ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Αξιολόγηση Επισημειώσεων με τη Μέθοδο της Συσταδοποίησης Κειμένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΧΑΡΑΛΑΜΠΟΥ ΠΑΠΑΚΩΝΣΤΑΝΤΙΝΟΥ

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

Αθήνα, Νοέμβριος 2017

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

Αξιολόγηση Επισημειώσεων με τη Μέθοδο της Συσταδοποίησης Κειμένων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΧΑΡΑΛΑΜΠΟΥ ΠΑΠΑΚΩΝΣΤΑΝΤΙΝΟΥ

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 29η Νοεμβρίου 2017.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Γιώργος Στάμου	Ανδρέας-Γεώργιος Σταφυλοπάτης	Κωνσταντίνα Νικήτα
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγήτρια Ε.Μ.Π.

Αθήνα, Νοέμβριος 2017

(Υπογραφή)

.....

Χαράλαμπος Παπακωνσταντίνου

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

Στην οικογένειά μου,
Ηλία, Δαρεία, Ράνια,
που με στηρίζουν όλα αυτά τα χρόνια.

Ευχαριστίες

Νιώθω την ανάγκη να εκφράσω την απεριόριστη ευγνωμοσύνη μου στον επιβλέποντα καθηγητή μου Γιώργο Στάμου για την εμπιστοσύνη που μου έδειξε, καθώς και στο Γιώργο Σιόλα, με τον οποίο είχα στενή συνεργασία και ο οποίος μου παρείχε αποτελεσματική καθοδήγηση. Και οι δύο με στήριξαν ανελλιπώς κατά τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας. Χωρίς τη βοήθειά τους, η ολοκλήρωσή της θα ήταν αδύνατη.

Επίσης, θα ήθελα να ευχαριστήσω τον καθηγητή Ανδρέα – Γεώργιο Σταφυλοπάτη και την καθηγήτρια Κωνσταντίνα Νικήτα, που δέχθηκαν να είναι μέλη της τριμελούς επιτροπής για την εξέταση αυτής της εργασίας.

Τέλος, ένα μεγάλο Ευχαριστώ σε όλους τους δασκάλους μου που από τα παιδικά μου χρόνια με έχουν βοηθήσει να φτάσω εδώ που είμαι σήμερα, ανάμεσά τους και η κ. Βίκυ. Σε όλους τους φίλους μου, ανάμεσά τους ο Γιάννης, ο Βασίλης, ο Νίκος, ο Στέφανος, η Πηνελόπη, η Υβόνη, για την ανυπολόγιστη βοήθειά τους σε επιστημονικό αλλά και συναισθηματικό επίπεδο. Στην οικογένειά μου, Ηλία, Δαρεία, Ράνια, που με βοήθησαν να γίνω ο άνθρωπος που είμαι σήμερα.

Περίληψη

Σε αυτήν την εργασία προτείνεται ένα σύστημα αξιολόγησης των επισημειώσεων ενός συνόλου εγγράφων. Αρχικά γίνεται μία συσταδοποίηση αναφοράς του συνόλου των εγγράφων σε επιμέρους συστάδες με βάση το περιεχόμενο των κειμένων, τα οποία αναπαρίστανται στο μοντέλο διανυσματικού χώρου με τεχνική αναπαράστασης *tf-idf*. Στη συνέχεια γίνεται νέα συσταδοποίηση των εγγράφων με βάση τις ετικέτες με τις οποίες αυτά έχουν επισημειωθεί. Ακολουθεί σύγκριση της νέας συσταδοποίησης τη συσταδοποίηση αναφοράς με χρήση το δείκτη Rand Index. Τέλος, οι επισημειώσεις των εγγράφων εμπλουτίζονται μέσω της αντιστοίχισης των ελεύθερων λέξεων – κλειδιών που έχουν ορίσει οι συγγραφείς των εγγράφων σε έννοιες της ιεραρχικής οντολογίας από την οποία προέρχονται και οι αρχικές επισημειώσεις. Γίνεται εκ νέου συσταδοποίηση με βάση τις εμπλουτισμένες ετικέτες και η τελική συσταδοποίηση συγκρίνεται και πάλι με τη συσταδοποίηση αναφοράς.

