΤΕ
ΙΝΑΝΤ	ΙΝΑΝΤΙΑ			
ΑΓΙ	ΑΓΙΟ	ΑΓΙΟΙ	ΑΓΙΟΥΣ	
ΑΝΘΡΩΠ	ΑΝΘΡΩΠ	ΑΝΘΡΩΠΙ	ΑΝΘΡΩΠΟ	ΑΝΘΡΩΠΟΙ
ΕΝΑΝΤΙ	ΕΝΑΝΤΙΟΥΣ			
ΠΙΕΝ	ΠΙΕΝΩ			
ΑΜ	ΑΜΑΝ	ΑΜΟΝ		
ΠΑΡ	ΠΑΡΑΔΟΣΗ	ΠΑΡΑΔΙΔΩ		
ΜΑΥΡ	ΜΑΥΡΑ	ΜΑΥΡΗ	ΜΑΥΡΟ	ΜΑΥΡΟΝ
ΜΑΤ	ΜΑΤΙ	ΜΑΤΙΑ	ΜΑΤΟΥΜΑΙ	
ΘΕΛ	ΘΕΛΑ	ΘΕΛΕ	ΘΕΛΕΙ	ΘΕΛΕΙΣ
ΚΑΝ	ΚΑΝΕ	ΚΑΝΑ	ΚΑΝΑΜΕ	ΚΑΝΑΝ
ΚΑΨ	ΚΑΨΕΙ	ΚΑΨΟΝ	ΚΑΨΩ	
ΜΩΡ	ΜΩΡΑ	ΜΩΡΕ	ΜΩΡΗ	ΜΩΡΙΑ
ΜΑΝ	ΜΑΝΑ	ΜΑΝΑΔΕΣ	ΜΑΝΑΝ	ΜΑΝΑΣ
ΑΙΝΤ	ΑΙΝΤΕ			
ΚΑΛ	ΚΑΛΕ	ΚΑΛΕΙ	ΚΑΛΗ	ΚΑΛΗΣ

Πίνακας 5.7: Αντιστοίχιση των όρων/stem terms που βρέθηκαν στα clusters 1-5 στο λεξιλόγιο του stem vocabulary της συλλογής των 388 τραγουδιών μέσω της διαδικασίας inverse stemming process

Παρατηρώντας τα δεδομένα μας συμπεραίνουμε πως θα ήταν αρκετά ενδιαφέρον εάν είχαμε την δυνατότητα να θέτουμε ερωτήματα του τύπου όπως π.χ. *"Ποια είναι τα πιο αντιπροσωπευτικά τραγούδια για μια θεματική περιοχή"*; Για να είμαστε σε θέση να απαντούμε σε τέτοιου είδους ερωτήματα θα πρέπει να γνωρίζουμε εκ των προτέρων ποιες θεματικές ενότητες υπάρχουν στην συλλογή και πως θα ήταν εφικτή η ανάκτηση τραγουδιών με παρόμοιο εννοιολογικό περιεχόμενο.

Για την επίτευξη αυτών των στόχων θα χρησιμοποιήσουμε μια ευρέως γνωστή και αποδεδειγμένη τεχνική που προέρχεται από το πεδίο της ανάκτησης πληροφοριών. Η διεξαγωγή των πειραμάτων, τα αποτελέσματα τους καθώς και η επιβεβαίωση αυτών όπως και η περιγραφή της τεχνικής αναλύονται στο κεφάλαιο που ακολουθεί.

5.4 Ανάκτηση Τραγουδιών Μέσω Σημασιολογικού Περιεχομένου (LSA)

5.4.1 Εισαγωγή

Η λανθάνουσα σημασιολογική ευρετηρίαση (LSI/LSA) είναι μια μέθοδος ανάκτησης και ευρετηρίασης που χρησιμοποιεί μια μαθηματική τεχνική που ονομάζεται αποσύνθεση ιδιαζουσών τιμών (SVD) για την αναγνώριση προτύπων στις σχέσεις μεταξύ των όρων και των εννοιών που περιέχονται σε ένα μια δομημένο συλλογή κειμένου.

Η LSI/LSA βασίζεται στην αρχή ότι οι λέξεις που χρησιμοποιούνται στο ίδιο γενικό πλαίσιο τείνουν να έχουν παρόμοιες σημασίες. Ένα βασικό χαρακτηριστικό της LSI/LSA είναι η ικανότητά να εξάγει το εννοιολογικό περιεχόμενο του σώματος ενός κειμένου με τη δημιουργία συσχετίσεων μεταξύ των όρων αυτών που εμφανίζονται σε παρόμοια πλαίσια.

Η LSI/LSA είναι επίσης μια εφαρμογή ανάλυσης αναλογίας, μια πολυμεταβλητή στατιστική τεχνική που αναπτύχθηκε από τον Jean - Paul Benzecri στις αρχές του 1970, σε έναν πίνακα συνάφειας δημιουργούμενο από καταγραφή όρων/λέξεων σε έγγραφα.

Ονομάζεται Λανθάνουσα Σημασιολογική Ευρετηρίαση LSI, λόγω της ικανότητάς της να συσχετίζει σημασιολογικά σχετικούς όρους που είναι λανθάνοντες/latent σε μια συλλογή κειμένου, εφαρμόστηκε για πρώτη φορά σε κείμενο στα Bell Laboratories τέλη του 1980. Η μέθοδος που ονομάζεται επίσης λανθάνουσα σημασιολογική ανάλυση (LSA), αποκαλύπτει την υποκείμενη λανθάνουσα σημασιολογική δομή στη χρήση των λέξεων σε ένα σώμα κειμένου και πώς μπορεί να χρησιμοποιηθεί για να εξάγουμε το νόημα του κειμένου ως απάντηση στα ερωτήματα των χρηστών, που συνήθως αναφέρονται ως εννοιολογικές αναζητήσεις.

Ερωτήματα, ή εννοιολογικές αναζητήσεις, που έχουν τεθεί σε μια σειρά από έγγραφα μέσω της LSI θα επιστρέψουν αποτελέσματα που είναι εννοιολογικά παρόμοια με την έννοια των κριτηρίων αναζήτησης, ακόμη και αν τα αποτελέσματα δεν μοιράζονται μια συγκεκριμένη λέξη ή λέξεις με τα κριτήρια αναζήτησης.

5.4.2 Επισκόπηση Πειράματος

Η υλοποίηση της τεχνικής LSA για ανάκτηση παρόμοιων τραγουδιών ως προς το εννοιολογικό τους περιεχόμενο στα ερωτήματα των χρηστών εφαρμόστηκε σε περιβάλλον MATLAB. Δεδομένου ενός ερωτήματος $\vec{q}$ ο χρήστης αναμένει ως απάντηση ένα διάνυσμα $\vec{a}$ όπου κάθε εγγραφή δηλώνει την ομοιότητα ανάμεσα στους όρους του ερωτήματος και τα τραγούδια της συλλογής. Ένα απτό παράδειγμα ενός ερωτήματος και της αντίστοιχης απάντησης δίνονται παρακάτω.

$$\vec{q} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}^T \tag{5.3}$$

$$\vec{a} = \begin{bmatrix} 0.1254 & 0.2548 & 0.3567 & 0.7516 & 0.6247 & 0.8320 & 0.7915 \end{bmatrix} \tag{5.4}$$

Όπου κάθε στήλη του διανύσματος $\vec{q}$ εκφράζει και έναν όρο του ερωτήματος που θέτει ο χρήστης. Για παράδειγμα για να απαντήσουμε στο ερώτημα: ***"Ποια τραγούδια μιλάνε για αγάπη"***; Τότε θα έπρεπε να εντοπίσουμε ένα τραγούδι από την συλλογή που να αναφέρεται στην αγάπη και να χρησιμοποιήσουμε αυτό ως ερώτημα/είσοδο στην LSI. Όπου υπάρχει ο αριθμός 1 στο διάνυσμα $\vec{q}$ δηλώνει πως λαμβάνονται υπόψη μόνο οι αντίστοιχοι όροι του τραγουδιού κατά την διαδικασία του ερωτήματος στην LSI. Το διάνυσμα $\vec{a}$ είναι μήκους **388** όσο και το πλήθος των τραγουδιών στη συλλογή και κάθε στήλη δηλώνει την ομοιότητα ανάμεσα στο ερώτημα και το αντίστοιχο τραγούδι της συλλογής.

5.4.3 Δεδομένα

Τα δεδομένα αφορούν τη συλλογή των 388 τραγουδιών από τις 10 διαφορετικές γεωγραφικές περιοχές καθώς και του stem vocabulary που περιλαμβάνει κάθε πιθανό όρο που παρουσιάζεται σε αυτά τα 388 τραγούδια και είναι διαστάσεων 3815×1. Οπότε ουσιαστικά αναφερόμαστε σε έναν term document matrix διαστάσεων $W^{3815 \times 388}$.

5.4.4 Ερωτήματα προς Διερεύνηση

Τα ερωτήματα που τέθηκαν κατά την διάρκεια αυτού του πειράματος ήταν τα ακόλουθα:

i. Ποια είναι τα πιο αντιπροσωπευτικά τραγούδια για μια θεματική; π.χ. Ποια τραγούδια μιλούν καλύτερα για την ξενιτιά;

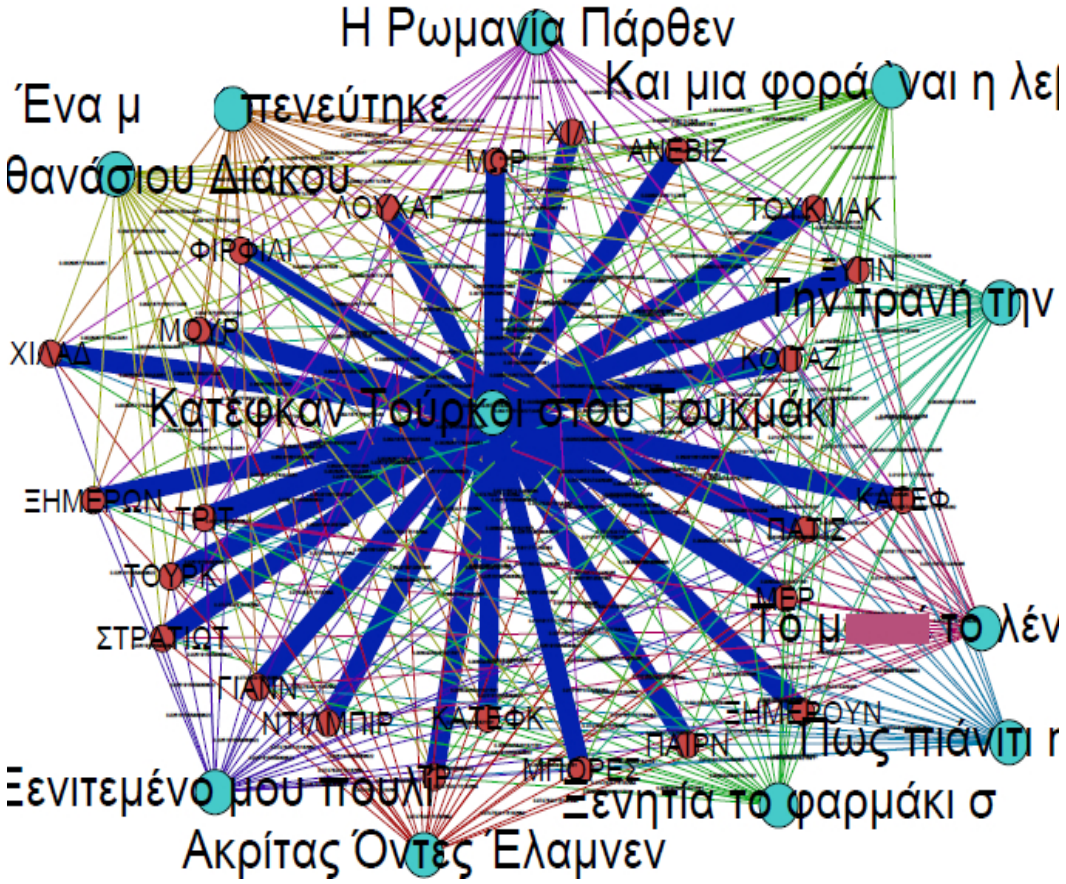
ii. Ποιες είναι οι πιο σημαντικές θεματικές των ποντιακών τραγουδιών; Για ποιες θεματικές μιλούν τα ποντιακά τραγούδια; (π.χ. ανδριωσύνη, αγάπη, ξενιτιά...)

5.4.5 Αποτελέσματα

5.4.5.1 Επιβεβαίωση/Groundtruth

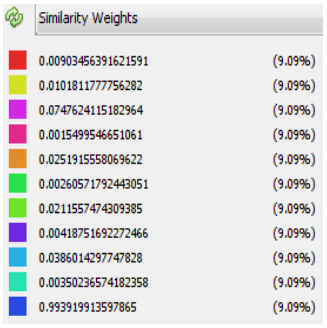
Προτού ξεκινήσουμε να δούμε τα αποτελέσματα των ερωτήσεων που τέθηκαν παραπάνω θα προσπαθήσουμε να αφιερώσουμε λίγο χρόνο στην διερεύνηση της ορθής λειτουργίας της τεχνικής LSI. Για τον σκοπό αυτό έχουμε χωρίσει ένα μικρό σύνολο δεδομένων (26 τραγούδια) σε 4 θεματικές ενότητες (*ανδριωσύνης, αγάπης, ξενιτιάς, αποκριάς*) από 6 τραγούδια αντίστοιχα και ουσιαστικά θέτουμε κάθε φορά ένα ερώτημα με ένα από τα τραγούδια της κάθε θεματικής περιοχής και αναμένουμε τα αποτελέσματα να επιστρέψουν από την συλλογή των 24 τα τραγούδια εκείνα που είναι παρόμοια ως προ το εννοιολογικό τους περιεχόμενο.

Να σημειώσουμε πως τα αποτελέσματα της τεχνικής LSI από το MATLAB εξάγονται σε μορφή .csv που περιέχει την λίστα ακμών ανάμεσα στις συσχετίσεις των όρων του ερωτήματος και των υπόλοιπων 24 τραγουδιών της τεχνητής συλλογής που κατασκευάσαμε. Έπειτα το .csv αρχείο με την λίστα ακμών εισάγετε στο δωρεάν διαθέσιμο λογισμικό για ανάλυση δικτύων NodeXL [νοδ] από το οποίο και εξάγουμε ένα τελικό αρχείο με την κατάληξη .graphml, το οποίο είναι αυτό που ουσιαστικά εισάγουμε στο επόμενο δωρεάν διαθέσιμο λογισμικό ανάλυσης γράφων Gephi [γεπ] προς περαιτέρω επεξεργασία και απεικόνιση.