Λέξεις Κλειδιά: συσταδοποίηση κειμένων, λέξη κλειδί, επισημείωση, ιεραρχική οντολογία, αξιολόγηση επισημειώσεων

Abstract

In this thesis, a system for the evaluation of the tags of a document dataset is recommended. Initially, a reference clustering of the document dataset into clusters is performed, based on the content of the documents, which are represented in the vector space model with the *tf-idf* technique. Subsequently, a new clustering of the document set is performed, based on the labels the documents have been tagged with. Next, the new clustering is compared to the initial reference clustering with the Rand Index similarity measure. Finally, the labels of the documents are enriched by matching the authors' free keywords to concepts of the hierarchical ontology the initial document labels originate from. A final clustering of the document set is performed, based on the enriched labels of the documents, and it is compared again with the initial reference clustering.

Keywords: document clustering, keyword, tag, hierarchical ontology, tag evaluation

Πίνακας περιεχομένων

1 Εισαγωγή 21

1.1 Αναζήτηση πληροφοριών στον Παγκόσμιο Ιστό 21

1.1.1 Συνεισφορά 22

1.2 Οργάνωση κειμένου 22

2 Σχετικές εργασίες 23

2.1 Συσταδοποίηση εγγράφων 23

2.2 Σύγκριση συσταδοποιήσεων 24

2.3 Εξαγωγή λέξεων – κλειδιών 25

3 Θεωρητικό υπόβαθρο 26

3.1 Οντολογίες 26

3.1.1 Περιγραφές ατόμων 26

3.1.2 Περιγραφές κατηγοριών 27

3.1.3 Περιγραφές σχέσεων 28

3.2 Γλώσσες αναπαράστασης γνώσης 29

3.2.1 RDF 29

3.2.2 OWL 29

3.2.3 SKOS 30

3.3 Συσταδοποίηση 30

3.4 Μέθοδοι συσταδοποίησης 31

3.4.1 Συσταδοποίηση k-μέσων 31

3.4.2 Ιεραρχική Συσταδοποίηση 33

3.4.3 Συσταδοποίηση πυκνότητας 34

3.5 Συσταδοποίηση κειμένων 37

3.5.1 Αναπαράσταση κειμένων στο μοντέλο διανυσματικού χώρου 37

3.5.1.1 Μοντέλο bag-of-words 38

3.5.1.2 Μοντέλο βαρών tf-idf 39

3.5.2 Απόσταση κειμένων στο μοντέλο διανυσματικού χώρου 40

4 Dataset 41

4.1 Σύστημα Ταξινόμησης ACM 41
4.2 Ψηφιακή Βιβλιοθήκη ACM 41
4.3 Δημιουργία του dataset 42
4.3.1 Αναζήτηση και λήψη εγγράφων 42
4.3.2 Διαλογή εγγράφων 43
4.3.3 Μετατροπή σε καθαρό κείμενο 43
5 Αξιολόγηση Επισημειώσεων με τη Μέθοδο της Συσταδοποίησης 44
5.1 Δημιουργία αρχικής συσταδοποίησης με βάση το καθαρό κείμενο 44
5.1.1 Hierarchical Clustering 45
5.1.2 DBScan 46
5.2 Συσταδοποίηση με βάση τις ελεύθερες λέξεις – κλειδιά 46
5.3 Συσταδοποίηση με βάση τις ετικέτες από το Σύστημα Ταξινόμησης ACM 46
5.3.1 Απόσταση ετικετών 47
5.3.1.1 Απλή απόσταση ετικετών (Naive Tag Distance) 47
5.3.1.2 Απόσταση ετικετών με προσαρμοσμένα βάρη (Adjusted Tag Distance) . 48
5.3.2 Απόσταση εγγράφων 49
5.3.2.1 Ελάχιστη απόσταση ετικετών 49
5.3.2.2 Μέγιστη απόσταση ετικετών 50
5.3.2.3 Μέση απόσταση ετικετών 50
5.3.3 Συσταδοποίηση 50
5.4 Εμπλουτισμός επισημειώσεων και νέα συσταδοποίηση 51
5.4.1 Αντιστοίχιση ελεύθερων λέξεων – κλειδιών σε ετικέτες του Συστήματος Ταξινόμησης ACM 51
5.4.2 Συσταδοποίηση 52
6 Αποτελέσματα 53
6.1 Μέτρο σύγκρισης συσταδοποιήσεων 53
6.1.1 Rand Index 53
6.2 Αξιολόγηση ελεύθερων λέξεων – κλειδιών 54
6.3 Αξιολόγηση ετικετών μέσω ιεραρχικής συσταδοποίησης 54
6.3.1 Απλή απόσταση ετικετών 55
6.3.2 Μέγιστη απόσταση ετικετών 55

6.3.3 Μέση απόσταση ετικετών *55*
6.3.4 Απλή απόσταση ετικετών με προσαρμοσμένα βάρη *55*
6.3.5 Μέγιστη απόσταση ετικετών με προσαρμοσμένα βάρη *56*
6.3.6 Μέση απόσταση ετικετών με προσαρμοσμένα βάρη *56*
6.3.7 Σύνοψη *56*
6.4 Αξιολόγηση ετικετών μέσω συσταδοποίησης πυκνότητας 59
7 Τεχνικές λεπτομέρειες 60
7.1 SQLite 60
7.2 Tika 61
7.3 LingPipe 61
7.4 psjava 62
7.5 Apache Commons: Math 62
7.6 Maven 62
7.7 XML 62
8 Επίλογος 64
8.1 Σύνοψη και συμπεράσματα 64
8.2 Μελλοντικές επεκτάσεις 64
9 Βιβλιογραφία 66

Κατάλογος σχημάτων

Σχήμα 3.1: Παράδειγμα εκτέλεσης αλγορίθμου k-means 32
Σχήμα 3.2: Ορισμοί απόστασης συστάδων στην Ιεραρχική Συσταδοποίηση 34
Σχήμα 3.3: Παράδειγμα Ιεραρχικής Συσταδοποίησης 34
Σχήμα 3.4: Παράδειγμα διαχωρισμού σημείων κατά την εκτέλεση αλγορίθμου DBSCAN ... 35
Σχήμα 3.5: Σύγκριση εφαρμογής αλγορίθμων k-means και DBSCAN σε κοινά σύνολα δεδομένων 37
Σχήμα 5.1: Γράφος απλής απόστασης ετικετών 48
Σχήμα 5.2: Γράφος απόστασης ετικετών με προσαρμοσμένα βάρη 49
Σχήμα 6.1: Σύγκριση συσταδοποιήσεων με απλή απόσταση ετικετών 56
Σχήμα 6.2: Σύγκριση συσταδοποιήσεων με μέγιστη απόσταση ετικετών 57
Σχήμα 6.3: Σύγκριση συσταδοποιήσεων με μέση απόσταση ετικετών 57
Σχήμα 6.4: Σύγκριση συσταδοποιήσεων με απόσταση ετικετών με προσαρμοσμένα βάρη ... 58
Σχήμα 6.5: Σύγκριση συσταδοποιήσεων με μέγιστη απόσταση ετικετών με προσαρμοσμένα βάρη 58
Σχήμα 6.6: Σύγκριση συσταδοποιήσεων με μέση απόσταση ετικετών με προσαρμοσμένα βάρη 59
Σχήμα 7.1: Δομή αποθήκευσης αποτελεσμάτων συσταδοποίησης 62

Κατάλογος πινάκων

Πίνακας 3.1: Παράδειγμα αναπαράστασης κειμένου στο μοντέλο bag-of-words 39
Πίνακας 5.1: Αλγόριθμοι συσταδοποίησης με βάση τις ετικέτες του Συστήματος Ταξινόμησης της ACM 51

1 Εισαγωγή

1.1 Αναζήτηση πληροφοριών στον Παγκόσμιο Ιστό

Ζούμε σε μία εποχή όπου οι διαθέσιμες πληροφορίες στον παγκόσμιο ιστό αυξάνονται με εκθετικούς ρυθμούς. Εκατομμύρια νέες σελίδες δημιουργούνται σε καθημερινή βάση, ενώ οι ήδη υπάρχουσες τροποποιούνται, ενημερώνονται, εμπλουτίζονται. Αν και η μεγάλη ευκολία δημιουργίας περιεχομένου στο Διαδίκτυο λειτουργεί θετικά στην καταχώρηση και τη διάδοση της γνώσης, αποτέλεσμα είναι επίσης ο τεράστιος όγκος δεδομένων, των οποίων η διαχείριση αποτελεί ένα σημαντικό ζήτημα προς επίλυση.

Οι πληροφορίες που είναι αποθηκευμένες στο Διαδίκτυο δεν μπορούν να φανούν χρήσιμες παρά μόνο εάν ο ενδιαφερόμενος χρήστης μπορεί να τις εντοπίσει σε σύντομο χρονικό διάστημα, διατυπώνοντας ένα κατάλληλο ερώτημα.