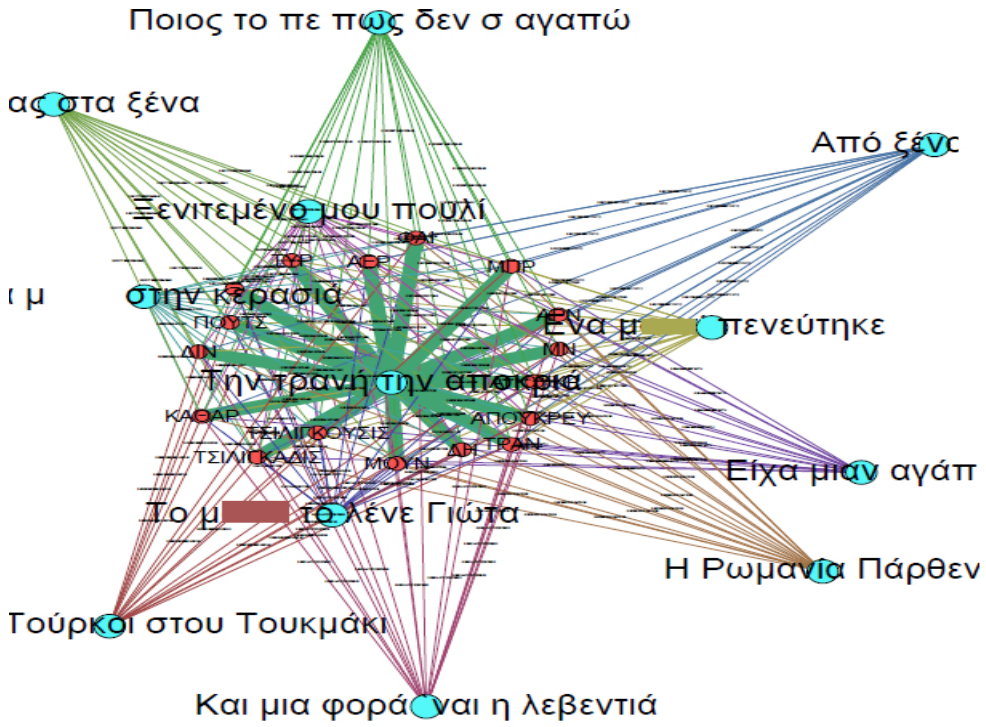
Σχήμα 5.3: Απεικόνιση ερωτήματος ανδριωσύνης και των τραγουδιών που βρίσκονται εννοιολογικά κοντά σε αυτό

Στο διάγραμμα που απεικονίζεται παραπάνω αλλά και στα υπόλοιπα 2 που ακολουθούν οι κόμβοι με κόκκινο χρώμα εκφράζουν τους όρους του ερωτήματος, που στην προκειμένη περίπτωση ως ερώτημα έχει χρησιμοποιηθεί το τραγούδι με τίτλο ***"Κατέφκαν Τούρκοι στου Τουκμάκι"*** ενώ οι κόμβοι με γαλάζιο χρώμα εκφράζουν τα τραγούδια της τεχνητής συλλογής που κατασκευάσαμε. Το πάχος των ακμών εκφράζει το ποσοστό ομοιότητας της "cosine similarity" ανάμεσα στους όρους του ερωτήματος και τα 24 τραγούδια της τεχνητής συλλογής, όσο μεγαλύτερο το πάχος τόσο μεγαλύτερο και το ποσοστό ομοιότητας μεταξύ τους. Επίσης το χρώμα των ακμών συνδέεται με την τιμή που αναθέτει η "cosine similarity" σε κάθε ζεύγος κόμβων. Προς μεγαλύτερη διευκόλυνση παραθέτουμε παρακάτω τον πίνακα κωδικοποίησης χρωμάτων των ακμών και τις αντίστοιχες τιμές της "cosine similarity" του γραφήματος στο ΣΧΗΜΑ 5.3 στην σελίδα 85.

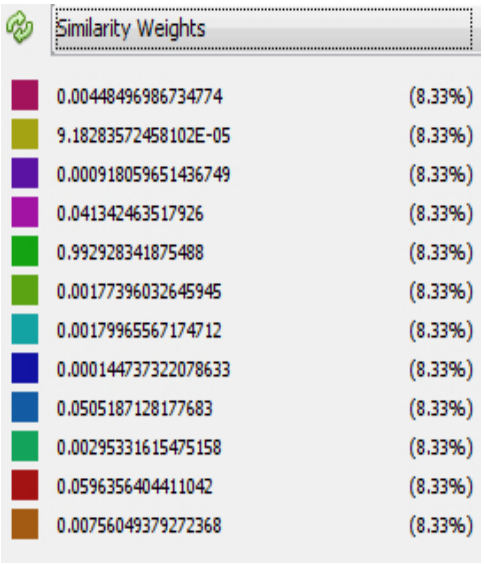


Σχήμα 5.4: Αντιστοίχιση χρωματικής περιγραφής των ακμών του γράφου στις αναδεικνυόμενες τιμές της cosine siilarity

Όπως παρατηρούμε από τα παραπάνω σχήματα η LSI και στις δυο περιπτώσεις κατάφερε να ανακτήσει και να αναθέσει μεγαλύτερο βάρος ακμών σε τραγούδια με παρόμοιο εννοιολογικό περιεχόμενο. Στο ΣΧΗΜΑ 5.3 θέσαμε ως ερώτημα ***"ποια είναι τα τραγούδια που περιγράφουν καλύτερα την θεματική περιοχή της ανδριωσύνης"***. Για το σκοπό αυτό χρησιμοποιήσαμε ως ερώτημα το τραγούδι με τίτλο ***"Κατέφκαν οι Τούρκοι στου Τουκμάκι"*** σε μια βάση αποτελούμενη από 24 τραγούδια από 4 διαφορετικές θεματικές περιοχές. Ως απάντηση λάβαμε τους εξής τίτλους τραγουδιών που εμφανίζονται κατά φθίνουσα σειρά ως προς την μετρική ομοιότητας συνημιτόνου στον παρακάτω πίνακα και δηλώνουν τα τραγούδια που χαρακτηρίζουν καλύτερα την θεματική περιοχή της ανδριωσύνης.



Σχήμα 5.5: Απεικόνιση ερωτήματος θεματικής περιοχής αποκριάς, χρησιμοποιώντας ως ερώτημα τους όρους του τραγουδιού με τίτλο "Την τρανή την αποκριά"



Σχήμα 5.6: Αντιστοίχιση χρωματικής περιγραφής των ακμών του γράφου που παρουσιάστηκε στο ΣΧΗΜΑ 5.5 στις αναδεικνυόμενες τιμές της cosine siilarity

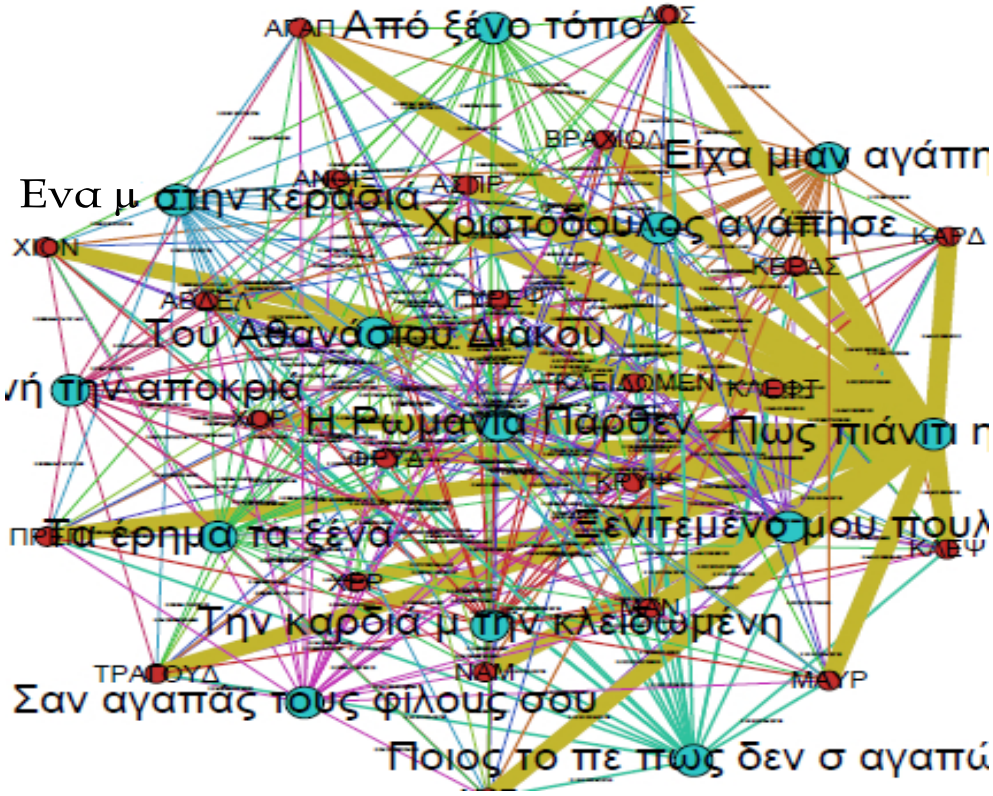
Τραγ. Ανδριωσύνης ΣΧΗΜΑ 5.3	cosine similarity weights
Κατέφκαν Τούρκοι στου Τουκμάκι	0.993919914
Ακρίτας Όντες Έλαμνεν	0.074762412
Η Ρωμανία Πάρθεν	0.03860143
Ξενιτεμένο μου πουλί	0.025191556
Το μ... το λένε Γιώτα	0.021155747
Πως πιάνιτι η αγάπη	0.010181178
Ξενητία το φαρμάκι σ	0.009034564
Ένα μ... πενεύτηκε	0.004187517
Την τρανή την αποκριά	0.003502366
Του Αθανάσιου Διάκου	0.002605718
Και μια φορά ᾽ναι η λεβεντιά	0.001549955

Πίνακας 5.8: Παρουσίαση τραγουδιών που εκφράζουν καλύτερα την θεματική περιοχή της ανδριωσύνης ταξινομημένα κατά φθίνουσα σειρά εμφάνισης ως προς τη μετρική ομοιότητας συνημιτόνου

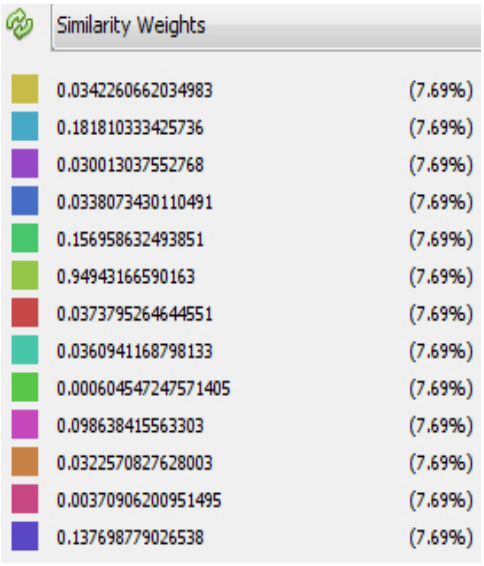
Κατά τον ίδιο τρόπο παρουσιάζονται πάλι ως προς φθίνουσα σειρά με βάση την μετρική ομοιότητας συνημιτόνου τα τραγούδια που εκφράζουν καλύτερα την θεματική περιοχή της αποκριάς και αποτυπώνονται στο ΣΧΗΜΑ 5.5. Ως ερώτημα χρησιμοποιήθηκε το τραγούδι με τίτλο ***"Την τρανή την αποκριά"***.

Τραγ. Αποκριάς ΣΧΗΜΑ 5.5	cosine similarity weights
Την τρανή την αποκριά	0.992928342
Ένα μ... στην κερασιά	0.05963564
Ένα μ... πενεύτηκε	0.050518713
Το μ... το λένε Γιώτα	0.041342464
Ξενιτεμένο μου πουλί	0.007560494
Είχα μιαν αγάπη	0.00448497
Η Ρωμανία Πάρθεν	0.002953316
Από ξένο τόπο	0.001799656
Σαν πας στα ξένα	0.00177396
Κατέφκαν Τούρκοι στου Τουκμάκι	0.00091806
Και μια φορά 'ναι η λεβεντιά	0.000144737
Ποιος το πε πως δεν σ αγαπώ	9.18284E-05

Πίνακας 5.9: Παρουσίαση τραγουδιών που εκφράζουν καλύτερα την θεματική περιοχή της αποκριάς ταξινομημένα κατά φθίνουσα σειρά εμφάνισης ως προς τη μετρική ομοιότητας συνημιτόνου



Σχήμα 5.7: Απεικόνιση ερωτήματος θεματικής περιοχής της αγάπης και των τραγουδιών που βρίσκονται εννοιολογικά κοντά σε αυτό