Οι κλασσικές μέθοδοι αναζήτησης περιλαμβάνουν στατιστική ανάλυση των κειμένων ως προς το λεξιλόγιό τους και τους όρους που εμφανίζονται μέσα σε αυτά και στη συνέχεια σύγκριση των όρων ενός ερωτήματος με αυτούς που εμφανίζονται στα υποψήφια ως απαντήσεις κείμενα. Οι συγκεκριμένες τεχνικές, αν και έχουν ήδη εξελιχθεί σημαντικά, παρουσιάζουν ορισμένα προβλήματα, όπως για παράδειγμα η περίπτωση όπου δύο διαφορετικές έννοιες, πιθανώς από δύο εντελώς διαφορετικούς τομείς ενδιαφέροντος, μοιράζονται την ίδια ονομασία σε μία φυσική γλώσσα.

Τα προβλήματα αυτά μπορούν να αντιμετωπιστούν μέσω της επισημείωσης των εγγράφων με έννοιες από μία ιεραρχική οντολογία αναπαράστασης της γνώσης [1]. Σε αυτήν την περίπτωση, είναι δυνατόν να αναγνωριστούν οι έννοιες της οντολογίας στις οποίες γίνεται αναφορά μέσα σε ένα ερώτημα, και στη συνέχεια να ερευνηθούν ως υποψήφια αποτελέσματα του ερωτήματος όσα έγγραφα έχουν επισημειωθεί με τις ίδιες έννοιες [2].

Σε αυτό το σημείο, γίνεται φανερή η ανάγκη αξιολόγησης των επισημειώσεων ενός συνόλου εγγράφων. Πράγματι, η σωστή ή λεπτομερέστερη επισημείωση των εγγράφων θα επιτρέπει και πιο αποτελεσματική απάντηση σχετικών ερωτημάτων. Στόχος, λοιπόν, αυτής

της εργασίας, είναι η ανάπτυξη ενός συστήματος αξιολόγησης των επισημειώσεων ενός συνόλου εγγράφων με έννοιες από μία ιεραρχική οντολογία αναπαράστασης γνώσης.

1.1.1 Συνεισφορά

Στα πλαίσια ανάπτυξης ενός εργαλείου αξιολόγησης των επισημειώσεων ενός συνόλου εγγράφων, έγιναν επίσης οι παρακάτω ενέργειες:

1. Μελετήθηκαν και αξιολογήθηκαν στην πράξη μέθοδοι συσταδοποίησης κειμένων.
2. Διατυπώθηκαν εναλλακτικοί ορισμοί της απόστασης δύο εννοιών μίας ιεραρχικής οντολογίας.
3. Διατυπώθηκαν εναλλακτικοί ορισμοί της απόστασης δύο εγγράφων σε συνάρτηση με την απόσταση των ετικετών – εννοιών με τις οποίες αυτά έχουν επισημειωθεί.
4. Διερευνήθηκαν μέθοδοι σύγκρισης εναλλακτικών συσταδοποιήσεων ενός συνόλου δεδομένων
5. Αναπτύχθηκε ένα σύστημα αξιολόγησης επισημειώσεων εγγράφων

1.2 Οργάνωση κειμένου

Στο Κεφάλαιο 2 παρουσιάζονται εργασίες σχετικές με το αντικείμενο αυτής της διπλωματικής. Στο Κεφάλαιο 3 γίνεται θεωρητική ανάλυση των αντικειμένων που πραγματεύεται η εργασία. Στο Κεφάλαιο 4 γίνεται παρουσίαση του συνόλου δεδομένων (dataset) που κατασκευάστηκε και χρησιμοποιήθηκε κατά τη διάρκεια των πειραμάτων. Στο Κεφάλαιο 5 αναλύονται οι πειραματικές διαδικασίες που πραγματοποιήθηκαν. Στο Κεφάλαιο 6 παρουσιάζονται τα αποτελέσματα των πειραμάτων. Στο Κεφάλαιο 7 γίνεται αναφορά σε αξιόλογες τεχνικές λεπτομέρειες που αφορούν την υλοποίηση των πειραμάτων. Στο Κεφάλαιο 8 συνοψίζονται τα συμπεράσματα της διπλωματικής εργασίας και γίνονται προτάσεις για μελλοντικές επεκτάσεις. Στο Κεφάλαιο 9 παρουσιάζεται η βιβλιογραφία που χρησιμοποιήθηκε για τη συγγραφή του παρόντος κειμένου.

2 Σχετικές εργασίες

Για την εκπόνηση των πειραμάτων και τη συγγραφή της παρούσας εργασίας, διερευνήθηκαν ορισμένες συγγενικές επιστημονικές περιοχές.

Για τη συσταδοποίηση των εγγράφων του συνόλου δεδομένων που κατασκευάστηκε, διερευνήθηκαν τεχνικές συσταδοποίησης εγγράφων.

Επίσης, έγινε μελέτη εναλλακτικών τρόπων ή μέτρων σύγκρισης συσταδοποιήσεων.

Τέλος, αναζητήθηκαν μέθοδοι αυτόματης εξαγωγής λέξεων – κλειδιών από ένα κείμενο με σκοπό τον εμπλουτισμό των επισημειώσεων των εγγράφων του συνόλου δεδομένων.

2.1 Συσταδοποίηση εγγράφων

Για τη συσταδοποίηση εγγράφων, προτείνεται ένα ολοκληρωμένο σύστημα από τους Andreas Hotho, Steffen Staab και Gerd Stumme [3]. Στη συγκεκριμένη μελέτη, για τη συσταδοποίηση αξιοποιείται επίσης η γνώση που είναι αποθηκευμένη σε μία οντολογία. Η διαδικασία συσταδοποίησης κειμένων που προτείνουν συνοψίζεται στα εξής βήματα:

1. Προεπεξεργασία των κειμένων
    a. Αφαίρεση κοινών λέξεων διάσπασης (stopwords), δηλαδή λέξεων που είναι συνηθισμένες σε μία πρόταση και δεν προσθέτουν πληροφορία, όπως για παράδειγμα οι λέξεις “the”, “a”, “in”, “to”, κ.τ.λ.
    b. Stemming, δηλαδή εύρεση της ρίζας των λέξεων που αποτελούν το κάθε έγγραφο
    c. Pruning, δηλαδή αφαίρεση των λέξεων με αρκετά χαμηλή συχνότητα εμφάνισης
    d. Υπολογισμός βαρών *tf-idf*
2. Ενσωμάτωση της γνώσης από την οντολογία στα κείμενα
    a. Αντικατάσταση συνωνύμων όρων από τους αρχικούς
    b. Τεχνικές αποσαφήνισης νοήματος όρων που αντιστοιχούν σε πολλαπλές έννοιες

3. Συσταδοποίηση
    a. Χρήση απόστασης συνημιτόνου
    b. Συσταδοποίηση με χρήση του αλγορίθμου Bi-Section-K-Means [4]

Η παραπάνω τεχνική συσταδοποίησης παρουσιάζει αρκετά καλά αποτελέσματα. Ωστόσο, η υλοποίησή της θεωρήθηκε ότι δεν αποτελεί μέρος του αντικειμένου της παρούσας εργασίας και επομένως δεν υιοθετήθηκε.

2.2 Σύγκριση συσταδοποιήσεων

Ένα συνηθισμένο μέτρο σύγκρισης δύο συσταδοποιήσεων είναι ο δείκτης Rand Index, οποίος και χρησιμοποιήθηκε στην παρούσα εργασία και περιγράφεται αναλυτικά στην παράγραφο 6.1.1.

Ένα εναλλακτικό μέτρο σύγκρισης δύο συνόλων είναι ο δείκτης Jaccard Index, όπως περιγράφεται στο *Introduction to Data Mining* [5].

Ο δείκτης Jaccard Index ορίζεται ως εξής: έστω δύο σύνολα A και B. Τότε, ο δείκτης ομοιότητας Jaccard των δύο συνόλων είναι:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Η τιμή του δείκτη Jaccard Index μπορεί να πάρει τιμές από 0 μέχρι 1. Μεγαλύτερη τιμή του δείκτη συνεπάγεται μεγαλύτερη ομοιότητα των συνόλων A και B.

Το πλεονέκτημα του συγκεκριμένου δείκτη ομοιότητας είναι ότι για την σύγκριση των δύο συνόλων λαμβάνεται υπόψη το ποσοστό των κοινών τους στοιχείων σε σχέση με τα συνολικά τους στοιχεία και όχι ο απόλυτος αριθμός των κοινών στοιχείων των δύο συνόλων. Έτσι, είναι δυνατή η σύγκριση της ομοιότητας ζευγαριών συνόλων με μεγάλες αποκλίσεις μεγέθους (πλήθους στοιχείων).