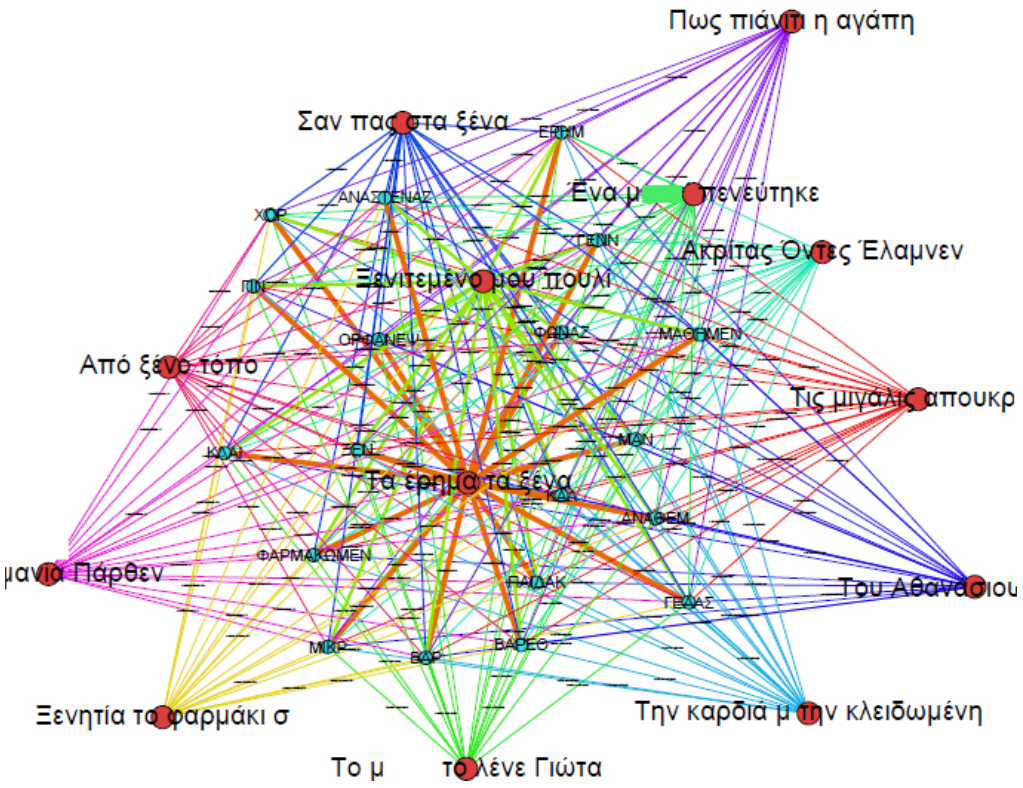


Σχήμα 5.8: Αντιστοίχιση χρωματικής περιγραφής των ακμών του γράφου που παρουσιάστηκε στο ΣΧΗΜΑ 5.7 στις αναδεικνυόμενες τιμές της μετρικής cosine siilarity

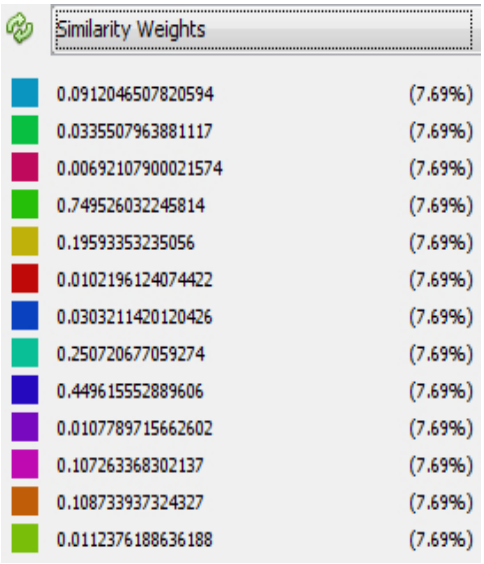
Τραγ. Αγάπης ΣΧΗΜΑ 5.7	cosine similarity weights
Πως πιάνιτι η αγάπη	0.949431666
Ποιος το πε πως δεν σ αγαπώ	0.181810333
Την καρδιά μ την κλειδωμένη	0.156958632
Είχα μιαν αγάπη	0.137698779
Από ξένο τόπο	0.098638416
Του Αθανάσιου Διάκου	0.037379526
Τα έρημα τα ξένα	0.036094117
Σαν αγαπάς τους φίλους σου	0.034226066
Χριστόδουλος αγάπησε	0.033807343
Ξενιτεμένο μου πουλί	0.032257083
Ένα μουνί στην κερασιά	0.030013038
Η Ρωμανία Πάρθεν	0.003709062
Την τρανή την αποκριά	0.000604547

Πίνακας 5.10: Παρουσίαση τραγουδιών που εκφράζουν καλύτερα την θεματική περιοχή της αγάπης ταξινομημένα κατά φθίνουσα σειρά εμφάνισης ως προς τη μετρική ομοιότητας συνημιτόνου

Στο ΣΧΗΜΑ 5.7 αλλά και στα αποτελέσματα του ΠΙΝΑΚΑ που παρουσιάζονται στο 5.10 που επέστρεψε η LSI, χρησιμοποιήθηκε ως ερώτημα το τραγούδι με τίτλο ***"Πως πιάνιτι η αγάπη"***. Στο τελευταίο γράφημα του κεφαλαίου επιβεβαίωσης της τεχνικής LSI που παρουσιάζεται στην επόμενη σελίδα οι κόμβοι που εκφράζουν τραγούδια εμφανίζονται με κόκκινο χρώμα ενώ οι όροι του ερωτήματος εμφανίζονται με γαλάζιο χρώμα. Ως ερώτημα για την ανάκτηση τραγουδιών από την θεματική περιοχή της ξενιτιάς χρησιμοποιήθηκε το τραγούδι με τίτλο ***"Τα έρημα τα ξένα"***.



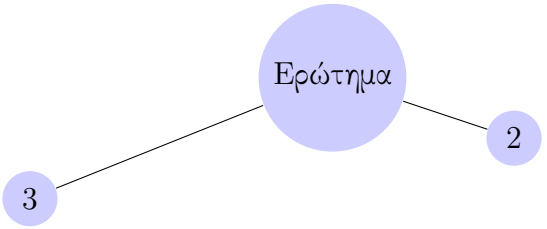
Σχήμα 5.9: Απεικόνιση ερωτήματος θεματικής περιοχής της ξενιτιάς και των τραγουδιών που βρίσκονται εννοιολογικά κοντά σε αυτό



Σχήμα 5.10: Αντιστοίχιση χρωματικής περιγραφής των ακμών του γράφου που παρουσιάστηκε στο ΣΧΗΜΑ 5.9 στις αναδεικνυόμενες τιμές της μετρικής cosine similarity

Τραγ. Ξενιτιάς ΣΧΗΜΑ 5.9	cosine similarity weights
Τα έρημα τα ξένα	0.749526032
Ξενιτεμένο μου πουλί	0.449615553
Σαν πας στα ξένα	0.250720677
Ακρίτας Όντες Έλαμνεν	0.195933532
Ένα μουνί πενεύτηκε	0.108733937
Τις μιγάλις απουκριές	0.107263368
Από ξένο τόπο	0.091204651
Η Ρωμανία Πάρθεν	0.033550796
Το μουνί το λένε Γιώτα	0.030321142
Του Αθανάσιου Διάκου	0.011237619
Πως πιάνιτι η αγάπη	0.010778972
Την καρδιά μ την κλειδωμένη	0.010219612
Ξενητία το φαρμάκι σ	0.006921079

Όπως διαπιστώνουμε από τα παραδείγματα που παρουσιάστηκαν μέχρι στιγμής σε αυτό το κεφάλαιο επιβεβαίωσης της τεχνικής LSI, τα αποτελέσματα θα μπορούσε να πει κανείς πως είναι ενθαρρυντικά, οπότε ακολουθώντας τον ίδιο συλλογισμό πλέον εκτελούμε ερωτήματα επιλέγοντας τραγούδια από τις 4 διαφορετικές θεματικές περιοχές (*αγάπης, ανδριωσύνης, αποκριάς, ξενιτιάς*) προς ολόκληρη την συλλογή των 388 τραγουδιών που διαθέτουμε. Με αυτό τον τρόπο απαντούμε στα ερωτήματα που τέθηκαν νωρίτερα στο κεφάλαιο ***"Ερωτήματα προς Διερεύνυση"***. Απαντώντας λοιπόν στο πρώτο ερώτημα του κεφαλαίου 5.4.4 λαμβάνουμε τα παρακάτω αποτελέσματα όπου πλέον οι κόμβοι με κόκκινο χρώμα δηλώνουν τα τραγούδια που χρησιμοποιήθηκαν ως ερωτήματα ενώ οι γαλάζιοι κόμβοι δηλώνουν τα τραγούδια που περιέχουν παρόμοιο σημασιολογικό περιεχόμενο με αυτά των ερωτημάτων. Και σε αυτήν την περίπτωση το πάχος της ακμής ανάμεσα σε δυο κόμβους εκφράζει το ποσοστό ομοιότητας της μετρικής cosine similarity, όπως και η απόσταση ανάμεσα σε δύο κόμβους εκφράζει πόσο όμοιοι είναι σε σημασιολογικό επίπεδο. Για παράδειγμα τρεις κόμβοι που συνδέονται μεταξύ τους με δυο ακμές όπου η μια εκ των δυο ακμών έχει μικρότερο μήκος και συνδέει δυο εκ των κόμβων δηλώνει πως το συγκεκριμένο ζεύγος κόμβων εκφράζει μεγαλύτερη ομοιότητα από το αντίστοιχο ζεύγος που συνδέεται διαμέσου μιας ακμής με μεγαλύτερο μήκος. Δείτε το παρακάτω παράδειγμα.



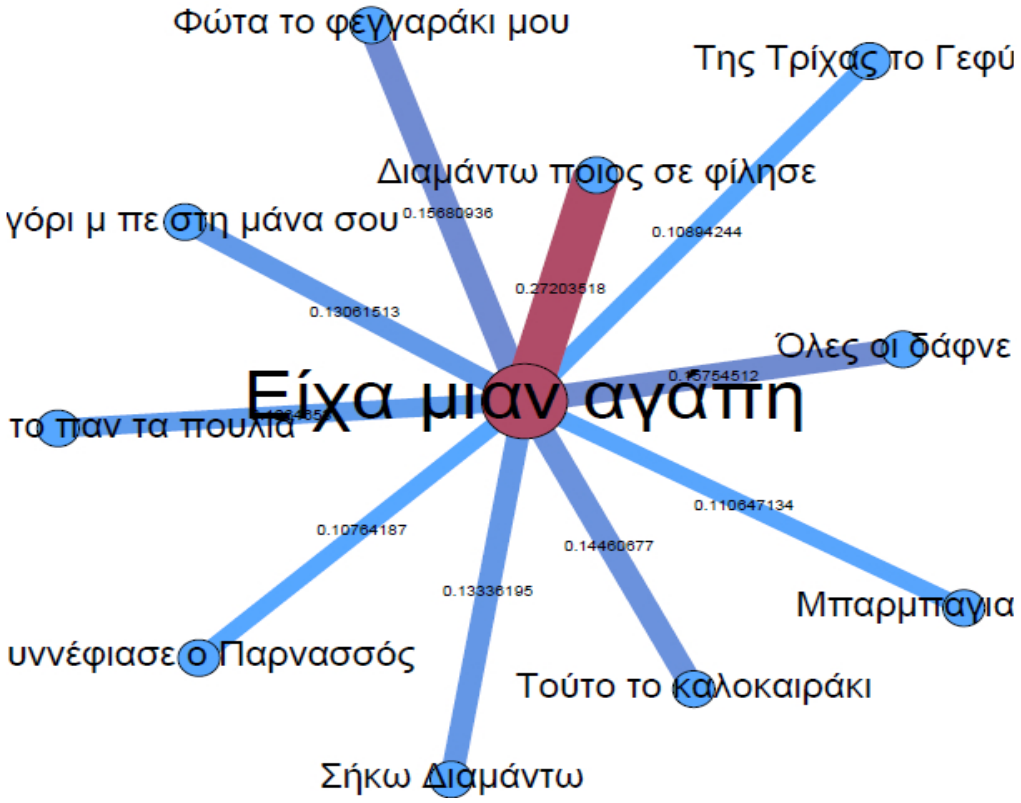
Σχήμα 5.11: Παράδειγμα επεξήγησης της σημασίας μήκους ακμής ανάμεσα σε δύο κόμβους. Ο κόμβος 2 εκφράζει τραγούδι που έχει μεγαλύτερη ομοιότητα με το ερώτημα από τον κόμβο 3

5.4.5.2 Εύρεση Αντιπροσωπευτικότερων Τραγουδιών μιας Θεματικής Περιοχής

Για την εύρεση των αντιπροσωπευτικότερων τραγουδιών μιας θεματικής περιοχής (*αγάπης, ανδριωσύνης, αποκριάς, ξενιτιάς*) θέσαμε 4 ερωτήματα τα οποία παρουσιάζονται στον παρακάτω πίνακα καθώς και η θεματική περιοχή στην οποία ανήκουν. Θα πρέπει να σημειώσουμε πως στους γράφους που ακολουθούν ακμές με ίδιο χρώμα εκφράζουν ίδιο cosine similarity. Επίσης έχει οριστεί ένα κατώφλι ώστε να εμφανίζονται τραγούδια που εκφράζουν ποσοστό ομοιότητας μεγαλύτερο του 10% με το ερώτημα.

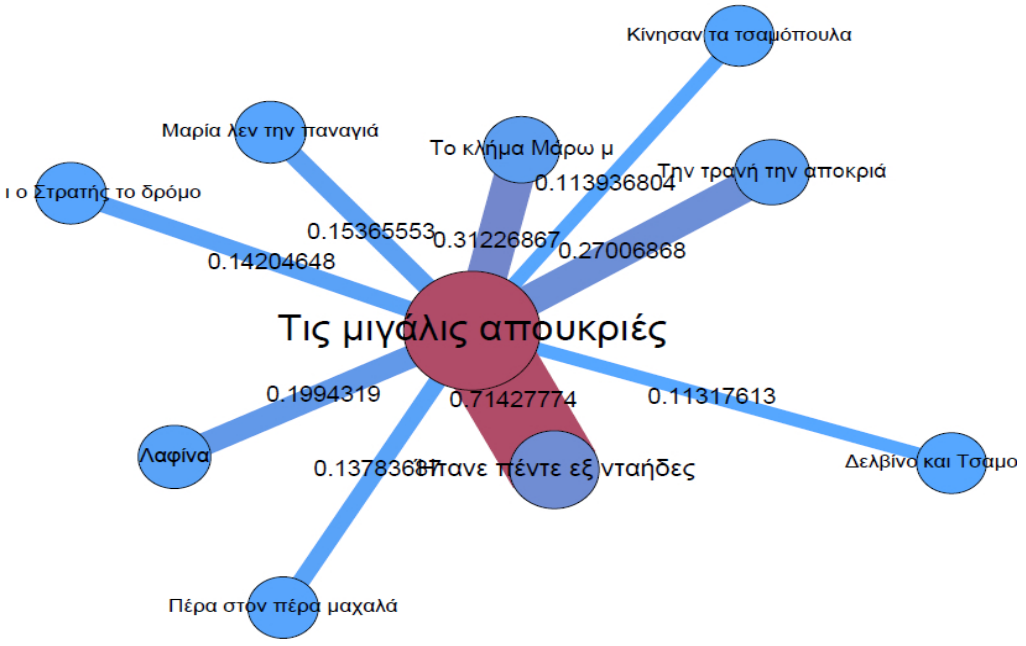
Τραγούδια που χρησιμοποιήθηκαν ως ερωτήματα πρς την LSI	Θεματική περιοχή στην οποία ανήκουν
1. Είχα μιαν αγάπη	Αγάπης
2. Ένας λεβέντης στο Μωριά	Ανδριωσύνης
3. Τις μεγάλης απουκριές	Αποκριάς
4. Εγώ είμαι προσφυγάκι	Ξενιτιάς

Πίνακας 5.11: Τραγούδια που χρησιμοποιήθηκαν ως ερωτήματα στην LSI και θεματικές στις οποίες ανήκουν

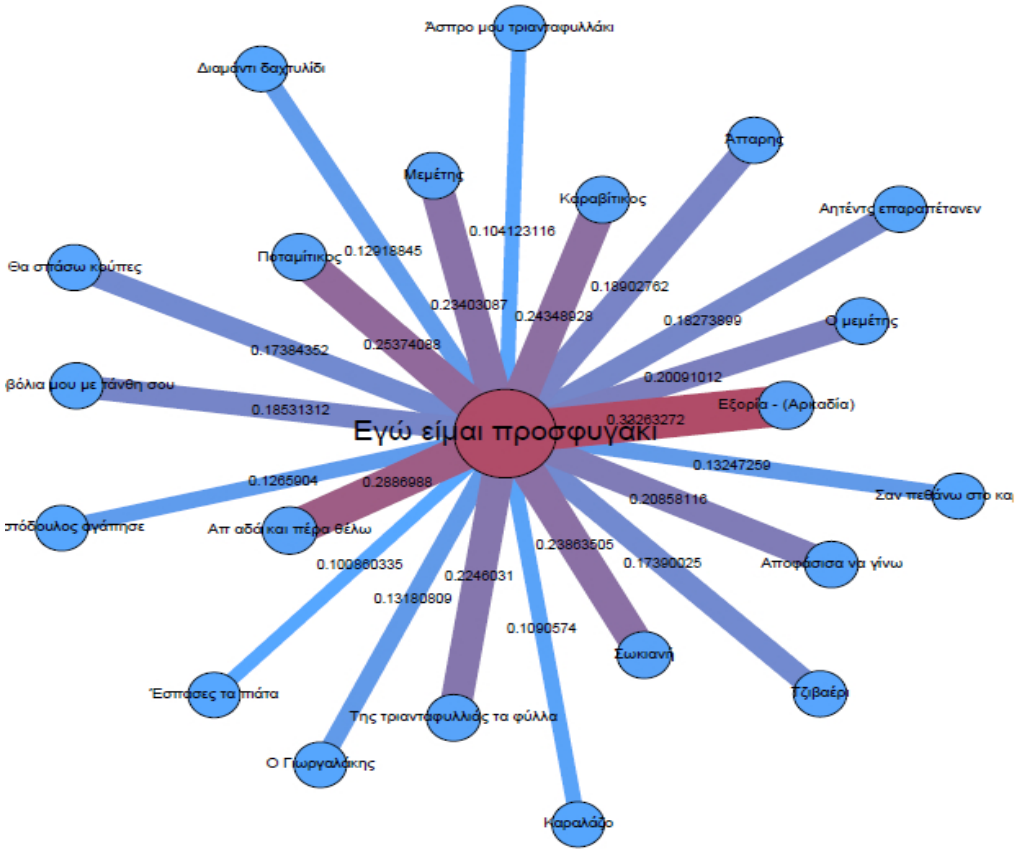


Σχήμα 5.12: Απεικόνιση ερωτήματος θεματικής περιοχής της ***αγάπης*** και των τραγουδιών που βρίσκονται εννοιολογικά κοντά σε αυτό, ως προς όλη την συλλογή των 388 τραγουδιών

Σχήμα 5.13: Απεικόνιση ερωτήματος θεματικής περιοχής της ***ανδριωσύνης*** και των τραγουδιών που βρίσκονται εννοιολογικά κοντά σε αυτό, ως προς όλη την συλλογή των 388 τραγουδιών



Σχήμα 5.14: Απεικόνιση ερωτήματος θεματικής περιοχής της ***αποκριάς*** και των τραγουδιών που βρίσκονται εννοιολογικά κοντά σε αυτό, ως προς όλη την συλλογή των 388 τραγουδιών



Σχήμα 5.15: Απεικόνιση ερωτήματος θεματικής περιοχής της ***προσφυγιάς*** και των τραγουδιών που βρίσκονται εννοιολογικά κοντά σε αυτό, ως προς όλη την συλλογή των 388 τραγουδιών

Εφόσον πλέον απαντήσαμε στο ερώτημα *"Ποια είναι τα πιο αντιπροσωπευτικά τραγούδια για μια θεματική"*; Το επόμενο βήμα είναι η διερεύνηση ομοιοτήτων μεταξύ τραγουδιών καθώς και η εύρεση τυχόν κοινοτήτων που μπορεί να προκύψουν.

5.4.5.3 Εύρεση Κοινοτήτων/Communities & Ομοιοτήτων Μεταξύ Τραγουδιών

Γνωρίζοντας πως η ιδανικότερη συνθήκη για την εύρεση ομοιοτήτων μεταξύ τραγουδιών του dataset θα ήταν μέσω της μετατροπής του $W^{t\times d}$ term document matrix σε έναν low rank approximation matrix $W_k^{d\times d}$. Από τη σχέση 3.12 στη σελίδα 26 γνωρίζουμε ότι $W_k = U_k \ S_k \ V_k^T$ οπότε:

$$W_k^{t\times d} = U_k \;\; S_k \;\; V_k^T = \left(U_k S_k^{\frac{1}{2}}\right)\left(S_k^{\frac{1}{2}} V_k^T\right) = \left(U_k S_k^{\frac{1}{2}}\right)^{t\times k}\left(V_k^T S_k^{\frac{1}{2}}\right)^{k\times d} \tag{5.5}$$

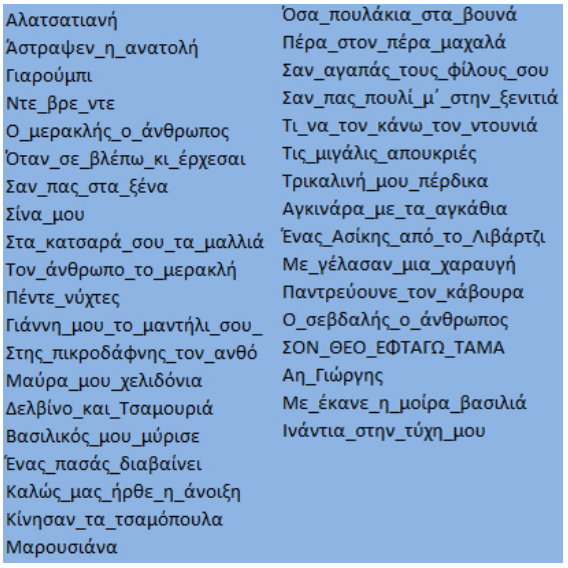
$$T_k^{t\times d} = \left(U_k S_k^{\frac{1}{2}}\right)^{t\times k}\left(V_k^T S_k^{\frac{1}{2}}\right)^{k\times d} \tag{5.6}$$

$$\tilde{D}_k^{k\times d} = \left(S_k^{\frac{1}{2}}\right)^{k\times k}\left(U_k^T\right)^{k\times t}\left(T_k\right)^{t\times d} \tag{5.7}$$

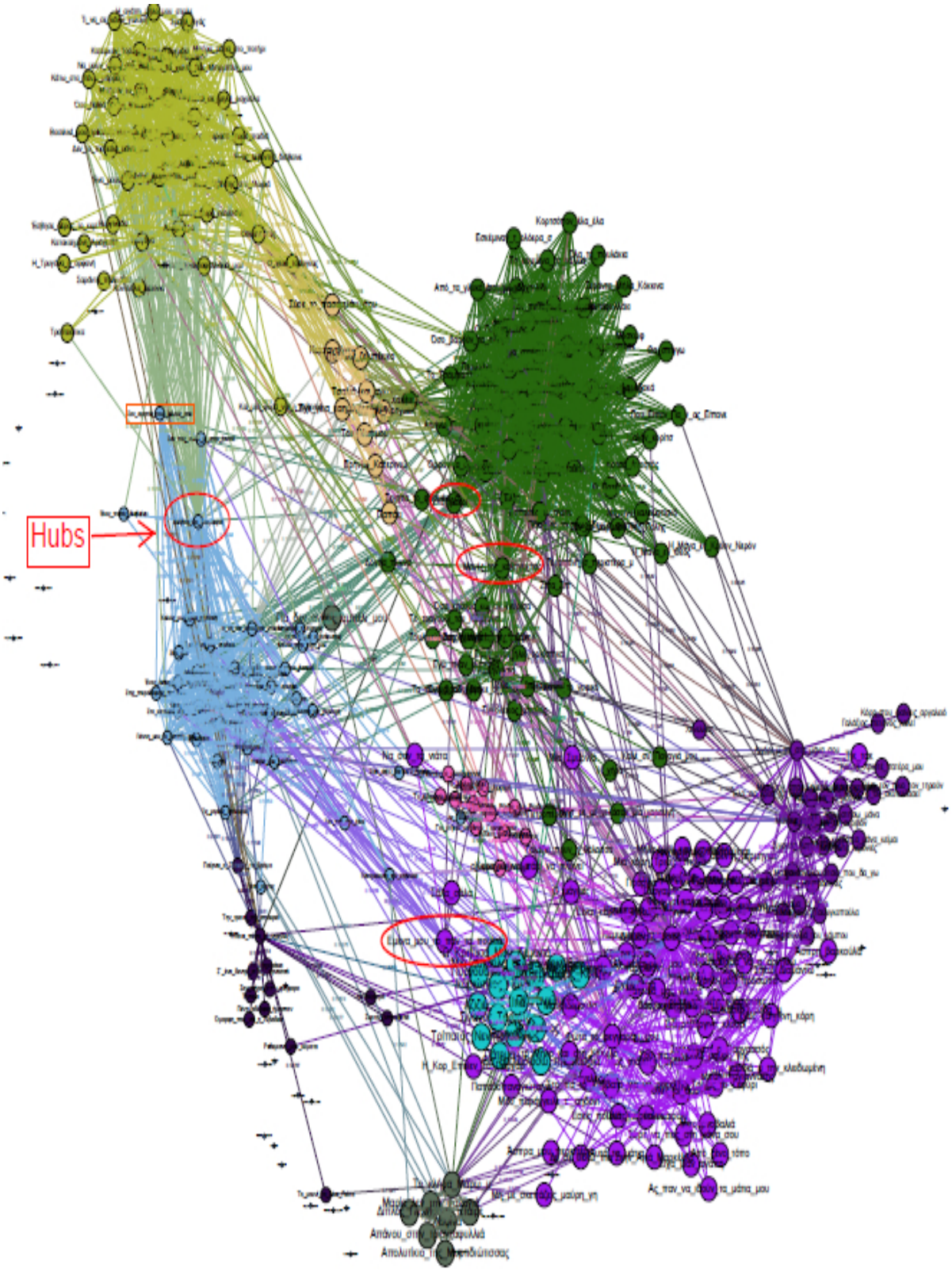
$$W_k^{d\times d} = D^T D \tag{5.8}$$

Σχήμα 5.16: Τύπος παραγωγής του πίνακα $W_k^{d\times d}$

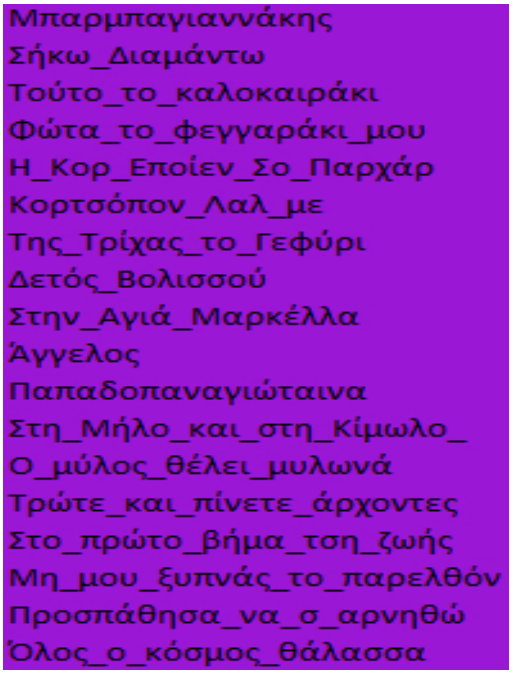
Έχοντας πλέον τον πίνακα $W_k^{d\times d}$ που δηλώνει similarities μεταξύ τραγουδιών τον εισάγουμε στο δωρεάν λογισμικό ανάλυσης δικτύων Gephi [γεπ] προς περαιτέρω επεξεργασία και ανάλυση. Για την διάταξη του γράφου χρησιμοποιήσαμε τον αλγόριθμο [φορ, φορ] ***Force Atlas 2 & Fruchterman-Reingold*** [φρυ] μέσα από το περιβάλλον του Gephi [γεπ], ενώ για την ανακάλυψη των communities μέσω clustering χρησιμοποιήσαμε τον αλγόριθμο Chinese Whispers [ςηι]. Τέλος το μέγεθος των κόμβων του γράφου καθορίστηκε μέσω του υπολογισμού μιας στατιστικής του γράφου που ονομάζεται "*eigencentrality*". Στα σχήματα που ακολουθούν παρουσιάζεται ολόκληρος ο γράφος καθώς και τα τραγούδια που ανήκουν σε κάθε cluster, χρησιμοποιώντας την ίδια χρωματική κωδικοποίηση που περιέχει ο γράφος στην σελίδα 79.