Ο δείκτης Jaccard Index αποτελεί δείκτη ομοιότητας δύο συνόλων, μπορεί όμως να χρησιμοποιηθεί και για τη σύγκριση δύο συσταδοποιήσεων X με n συστάδες και Y με m συστάδες ως εξής:

1. Κάθε συστάδα X_i της X συγκρίνεται με κάθε συστάδα Y_i της Y με μέτρο σύγκρισης το δείκτη Jaccard Index.
2. Κατασκευάζεται ο πίνακας σύγκρισης C όπου $C_{i,j} = J(X_i, Y_j)$.
3. Λαμβάνεται ως δείκτης ομοιότητας των δύο συσταδοποιήσεων ο μέσος όρος όλων των $C_{i,j}$:

$$r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} C_{i,j}}{n \cdot m}$$

2.3 Εξαγωγή λέξεων – κλειδιών

Οι Rose, Engel και Cramer [6] ανέπτυξαν τον αλγόριθμο Rapid Automatic Keyword Extraction (RAKE) ο οποίος στοχεύει στην αυτόματη εξαγωγή λέξεων – κλειδιών από ένα συγκεκριμένο έγγραφο. Τα βήματα του αλγορίθμου συνοψίζονται ως εξής:

1. Εύρεση υποψήφιων λέξεων – κλειδιών μέσω του διαχωρισμού του κειμένου του εγγράφου με βάση κάποιες προκαθορισμένες λέξεις διάσπασης (stopwords).
2. Βαθμολόγηση κάθε υποψήφιας λέξης – κλειδιού σύμφωνα με κάποιο από τα προτεινόμενα κριτήρια, όπως για παράδειγμα η συχνότητα εμφάνισης της λέξης εντός του εγγράφου.
3. Ταξινόμηση των υποψήφιων λέξεων – κλειδιών ως προς τη βαθμολογία του προηγούμενου βήματος σε φθίνουσα σειρά.
4. Επιλογή των N υποψήφιων λέξεων – κλειδιών με την καλύτερη βαθμολογία και χαρακτηρισμός τους ως λέξεις – κλειδιά.

Ο αλγόριθμος αυτός θα μπορούσε να χρησιμοποιηθεί στην παρούσα εργασία για τον εμπλουτισμό των ετικετών που έγινε στην παράγραφο 5.4. Ωστόσο, επιλέχθηκε η χρήση των υπαρχόντων ελεύθερων λέξεων – κλειδιών που προσδιόρισαν οι συγγραφείς των εγγράφων μέσω της αντιστοίχισής τους σε έννοιες του Συστήματος Ταξινόμησης της ACM (παράγραφος 4.1).

3 Θεωρητικό υπόβαθρο

Για την αναπαράσταση των επιμέρους πεδίων της Επιστήμης της Πληροφορικής στο πείραμα χρησιμοποιήθηκε μία Οντολογία.

Επίσης, χρησιμοποιήθηκαν τεχνικές Συσταδοποίησης κειμένων.

3.1 Οντολογίες

Στην Επιστήμη της Πληροφορικής, με τον όρο Οντολογία ονομάζεται ένα σύνολο τυποποιημένων προσδιορισμών τύπων, ιδιοτήτων και σχέσεων μεταξύ των οντοτήτων που υπάρχουν σε ένα τομέα ενδιαφέροντος [7].

Μία οντολογία αποτελεί μία μοντελοποίηση της γνώσης ενός συγκεκριμένου πεδίου ενδιαφέροντος. Επομένως, δεν έχει νόημα μία οντολογία να χαρακτηρίζεται ως «σωστή» ή «λανθασμένη», καθώς η λογική μοντελοποίησης είναι υποκειμενική και εξαρτάται από την εκάστοτε εφαρμογή για την οποία δημιουργήθηκε η οντολογία. Είναι δυνατόν για το ίδιο πεδίο ενδιαφέροντος να κατασκευαστούν διαφορετικές οντολογίες, που βασίζονται σε διαφορετικές μοντελοποιήσεις της γνώσης.

Οι οντολογίες παίζουν καθοριστικό ρόλο στην αναζήτηση πληροφοριών στο Σημασιολογικό Ιστό, καθώς σήμερα αποτελούν μία από τις κυριότερες μεθόδους κωδικοποίησης, αποθήκευσης, ανεύρεσης και παραγωγής γνώσης.

Σε μία συνηθισμένη οντολογία σχηματίζονται περιγραφές για τα παρακάτω:

- **Άτομα (Individuals)**, δηλαδή τα στοιχεία του κόσμου προς περιγραφή
- **Κατηγορίες**, δηλαδή σύνολα ατόμων που μοιράζονται μία κοινή ιδιότητα
- **Σχέσεις**, δηλαδή συσχετισμούς μεταξύ δύο ή περισσότερων ατόμων

3.1.1 Περιγραφές ατόμων

Το πρώτο βήμα για την κατασκευή μιας οντολογίας είναι η αναπαράσταση των ατόμων, δηλαδή των στοιχείων του κόσμου του οποίου πρόκειται να γίνει η περιγραφή. Καθένα από

αυτά τα στοιχεία αντιστοιχίζεται σε μία συμβολοσειρά, ώστε να μπορεί να ταυτοποιηθεί από αυτή τη συμβολοσειρά με τρόπο μονοσήμαντο [8].