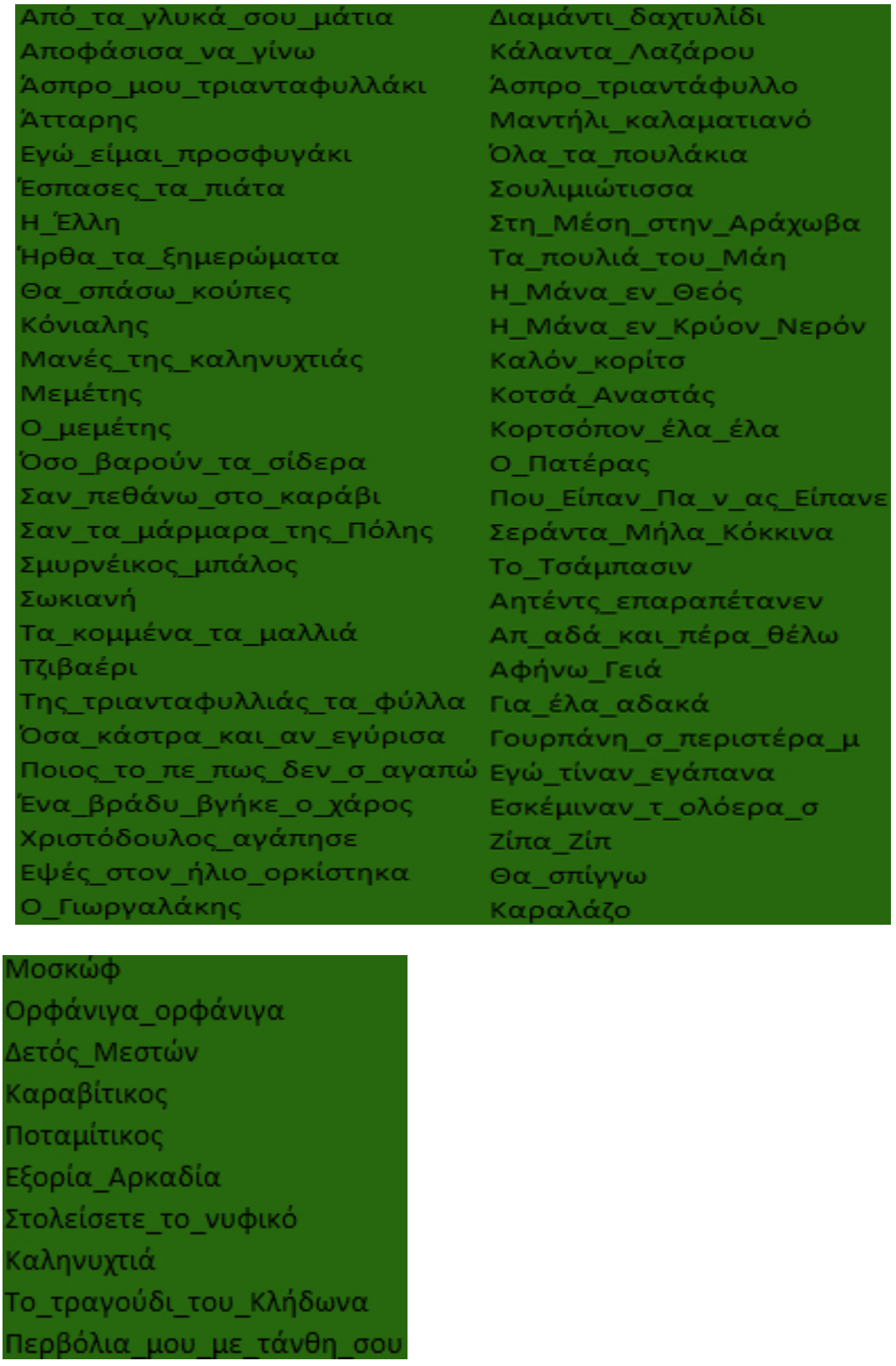
Σχήμα 5.17: Σύνολο **36** τραγουδιών που περιλαμβάνονται στο **cluster 1**



Σχήμα 5.18: Κοινότητες



Σχήμα 5.19: Σύνολο **72** τραγουδιών στο **cluster 2**



Σχήμα 5.20: Σύνολο **64** τραγουδιών στο **cluster 3**

Σχήμα 5.21: Σύνολο **28** τραγουδιών στο **cluster 4 (έντονο μοβ χρώμα)** & **35** στο **cluster 5 (λαχανί χρώμα)**

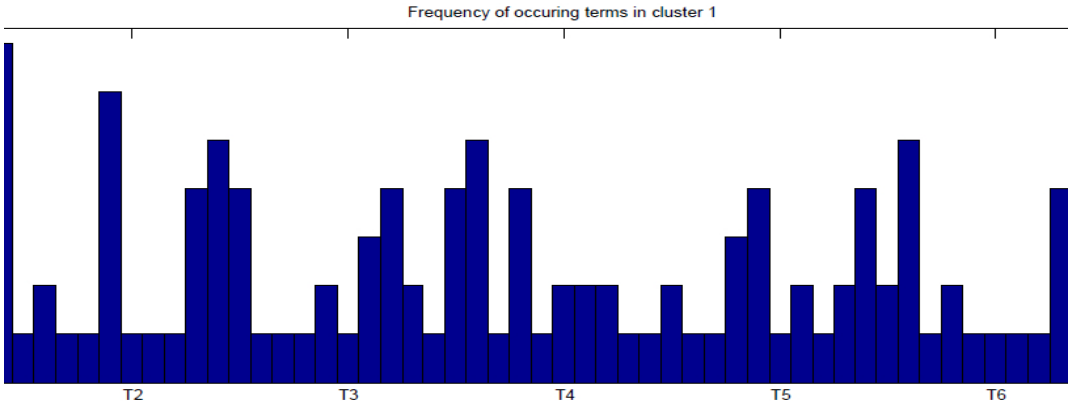
Σχήμα 5.22: Σύνολο **13** τραγουδιών στο (**cluster 6**), **12** στο (**cluster 7**), **10** στο (**cluster 8**), **9** στο (**cluster 9**) & **7** στο (**cluster 10**)

	$cluster_1$	$cluster_2$	$cluster_3$	$cluster_4$	$cluster_5$
Μικρασιάτικα	27.77%(10/36)	33.33%(24/72)	32.81%(21/64)	10.71%(3/28)	2.85%(1/35)
Θρακιώτικα	2.77%(1/36)	5.55%(4/72)	1.56%(1/64)	7.14%(2/28)	2.85%(1/35)
Ηπειρώτικα	11.11%(4/36)	6.94%(5/72)	3.12%(2/64)	7.14%(2/28)	20%(7/35)
Μακεδονίτικα		1.38%(1/72)	3.12%(2/64)	7.14%(2/28)	
Ρουμελιώτικα	2.77%(1/36)	11.11%(8/72)	3.12%(2/64)	7.14%(2/28)	22.85%(8/35)
Θεσσαλιώτικα	33.33%(12/36)	8.33%(6/72)	1.56%(1/64)	10.71%(3/28)	20%(7/35)
Μωραΐτικα	8.33%(3/36)	13.88%(10/72)	9.37%(6/64)	28.57%(8/28)	25.71%(9/35)
Ποντιακά	5.55%(2/36)	4.16%(3/72)	32.81%(21/64)	17.85%(5/28)	2.85%(1/35)
Νησιώτικα	2.77%(1/36)	8.33%(6/72)	10.93%(7/64)	3.57%(1/28)	
Κρητικά	5.55%(2/36)	6.94%(5/72)	1.56%(1/64)		2.85%(1/35)
Συνολ. Τραγ.	36	72	64	28	35

	$cluster_6$	$cluster_7$	$cluster_8$	$cluster_9$	$cluster_{10}$
Μικρασιάτικα	7.69%(1/13)	8.33%(1/12)		33.33%(3/9)	
Θρακιώτικα	15.38%(2/13)	8.33%(1/12)		11.11%(1/9)	
Ηπειρώτικα				11.11%(1/9)	14.28%(1/7)
Μακεδονίτικα	15.38%(2/13)				14.28%(1/7)
Ρουμελιώτικα	15.38%(2/13)			11.11%(1/9)	
Θεσσαλιώτικα	15.38%(2/13)				28.57%(2/7)
Μωραΐτικα	7.69%(1/13)	16.66%(2/12)	10%(1/10)	11.11%(1/9)	
Ποντιακά		33.33%(4/12)		11.11%(1/9)	14.28%(1/7)
Νησιώτικα	15.38%(2/13)	25%(3/12)	90%(9/10)	11.11%(1/9)	28.57%(2/7)
Κρητικά	7.69%(1/13)	8.33%(1/12)			
Συνολ. Τραγ.	13	12	10	9	7

Πίνακας 5.12: Κατανομή τραγουδιών ανά γεωγραφική περιοχή μέσα σε καθένα από τα 10 clusters

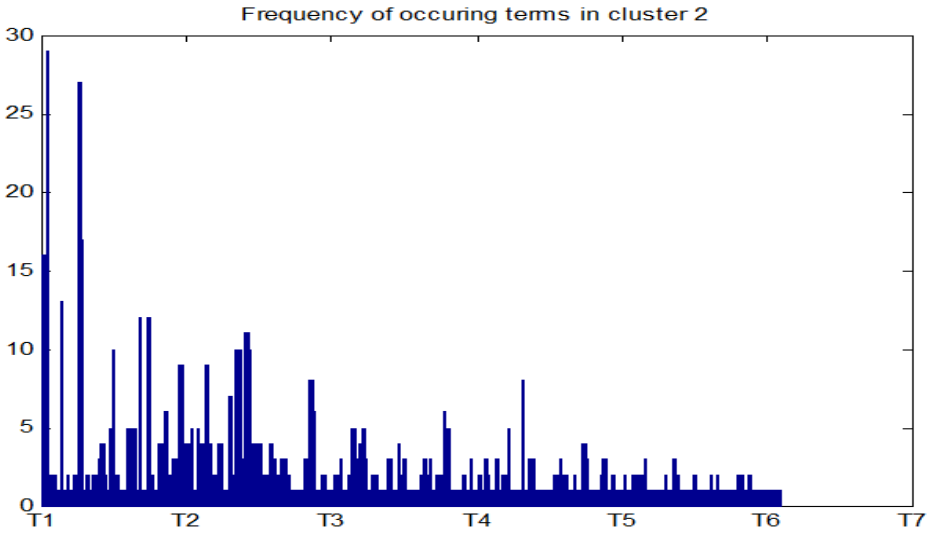
Οι μεταβλητές $T1, T2, T3, \ldots$ στο ΣΧΗΜΑ 5.23 δηλώνουν τους πιο πολυσύχναστους όρους στο **cluster 1**. Στην σελίδα 86 στον ΠΙΝΑΚΑ 5.13 παρουσιάζονται οι **5** πιο πολυσύχναστη όροι καθώς και η αντιστοίχιση των μεταβλητών $T1, T2, T3, \ldots$ στους πραγματικούς όρους.



Σχήμα 5.23: Συχνότητα εμφάνισης όρων στο cluster 1

Variables	Terms	Συχνότητα Εμφάνισης
T1	ΑΓΑΠ	22
T4	ΑΗΔΟΝ	7
T9	ΑΡΑΠ	6
T14	ΒΓ	5
T26	ΓΥΡΙΖ	5

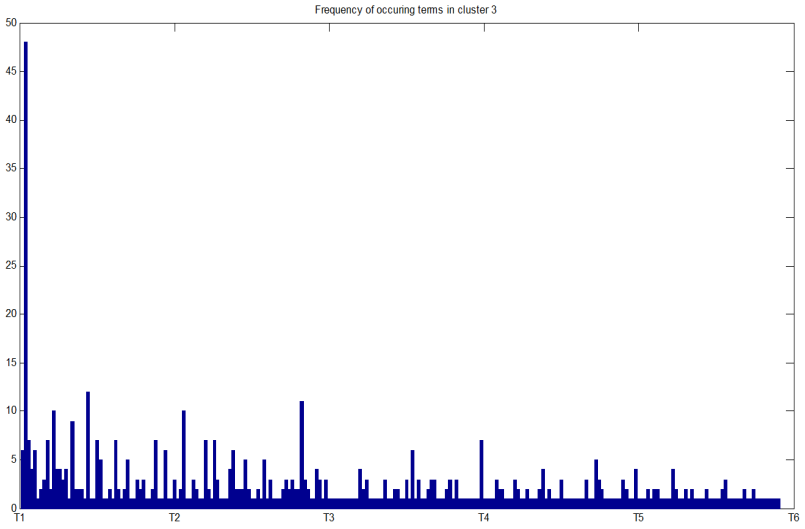
Πίνακας 5.13: Παρουσίαση των **5** πιο πολυσύχναστων όρων στο **cluster 1** σύμφωνα με το ΣΧΗΜΑ 5.23



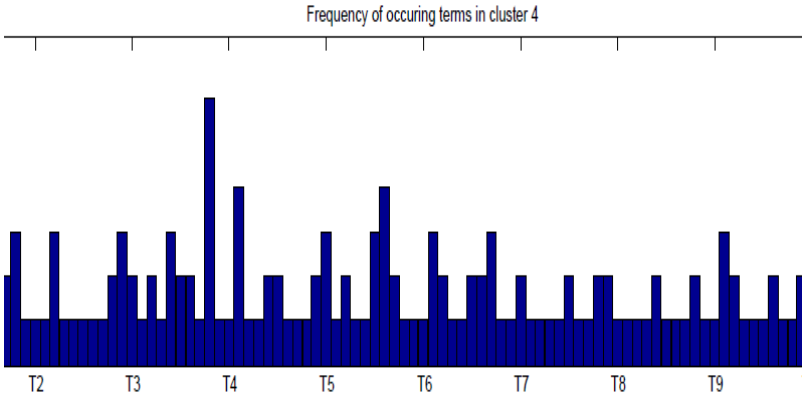
Σχήμα 5.24: Συχνότητα εμφάνισης όρων στο cluster 2

Variables	Terms	Συχνότητα Εμφάνισης
T2	ΑΓΑΠ	29
T13	ΜΑΤ	27
T14	ΜΑΥΡ	17
T1	ΠΟΥΛ	16
T7	ΘΕΛ	13

Πίνακας 5.14: Παρουσίαση των **5** πιο πολυσύχναστων όρων στο **cluster 2** σύμφωνα με το ΣΧΗΜΑ 5.24



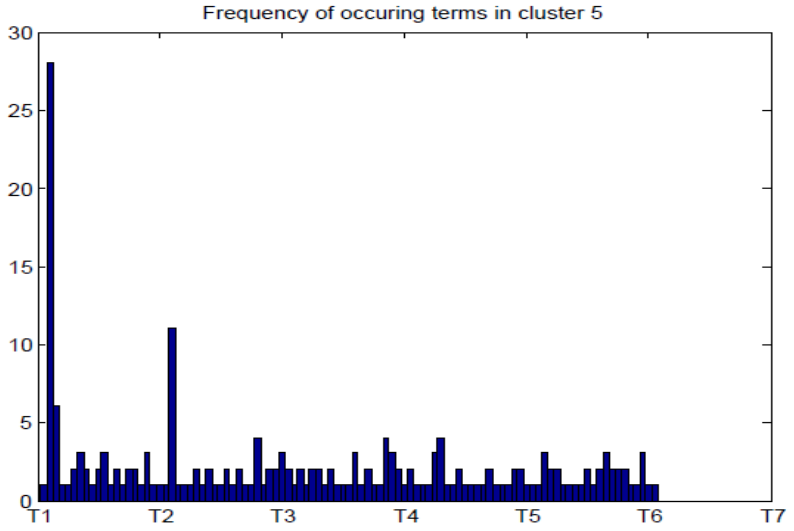
Σχήμα 5.25: Συχνότητα εμφάνισης όρων στο cluster 3



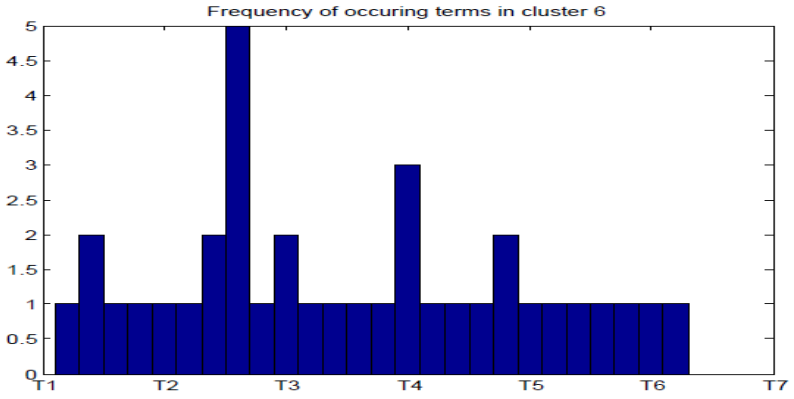
Σχήμα 5.26: Συχνότητα εμφάνισης όρων στο cluster 4

Variables	Terms	Συχνότητα Εμφάνισης	Variables	Terms	Συχνότητα Εμφάνισης
T2	ΑΜ	29	T1	ΜΑΝ	20
T13	ΑΦ	27	T6	ΠΑΛ	7
T14	ΕΙΠ	17	T28	ΚΑΝ	6
T1	ΠΟΛ	16	T31	ΜΑΤ	4
T7	ΠΕΘΑΝ	13	T46	ΠΑΙΔ	4

Πίνακας 5.15: Παρουσίαση των **5** πιο πολυσύχναστων όρων στο **cluster 3** (αριστερός υπο-πίνακας) σύμφωνα με το ΣΧΗΜΑ 5.25 & των **5** πιο πολυσύχναστων όρων στο **cluster 4** (δεξιός υπο-πίνακας) σύμφωνα με το ΣΧΗΜΑ 5.26



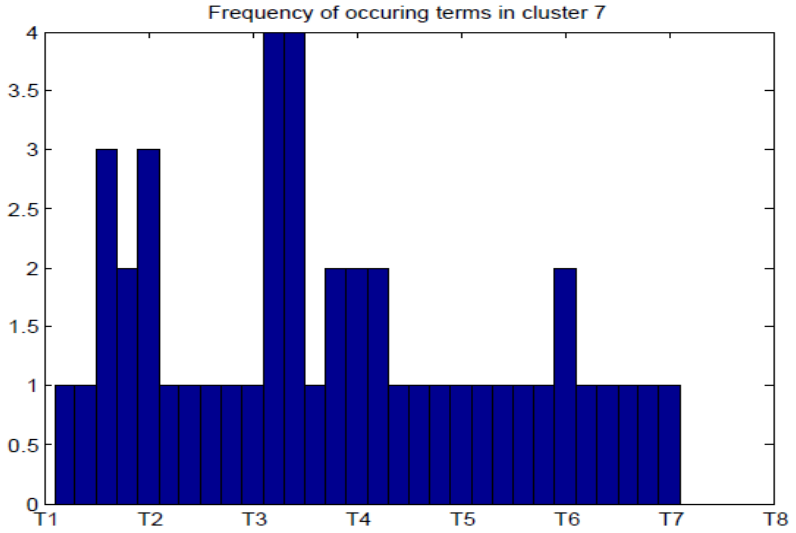
Σχήμα 5.27: Συχνότητα εμφάνισης όρων στο cluster 5



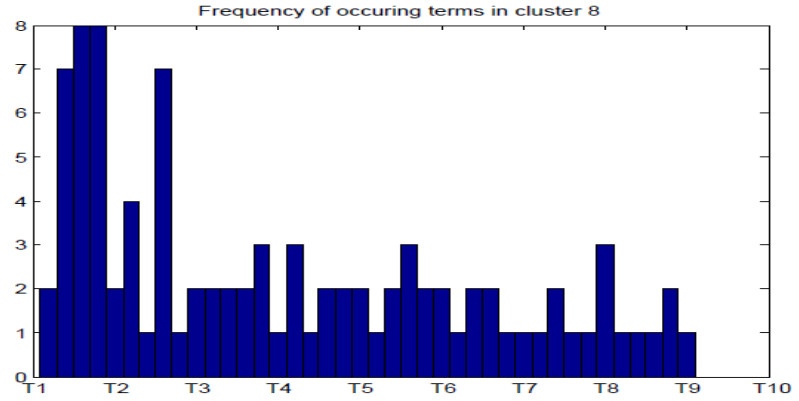
Σχήμα 5.28: Συχνότητα εμφάνισης όρων στο cluster 6

Variables	Terms	Συχνότητα Εμφάνισης	Variables	Terms	Συχνότητα Εμφάνισης
T2	ΜΩΡ	28	T8	ΠΗΡ	5
T22	ΛΕΒΕΝΤ	11	T15	ΒΡΙΣΚ	3
T3	ΑΓΑΠ	6	T7	ΠΑΙΡΝ	2
T36	ΔΕΛ	4	T2	ΑΔΕΡΦ	2
T57	ΓΕΡ	4	T10	ΜΠΡ	2

Πίνακας 5.16: Παρουσίαση των **5** πιο πολυσύχναστων όρων στο **cluster 5** (αριστερός υπο-πίνακας) σύμφωνα με το ΣΧΗΜΑ 5.27 & των **5** πιο πολυσύχναστων όρων στο **cluster 6** (δεξιός υπο-πίνακας) σύμφωνα με το ΣΧΗΜΑ 5.28



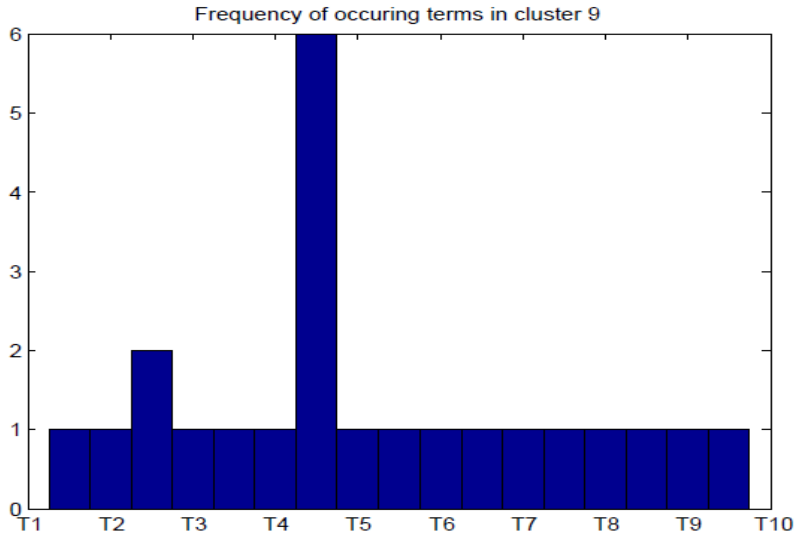
Σχήμα 5.29: Συχνότητα εμφάνισης όρων στο cluster 7



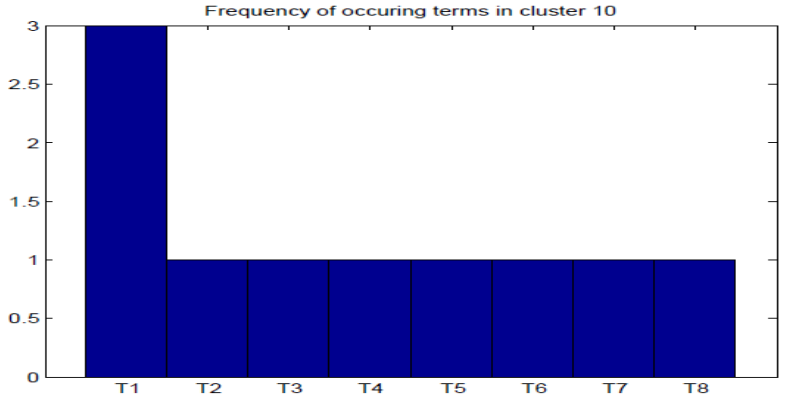
Σχήμα 5.30: Συχνότητα εμφάνισης όρων στο cluster 8

Variables	Terms	Συχνότητα Εμφάνισης	Variables	Terms	Συχνότητα Εμφάνισης
T11	ΠΕΡ	4	T3	ΧΟΡΕΥ	8
T12	ΝΕΡ	4	T4	ΑΝΤΡ	8
T5	ΠΑΛ	3	T2	ΧΟΡ	7
T16	ΦΙΛ	2	T8	ΓΥΝΑΙΚ	7
T25	ΗΛΙ	2	T6	ΓΙΑΤ	4

Πίνακας 5.17: Παρουσίαση των **5** πιο πολυσύχναστων όρων στο **cluster 7** (αριστερός υπο-πίνακας) σύμφωνα με το ΣΧΗΜΑ 5.29 & των **5** πιο πολυσύχναστων όρων στο **cluster 8** (δεξιός υπο-πίνακας) σύμφωνα με το ΣΧΗΜΑ 5.30



Σχήμα 5.31: Συχνότητα εμφάνισης όρων στο cluster 9



Σχήμα 5.32: Συχνότητα εμφάνισης όρων στο cluster 10

Variables	Terms	Συχνότητα Εμφάνισης	Variables	Terms	Συχνότητα Εμφάνισης
T7	ΓΕΙΤ	6	T1	ΜΑΡ	3
T3	ΑΓΑΠ	2	T2	ΤΑΞ	1
T9	ΦΕΡ	1	T4	ΚΛΗΜ	1
T11	ΠΑΤ	1	T5	ΣΤΑΦΥΛ	1
T12	ΡΑΧΟΥΛ	1	T6	ΔΙΠΛ	1

Πίνακας 5.18: Παρουσίαση των **5** πιο πολυσύχναστων όρων στο **cluster 9** (αριστερός υπο-πίνακας) σύμφωνα με το ΣΧΗΜΑ 5.31 & των **5** πιο πολυσύχναστων όρων στο **cluster 10** (δεξιός υπο-πίνακας) σύμφωνα με το ΣΧΗΜΑ 5.32

5.5 Περαιτέρω Συζήτηση

Σύμφωνα με τα πειράματα που διεξήχθησαν παραπάνω καθώς και άλλα πειράματα που έχουν διεξαχθεί κατά καιρούς από άλλους ερευνητές,αποδεικνύεται ότι η LSA παρουσιάζει έναν αριθμό πλεονεκτημάτων που την καθιστούν ανώτερη σε σύγκριση με τις παραδοσιακές τεχνικές αντιστοίχισης εφάμιλλων όρων term matching. Παρόλα αυτά η συγκεκριμένη τεχνική εξακολουθεί να έχει κάποιους περιορισμούς που θα συζητηθούν αργότερα.

5.5.1 Πλεονεκτήματα της LSA

1 **Καλύτερη Επίδοση**

Στο δεύτερο πείραμα μας, κύριος στόχος του οποίου ήταν να διαπιστώσουμε ποια από τις τεχνικές θα μπορούσε να ανακτήσει το μεγαλύτερο αριθμό των σχετικών εγγράφων,η προσέγγιση της LSA βρέθηκε να είναι ανώτερη σε σύγκριση με το απλό ταίριασμα εφάμιλλων όρο term matching. Σε γενικές γραμμές η εφαρμογή της LSA που χρησιμοποιείται στο MATLAB ανέκτησε κατά μέσο όρο περισσότερα από 42 έγγραφα (high recall), αν και μόνο τα πρώτα 10 έγγραφα λήφθηκαν υπόψη σε κάθε περίπτωση. Ήταν δύσκολο να μετρηθεί η ακριβής τιμές ακρίβειας και η ανάκλησης λόγω του περιορισμού των χρησιμοποιούμενων μέσων και το μέγεθος της συλλογής εγγράφων η οποία θα δώσει αποτελέσματα που δεν είναι ολοκληρωμένα.

Καλύτερα αποτελέσματα επιδόσεων θα μπορούσαν να έχουν επιτευχθεί αν το LSA MATLAB πείραμά μας επέτρεπε την αναζήτηση περισσότερων από ενός όρου λεξικούς χρησιμοποιώντας ένα μόνο ερώτημα και αν η συλλογή των εγγράφων ήταν αρκετά μεγάλη με πολλαπλά θέματα. Βασιζόμενη σε άλλα πειράματα που έχουν γίνει στη MED (μια συλλογή από ιατρικές περιλήψεις) και CISI (ένα σύνολο 1460 περιλήψεων επιστήμης πληροφορικής) μεταξύ άλλων συνόλων δεδομένων, όπως περιγράφεται στο [19],η απόδοση της LSA γενικά έχει υπολογιστεί ότι είναι ανώτερη σε σύγκριση με το απλό ταίριασμα εφάμιλλων όρων όσον αφορά την ανάκληση και την απόδοση ακριβείας (recall & precision).