Είναι προφανές ότι κάθε άτομο της οντολογίας, δηλαδή κάθε συμβολοσειρά – ονομασία στοιχείου, θα πρέπει να αντιστοιχεί σε ένα και μόνο ένα στοιχείο του πραγματικού κόσμου. Δεν είναι ανάγκη, όμως, να ισχύει και το αντίστροφο. Αντιθέτως, είναι δυνατόν ένα στοιχείο του πραγματικού κόσμου να εμφανίζεται εντός της οντολογίας με πολλά διαφορετικά ονόματα. Με άλλα λόγια, είναι συνηθισμένο στις οντολογίες να **μην** ισχύει η *υπόθεση μοναδικού ονόματος* (unique name assumption – UNA).

Δεδομένων των παραπάνω, είναι χρήσιμη η ύπαρξη της δυνατότητας δήλωσης της ισότητας ή μη δύο ατόμων μιας οντολογίας, δηλαδή η δήλωση εάν δύο άτομα αναφέρονται στο ίδιο στοιχείο του πραγματικού κόσμου ή όχι.

3.1.2 Περιγραφές κατηγοριών

Τα άτομα μιας οντολογίας που μοιράζονται μία κοινή ιδιότητα σχηματίζουν μία *κατηγορία* ή *κλάση* στην οντολογία αυτή. Όπως και τα άτομα, έτσι και οι κατηγορίες ονοματίζονται εντός της οντολογίας με τη χρήση μίας συμβολοσειράς [8].

Οι κατηγορίες μίας οντολογίας θυμίζουν τη μαθηματική έννοια του συνόλου. Δηλαδή, μία κατηγορία μίας οντολογίας αντιστοιχεί στο σύνολο των στοιχείων του πραγματικού κόσμου που μοιράζονται την αντίστοιχη ιδιότητα. Επομένως, είναι δυνατόν να προκύψουν συσχετισμοί μεταξύ κατηγοριών που αντιστοιχούν σε πράξεις μεταξύ μαθηματικών συνόλων. Παρακάτω περιγράφονται ορισμένοι ενδεικτικοί τέτοιοι συσχετισμοί:

- Έστω ότι σε μία οντολογία υπάρχει μία κλάση Β της οποίας τα άτομα είναι **υποσύνολο** του συνόλου των ατόμων μίας κλάσης Α. Με άλλα λόγια, όλα τα στοιχεία που ανήκουν στην κατηγορία Β ανήκουν ταυτόχρονα και στην κατηγορία Α. Τότε η κατηγορία Β λέγεται *υποκατηγορία* ή *υποέννοια* της κατηγορίας Α. Για παράδειγμα, σε μία υποθετική οντολογία που περιλαμβάνει τις κλάσεις *Άνδρας* και *Άνθρωπος*, όλα τα άτομα της κλάσης *Άνδρας* θα ανήκουν και στην κλάση *Άνθρωπος*. Επομένως, η κατηγορία *Άνδρας* είναι υποκατηγορία της κλάσης *Άνθρωπος*.
- Έστω ότι σε μία οντολογία υπάρχουν δύο κλάσεις Α, Β των οποίων τα άτομα αποτελούν δύο σύνολα **ξένα** μεταξύ τους. Δηλαδή, δεν υπάρχει άτομο της κλάσης Α που να ανήκει και στην κλάση Β και αντιστρόφως. Σε αυτήν την περίπτωση, οι κλάσεις Α και Β ονομάζονται *ξένες μεταξύ τους*. Για παράδειγμα, δεν γίνεται ένα άτομο της προηγούμενης υποθετικής οντολογίας να ανήκει ταυτόχρονα στις κλάσεις *Άνδρας* και *Γυναίκα*. Επομένως, οι κλάσεις *Άνδρας* και *Γυναίκα* είναι ξένες μεταξύ τους.

- Τέλος, έστω ότι σε μία οντολογία υπάρχουν δύο κλάσεις Α, Β οι οποίες περιλαμβάνουν ακριβώς τα ίδια άτομα, περιγράφουν δηλαδή την ίδια ιδιότητα των ατόμων με αποτέλεσμα τα αντίστοιχα σύνολα στοιχείων του πραγματικού κόσμου να **ταυτίζονται**. Σε αυτήν την περίπτωση θεωρείται ότι οι κλάσεις Α και Β *ταυτίζονται*.