Χρησιμοποιώντας για παράδειγμα, τη MED, η LSA παρουσίασε μια βελτίωση ακριβείας 13% σε σύγκριση με την τεχνική term matching. Η ανωτερότητα της LSA εντοπίζεται στην ικανότητά της να ταιριάζει σωστά τα ερωτήματα με τα σχετικά έγγραφα με βάση την τοπική τους εννοιολογική σημασία, ακόμη και αν

τα ερωτήματα και τα έγγραφα χρησιμοποιούν διαφορετικούς όρους. Το συνυφασμένο στάδιο προ επεξεργασίας της LSA επίσης βελτιώνει την απόδοση της καθώς λαμβάνει υπόψη τη συνολική κατανομή ενός όρου/λέξης σύμφωνα με το περιεχόμενο/πλαίσιο χρήσης της, ανεξάρτητα από συσχετισμούς που υπάρχουν με άλλους όρους/λέξεις.

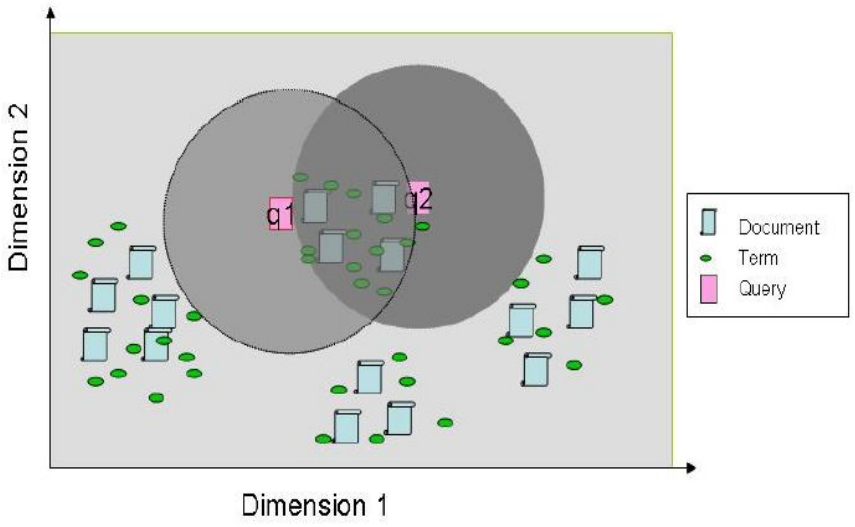
2 **Συνωνυμία (Synonymy)**

Η τεχνική LSA σε αντίθεση με τις παραδοσιακές μεθόδους ταιριάσματος όρων είναι σε θέση να αντιμετωπίσει το πρόβλημα της συνωνυμίας που προκύπτει ως ένα βαθμό. Οι επιμέρους όροι [19] αντικαθίσταται ως δείκτες των εγγράφων από ανεξάρτητες "τεχνητές έννοιες"που καθορίζονται από οποιεσδήποτε λέξεις/όρους (ή έγγραφα) ή συνδυασμούς αυτών. Αυτό επιτρέπει την ανάκτηση των σχετικών εγγράφων που δεν περιέχουν τους όρους του ερωτήματος, ή των οποίων οι περιεχόμενοι όροι προσδιορίζεται από άλλους όρους στο ερώτημα ή το έγγραφο, αλλά όχι και στα δύο.

Στην LSA,οι όροι/λέξεις που συνυπάρχουν λόγο παρόμοιας σημασίας, βρίσκονται κοντά η μια στην άλλη στο χώρο μειωμένης διάστασης ακόμη κι αν δεν μπορούν να συνυπάρχουν στα ίδια έγγραφα. Αυτό καθιστά δυνατή την ανάκτηση γειτονικών εγγράφων (το συνημίτονο της γωνίας που ορίζει την έννοια της γειτνίασης είναι σχετικό βάση των διαστάσεων και του σύνολου των δεδομένων που χρησιμοποιούνται) κάνοντας χρήση οποιουδήποτε από αυτούς τους όρους. Στο 4ο πείραμα μας, διαπιστώθηκε ότι ορισμένοι όροι του ερωτήματος οδήγησαν στην ανάκτηση όχι μόνο των εγγράφων που περιέχουν αυτούς τους όρους, αλλά και έγγραφα που περιέχουν όμοιους ή παρεμφερείς όρους, π.χ. για τον όρο soccer ανακτήθηκαν επίσης έγγραφα που περιέχουν τον όρο football.

Αυτό σημαίνει ότι ο όρος soccer και football βρίσκονται κοντά ο ένας στον άλλο στο σημασιολογικό χώρο και ως εκ τούτου θα ανακτήσει παρόμοια έγγραφα. Αυτή η σχέση μεταξύ soccer και football προέρχεται από τη μεταβατική σχέση [20]: ο όρος soccer συνυπάρχει με τον όρο football και ο όρος football συνυπάρχει με τον όρο human. Η εξάλειψη λέξεων στίξις stemming που διενεργήθηκε στο στάδιο της προ επεξεργασίας της LSA βοηθά επίσης στο πρόβλημα χειρισμού της συνωνυμίας, δεδομένου ότι βοηθά στη σύλληψη των πιθανών συνώνυμων όρων. Στην παρκάτω εικόνα τα ερωτήματα $q1$ και $q2$ περιέχει πανομοιότυπους όρους και ως εκ τούτου θα ανακτήσουν παρόμοια έγγραφα τα οποία διαφορετικά

δεν θα ήταν εφικτό με τις παραδοσιακές τεχνικές ταιριάσματος εφάμιλλων όρων term matching.



Σχήμα 5.33: Παράδειγμα χειρισμού της συνωνυμίας (synonymy) από την LSA χρησιμοποιώντας τα ερωτήματα $q1$ και $q2$. Οι γκρι κύκλοι δείχνουν έγγραφα που είναι πιθανό να ανακτηθούν από κάθε ερώτημα. Λόγω της τομής των 2 εγγράφων, 4 είναι αυτά τα έγγραφα που θα ανακτηθούν είτε από το ένα ερώτημα είτε από το άλλο.

3 Ελαχιστοποίηση Αποθήκευσης Χώρου

Συγκρίνοντας την προσέγγιση της LSA με το διανυσματικό μοντέλο χώρου για έγγραφα, αντιλαμβανόμαστε πως η LSA εμφανίζει καλύτερη αξιοποίηση του αποθηκευτικού χώρου.

Η κατάργηση των κοινών όρων (όροι με λιγότερη διακριτική ισχύ), η διαδικασία προ επεξεργασίας αφαίρεσης λέξεων/όρων στίξις stemming process που παρουσιάστηκε στο 1ο πείραμα καθώς και η μείωση διάστασης που είναι εγγενής στην προσέγγιση της LSA συμβάλλουν σημαντικά στην μειώσει των απαιτήσεων αποθήκευσης μιας μεγάλης συλλογής εγγράφων. Στο 1ο πείραμα μας, διαπιστώθηκε ότι ο term document matrix $A^{t\times d}$ περιέχει πολλαπλές μηδενικές εγγραφές (περίπου 89%).

Στο διανυσματικό μοντέλο χώρου αυτός ο πίνακας θα έπρεπε να αποθηκευτεί με όλα αυτά τα μηδενικά. Το συγκεκριμένο πρόβλημα διαχειρίζεται καλύτερα μέσω της μείωσης διαστάσεων από την LSA που επίσης καταργεί το θόρυβο στο σημασιολογικό χώρο και καθιστά την LSA κατά μακράν καλύτερη του VSM στην ανάκτηση εγγράφων.

4 **Η LSA Μιμείται την Ανθρώπινη Γνώση**

Τα μέτρα ομοιότητας με βάση την LSA έχουν αποδειχθεί ότι μιμούνται την ανθρώπινη κρίση. Στο [6], χρησιμοποιήθηκε ένα LSA μοντέλο για την προσομοίωση της ανθρώπινης απόδοσης χρησιμοποιώντας το TOEFL (Test of English as a Foreign Language) και συγκρίνοντας τα αποτελέσματα με το μέσο όρο των εξεταζομένων. Η LSA βρέθηκε να πετυχαίνει 65% σωστή βαθμολογία στις απαντήσεις που δόθηκαν η οποία ήταν ταυτόσημη με το μέσο όρο βαθμολογίας ενός μεγάλου δείγματος φοιτητών από μη αγγλόφωνες χώρες που λαμβάνουν την συγκεκριμένη εξέταση προκειμένου να υποβάλλουν αίτηση για σπουδές σε αγγλόφωνες χώρες.

5.5.2 Περιορισμοί της τεχνικής LSA

1 **Επικαιροποίηση - Ενημέρωση**

Ένα πρόβλημα που μπορεί να αντιμετωπίσει κανείς όσον αφορά την προσέγγιση LSA είναι η ενημέρωση του term document matrix όταν προστίθενται νέα έγγραφα στην συλλογή. Η ενημέρωση συνήθως περιλαμβάνει τον εκ νέου υπολογισμό της SVD όποτε καταφθάνουν νέα έγγραφα, αλλά λόγω των χρονικών περιορισμών και των υπολογιστικών πόρων (μνήμη και ταχύτητα επεξεργαστή) που απαιτούνται, ο εκ νέου υπολογισμός της SVD γίνεται πολύ δαπανηρός.

Μια προσέγγιση αναδίπλωσης νέων εγγράφων και ερωτημάτων έχει προταθεί στο [19], όπου νέα έγγραφα τοποθετούνται στα κέντρα βάρους των όρων/λέξεων που περιλαμβάνουν ενώ νέοι όροι/λέξεις τοποθετούνται στα κέντρα βάρους των εγγράφων στα οποία εμφανίζονται, αλλά παρόλα αυτά εξακολουθεί να παραμένει άγνωστο για πόσες φορές αυτό μπορεί να επαναληφθεί χωρίς να χρειαστεί να επαναλάβουμε τον πλήρη υπολογισμό της SVD.

Αυτό συμβαίνει επειδή η προσθήκη νέων εγγράφων μέσω αυτής της προσέγγισης "αναδίπλωση" αδυνατεί να συλλάβει περιστατικά συσχετίσεων στα πρόσφατα έγγραφα που προστέθηκαν και αγνοεί νέους όρους/λέξεις που μπορεί να περιέχονται στα έγγραφα αυτά οδηγώντας σε υποβάθμιση της ποιότητας των αποτελεσμάτων της LSA.

2 Προσδιορισμός των k (παραγόντων) κατά τη Μείωση Διαστάσεων

Για προσεγγίσεις low-rank, θα πρέπει να επιλεγούν k διαστάσεις για την εκπροσώπηση των εγγράφων, όρων/λέξεων, καθώς και ερωτημάτων. Στα πειράματα που διενεργήθηκαν, το k επιλέχθηκε εμπειρικά να είναι $k = 10$ μετά από μια σειρά δοκιμών που έγιναν με βάση την εξεύρεση βέλτιστη αντιπροσώπευσης εγγράφων. Στις δοκιμές, η αναζήτηση του όρου ιατρικής με τιμές για k να ορίζονται ως 50, 40, 30, 20, 10 και 5 ανακτήθηκαν αντίστοιχα 4, 6, 6, 8, 8 και 7 σχετικά έγγραφα, καθιστώντας έτσι το 10 ως μια κατάλληλη τιμή για το k.

Η επιλογή του k ως 10 κινητοποιήθηκε επίσης από τις αναμενόμενες θεματικές ομάδες (clusters) στο μειωμένο σημασιολογικό χώρο βασιζόμενο στις 10 θεματικές ενότητες από τις οποίες προέρχονται τα έγγραφα. Ως εκ τούτου, εξακολουθεί να μην υπάρχει συγκεκριμένη τεχνική ή αλγόριθμος για την επιλογή του αριθμού k διατηρήσεις των διαστάσεων και ως συνέπεια παραμένει ένα εμπειρικό ζήτημα. Αν το k είναι πολύ μεγάλο μπορεί να έχουμε περισσότερο θόρυβο στο διανυσματικό χώρο, ενώ πολύ χαμηλό k μπορεί να οδηγήσει αφαίρεση σημαντικών πληροφοριών από παράγοντες.

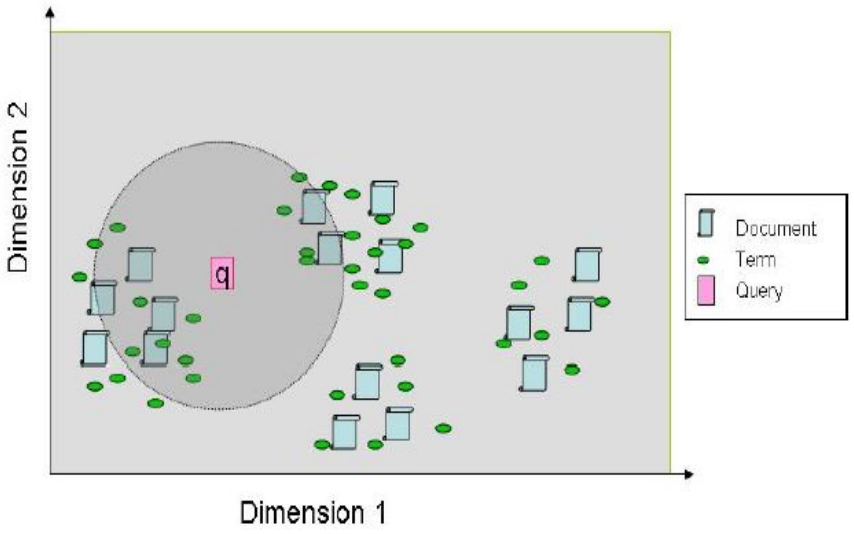
3 Πολυσημία (Polysemy)

Στην LSA, κάθε όρος αντιστοιχίζεται ως ένα σημείο στον λανθάνοντα σημασιολογικό χώρο. Για έναν όρο με πολλαπλές σημασίες, η αντιστοίχιση γίνεται με βάση το σταθμισμένο μέσο όρο των διαφορετικών εννοιών. Αυτό σημαίνει ότι μια σοβαρή στρέβλωση μπορεί να συμβεί όταν κάποιες από τις πραγματικές έννοιες δεν είναι σαν τη μέση έννοια.

Αυτό μπορεί να οδηγήσει είτε στην ανάκτηση υψηλού αριθμού εγγράφων που έχουν διαφορετικές θεματικές σημασίες, κάποια σχετικά και κάποια άσχετα προς το χρήστη (σε περίπτωση που οι επηρεαζόμενες ομάδες/clusters βρίσκονται κοντά) ή ακόμη στην ανάκτηση πολύ μικρού αριθμού ή και καθόλου εγγράφων «σε περίπτωση που οι επηρεαζόμενες ομάδες/clusters βρίσκονται μακριά μεταξύ τους).

Στο πείραμα 4, η αναζήτηση για τον όρο "τύπο"ανέκτησε μερικά έγγραφα σχετικά με πολιτική και μερικά για τον αθλητισμό. Λαμβάνοντας υπόψη αυτά τα

αποτελέσματα, είναι δύσκολο να απαντήσει κανείς αυτόματα ποιες θεματικές ενότητες ικανοποιούν το αναμενόμενο νόημα του χρήστη και ως εκ τούτου να ανακτήσετε μόνο τα πολύ σχετικά έγγραφα που βασίζονται στην εν λόγω έννοια. Το παράδειγμα του προβλήματος της πολυσημίας παρουσιάζεται παρακάτω, όπου το ερώτημα (q) βρίσκεται σχεδόν ανάμεσα σε δύο ομάδες/clustres (θεματικές ενότητες), συνεπώς, καταλήγοντας στην ανάκτηση διαφορετικών εγγράφων όσον αφορά την θεματική τους ενότητα.



Σχήμα 5.34: Παράδειγμα χειρισμού της πολυσημίας (polysemy) από την LSA

Η λύση στο πρόβλημα της πολυσημίας θα ήταν να συνειδητοποιήσουμε ότι ένας όρος/λέξη έχει πολλές διακριτές σημασίες και να τον υπό κατηγοριοποιήσουμε και τοποθετήσουμε σε διάφορα σημεία στον λανθάνοντα σημασιολογικό χώρο. Αλλά, εντούτοις, δεν έχει επιτευχθεί στο συγκεκριμένο έργο.

4 Ταξινόμηση Όρων/Λέξεων

Η φύση του μοντέλου "bag of words" της LSA δεν κάνει χρήση της κατάταξης/σειράς εμφάνισης των λέξεων που μπορεί να οδηγήσει στην άγνοια των συντακτικών, λογικών και ρεαλιστικών μη γλωσσικών συνεπαγωγών. Συνεπώς παραλείπει τις έννοιες που παράγονται από τη σειρά των λέξεων για να επικεντρωθεί μόνο στο πώς συσχετίζονται οι διαφορές στην επιλογή των λέξεων και οι διαφορές στις έννοιες κειμένου. Αυτό μπορεί να οδηγήσει σε ατελή αναπαράσταση και σφάλματα ανάκτησης. Για παράδειγμα, τα δύο παρακάτω έγγραφα θα τύγχαναν ίσης μεταχείρισης:

"***Το παλιό τοπίο στη μικρή πόλη***
Η παλιά μικρή πόλη στο τοπίο"

Αυτό το πρόβλημα μπορεί να αντιμετωπιστεί σε κάποιο βαθμό από τη χρήση biwords και ευρετηρίων/indexes φράσεων. Αυτό περιλαμβάνει προσθήκη biwords (2 όροι/λέξεις που έχουν μια κοινή έννοια και δεν θα ήταν δυνατόν να το καταλάβουμε εάν είχαν αντιμετωπιστεί ξεχωριστά) και φράσεων (διάφοροι όροι που έχουν μια κοινή έννοια δεν θα ήταν δυνατόν να το καταλάβουμε εάν είχαν αντιμετωπιστεί ξεχωριστά) στον term document matrix. Για να επιτευχθεί αυτό όλα τα biwords και φράσεις που εμφανίζονται στα έγγραφα πρέπει να είναι γνωστά πριν ξεκινήσει η διαδικασία. Αυτή η διαδικασία είναι μάλλον δαπανηρή ως προς την υλοποίηση της.

Κεφάλαιο 6

Συμπεράσματα & Μελλοντικές Βελτιώσεις

6.1 Συμπεράσματα

Σε αυτή την εργασία διερευνήθηκε η εφαρμογή της LSA ως μέθοδος ανακάλυψης γνώσης από δεδομένα κειμένου, βάση της επίδοσης της και συγκρινόμενη με την τεχνική ανάκτησης εγγράφων term-matching. Από τα πειράματα και τα αποτελέσματα που ολοκληρώθηκαν στο κεφάλαιο 5 συμπεραίνουμε πως η LSA έχει μεγαλύτερη απόδοση από την τεχνική term-matching και επίσης λόγω της εύκολης υλοποίησης της τυγχάνει ευρείας εφαρμογής. Ακόμη ένα πλεονέκτημα που διευρύνει την χρησιμότητα της LSA είναι η ικανότητα της να διαχειρίζεται το πρόβλημα της πολυσημίας και συνωνυμίας που προκύπτει πολύ συχνά σε δεδομένα κειμένου. Παρόλα αυτά ένα αντίστοιχα μεγάλο μειονέκτημα της LSA είναι η πολυπλοκότητα εφαρμογής της σε συλλογή με πολύ μεγάλο αριθμό εγγράφων λόγω της αποσύνθεσης SVD η οποία είναι αρκετά χρονοβόρα $O(n^3)$ και σε ορισμένες περιπτώσεις καθιστά την εφαρμογή της LSA ανέφικτη.

6.2 Μελλοντικές Βελτιώσεις

Έχοντας διερευνήσει την εφαρμογή της LSA στην ανάκτηση εγγράφων οι ακόλουθοι περιορισμοί πρέπει ακόμη να επιδιωχθούν σε μελλοντικές εργασίες.

a) **Επίλυση πολυσημίας**

Παρά το γεγονός ότι τα προβλήματα της πολυσημίας και συνωνυμίας στην τεχνική αντιστοίχισης λέξεων-κλειδιών ήταν αυτά που λειτούργησαν ως κίνητρο για

την ανάπτυξη της τεχνικής LSA, το πρόβλημα της πολυσημίας δεν έχει ακόμη επιλυθεί. Θα πρέπει να γίνουν προσπάθειες στο μέλλον αντιστοίχισης των όρων που εκφράζουν διαφορετικές έννοιες ως προς το περιεχόμενο τους ως διακριτά σημεία στο χώρο της λανθάνουσας σημασιολογικής ανάλυσης βάση της σημασίας τους.

b) **Καθορισμός του αριθμού των παραγόντων**

Μια πιθανή λύση εξακολουθεί να είναι απαραίτητη για την αντιμετώπιση του προβλήματος της επιλογής των k διαστάσεων προς διατήρηση στο χώρο μειωμένης διάστασης. Επί του παρόντος, αυτό γίνεται εμπειρικά ανάλογα σχετικά με τις μεθόδους που χρησιμοποιούνται για την αξιολόγηση των αποτελεσμάτων της ανάκτησης.

c) **Διάταξη λέξεων**

Μια λύση για την αντιμετώπιση της σειρά των λέξεων στην LSA θα πρέπει να διερευνηθεί περαιτέρω. Αυτό θα βοηθήσει την LSA στο να κερδίσει σημαντικές γνωστικές ικανότητες που τα ανθρώπινα όντα διαθέτουν και χρησιμοποιούν για να κατασκευάζουν και να εφαρμόζουν γνώσεις από προηγούμενη εμπειρία, και ιδίως την ικανότητα να χρησιμοποιούν λεπτομερείς και περίπλοκες πληροφορίες διάταξης, όπως αυτή που εκφράζεται από τη σύνταξη και χρησιμοποιείται στη λογική.

d) **Probabilistic Latent Semantic Analysis (pLSA)**

Η pLSA χρησιμοποιεί ένα παραγωγικό πρότυπο λανθάνουσας κατηγορίας για την εκτέλεση ανάμειξης πιθανοκρατικής αποσύνθεσης. Έχει διατυπωθεί ότι αυτή η προσέγγιση [21] οδηγεί σε έναν σωστότερο/ηθικότερο κανόνα πρόσβασης με γερά θεμέλια στην στατιστική συμπερασματολογίας. Από την άλλη πλευρά [22], τα μεθοδολογικά θεμέλια της LSA παραμένουν σε μεγάλο βαθμό μη ικανοποιητικά και ελλιπή. Η έρευνα θα πρέπει να επικεντρωθεί περισσότερο στη σύγκριση των δύο προσεγγίσεων, LSA και pLSA για την εύρεση εργαλείων καλύτερης επίδοσης για την ανάκτηση πληροφοριών.

e) **Συνδυασμός τεχνικών**

Μια καλύτερη προσέγγιση ως προς την ανακάλυψη σημασιολογικών εννοιών ίσως θα ήταν ο συνδυασμός της pLSA μαζί με την επεξεργασία σημάτων για την ανάκτηση εγγράφων/τραγουδιών βάση περιεχομένου. Στην συγκεκριμένη περίπτωση αναμένεται ο συνδυασμός της pLSA & τεχνικών μηχανικής μάθησης π.χ. deep learning να επιφέρει μεγαλύτερη ακρίβεια επίδοσης στην ανάκτηση εγγράφων αλλά και στην ανακάλυψη περισσότερων θεματικών περιοχών.

6.3 Εργαλεία που Χρησιμοποιήθηκαν

Τα παρακάτω περιβάλλοντα και εργαλεία χρησιμοποιήθηκαν για την υλοποιήσει της συγκεκριμένης εργασίας καθώς και των πειραμάτων που διεξήχθησαν στο κεφάλαιο 5.

a) MATLAB

   1) LSA

   2) k-means

   3) MDS & dimensionality reduction

b) NodeXL & Gephi

   Ανάλυση κοινωνικών δικτύων και γράφων.

Βιβλιογραφία

[Fruchterman and Reingold] T.M.J. Fruchterman and E.M. Reingold. Fruchterman-reingold algorithm. http://mathcs.pugetsound.edu/~jross/courses/archive/s13/cs261/lab/k/fruchterman91graph.pdf, visited 2013-10-20.

[Jacomy et al.] M. Jacomy, S. Heymann, T. Venturini, and M. Bastian. Force-directed graph drawing algorithm. http://webatlas.fr/tempshare/ForceAtlas2_Paper.pdf, visited 2013-10-12.

[1] C. D. Manning, P. Raghavan, and H. Schutze. Introduction to information retrieval. *Cambridge University Press, New York*, 2:134–160, 428–444, April 2009.

[2] P. Gajer and S. G. Kobourov. Graph drawing with intelligent placement. *Graph Algorithms and Applications*, 3:203–224, March 2002.

[3] S. G. Kobourov. Spring embedders and force directed graph drawing algorithms. *arXiv:1201.3011v1 cs.CG*, pages 1–23, January 2012.

[4] B. Rosario. Latent semantic indexing: An overview. *INFOSYS 240 Spring 2000 Final Paper*, December 2000.

[nod] Nodexl: Network overview, discovery and exploration for excel - home. http://nodexl.codeplex.com/, visited 2013-10-23.

[gep] Gephi, an open source graph visualization and manipulation software. https://gephi.org/, visited 2013-10-23.

[5] I. Borg and P. J. F. Groenen. Modern multidimensional scaling: Theory and applications. *Springer Series in Statistics*, 2:37–41, 2005.

[6] T. K. Landauer, P.W. Foltz, and D. Laham. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284, Januray 1998.

[7] K. Yang. Latent semantic indexing and linear relevance feedback in text information retrieval theory. *Cambridge University Press, New York*, May 1999.

[8] J. R. Bellegarda. Latent semantic mapping - principles and applications. *Principles and Applications*, March 2007.

[9] P.W. Foltz. Using semantic indexing for information filtering. *Proceedings of the Conference on Office Information Systems, Cambridge*, MA:40–47, August 2007.

[10] G. Ntais. Development of a stemmer for the greek language. *Department of Computer and Systems Sciences at Royal Institute of Technology Stockholm University*, pages 1–46, 2006. http://people.dsv.su.se/~hercules/papers/ntais_greek_stemmer_thesis_final.pdf.

[11] P. Stoica and R. Moses. Spectral analysis of signals. *Library of Congress Cataloging-in-Publication Data Spectral Analysis of Signals*, Januray 2005.

[12] R. Balakrishnan and K. Ranganathan. A textbook of graph theory. *Springer Universitext*, 2:13–18, October 2012.

[13] J. Lamping, R. Rao, and P. Pirolli. A focus plus context technique based on hyperbolic geometry for visualizing large hierarchies. *In Proceedings of Computer Human Interaction*, pages 401–408, 1995.

[14] T. Munzner, L. Lavagno, and W. Reisig. H3: Laying out large directed graphs in 3d hyperbolic space. *Proceedings of IEEE Symposium on Information Visualization*, pages 2–10, 1997.

[15] S. G. Kobourov and K. Wampler. Non-euclidean spring embedders. *IEEE Transactions on Visualization and Computer Graphics*, 6:757–767, November 2005.

[16] P. Eades. A heuristic for graph drawing. *Congressus Numerantium*, 42:149–160, 1984.

[17] A. Noack. Energy models for graph clustering. *Graph Algorithms Applications*, 11:453–480, 2007.

[18] A. Noack. Modularity clustering is force-directed layout. *Physical Review E*, 79, 2009.

[drt] drtoolbox: Matlab dimensionality reduction toolbox. http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html, visited 2013-10-24.

[for] Force-directed graph drawing algorithm. . http://en.wikipedia.org/wiki/Force-directed_graph_drawing.

[for] ForceAtlas2 new home-brew layout | gephi. . https://gephi.org/2011/forceatlas2-the-new-version-of-our-home-brew-layout/.

[fru] Fruchterman reingold wiki. http://wiki.gephi.org/index.php/fruchterman-reingold. http://wiki.gephi.org/index.php/Fruchterman-Reingold, visited 2013-10-29.

[chi] Chinese whispers: Clustering algorithm. http://dl.acm.org/citation.cfm?id=1654774, visited 2013-10-11.

[19] S. Deerwester, S.T. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *American Society for Information Science*, pages 391–407, March 1990.

[20] A. Kontostathis and W. M. Pottenger. A framework for understanding latent semantic indexing (lsi) performance. *Elsevier Science*, June 2004.

[21] T. Hofmann. Probabilistic latent semantic analysis. *In Proceedings of 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, February 1999.

[22] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, September 2001.