Μέσα σε μία οντολογία μπορούν επίσης να χρησιμοποιηθούν *κατασκευαστές* κατηγοριών. Πρόκειται για τελεστές που αντιστοιχούν σε λογικές πράξεις και οδηγούν στη δημιουργία νέων κατηγοριών βάσει άλλων υπαρχόντων. Παρακάτω αναφέρονται κάποια ενδεικτικά παραδείγματα:

- **Διάζευξη** δύο κατηγοριών Α, Β ονομάζεται το σύνολο των ατόμων της οντολογίας που παρουσιάζουν τουλάχιστον μία από τις δύο ιδιότητες που περιγράφουν οι κατηγορίες Α, Β, ανήκουν δηλαδή σε τουλάχιστον μία από τις δύο κατηγορίες. Πρόκειται για την **ένωση** των αντίστοιχων συνόλων των στοιχείων του πραγματικού κόσμου.
- **Σύζευξη** δύο κατηγοριών Α, Β είναι το σύνολο των ατόμων που παρουσιάζει και τις δύο ιδιότητες που περιγράφουν οι κατηγορίες Α, Β αντίστοιχα, ανήκουν δηλαδή και στις δύο κατηγορίες. Πρόκειται για την **τομή** των αντίστοιχων συνόλων.
- **Άρνηση** μίας κατηγορίας Α είναι το σύνολο των ατόμων τις οντολογίας που **δεν** παρουσιάζουν την ιδιότητα που περιγράφει η κατηγορία Α, είναι δηλαδή όλα τα άτομα που δεν ανήκουν στην κατηγορία Α. Πρόκειται για το **συμπλήρωμα** του αντίστοιχου συνόλου.

3.1.3 Περιγραφές σχέσεων

Σε μία οντολογία, με τον όρο *Σχέση* ή *Ρόλος* αναφέρεται ο συσχετισμός μεταξύ δύο (συνήθως) ατόμων της οντολογίας.

Όπως και με τις κατηγορίες, είναι δυνατόν να παρατηρηθούν σχέσεις μεταξύ των ρόλων μίας οντολογίας [8]. Παρακάτω περιγράφονται οι συνηθέστερες περιπτώσεις.

Όταν ο συσχετισμός ενός ατόμου με ένα άλλο άτομο μέσω ενός ρόλου *ρόλος1* συνεπάγεται απαραιτήτως και το συσχετισμό του με το ίδιο άτομο μέσω ενός άλλου ρόλου *ρόλος2*, τότε ο *ρόλος1* ονομάζεται **υπορόλος** του ρόλου *ρόλος2*. Για παράδειγμα, έστω ότι ένα άτομο Α σχετίζεται με ένα άτομο Β με το ρόλο *έχειΠατέρα*. Αυτόματως, το άτομο Α σχετίζεται με το άτομο Β και με το ρόλο *έχειΓονιό*. Επομένως, ο ρόλος *έχειΠατέρα* είναι υπορόλος του ρόλου *έχειΓονιό*.

Όταν ο συσχετισμός ενός ατόμου Α με ένα άτομο Β μέσω ενός ρόλου *ρόλος1* συνεπάγεται το συσχετισμό του ατόμου Β με το άτομο Α μέσω ενός ρόλου *ρόλος2*, τότε οι

ρόλοι *ρόλος1*, *ρόλος2* ονομάζονται **ανάστροφοι**. Για παράδειγμα, έστω ότι ένα άτομο Α σχετίζεται με ένα άτομο Β με το ρόλο *έχειΓονιό*. Ως συνέπεια, το άτομο Β σχετίζεται με το άτομο Α με το ρόλο *έχειΠαιδί*. Επομένως, οι ρόλοι *έχειΓονιό* και *έχειΠαιδί* είναι ανάστροφοι.

3.2 Γλώσσες αναπαράστασης γνώσης

Γίνεται φανερό από τα προηγούμενα ότι είναι ανάγκη να αναπτυχθούν γλώσσες οι οποίες να παρέχουν τις δυνατότητες μοντελοποίησης γνώσης που έχει μία οντολογία, ενώ ταυτόχρονα να είναι εύκολα κατανοητές από τον άνθρωπο αλλά και επεξεργάσιμες από τον υπολογιστή.

Παρακάτω περιγράφονται οι γλώσσες RDF και OWL. Επίσης γίνεται μία σύντομη περιγραφή του SKOS.

3.2.1 RDF

Η RDF (Resource Description Framework) είναι η πρώτη γλώσσα που υιοθετήθηκε από τον οργανισμό W3C ως μοντέλο αναπαράστασης μεταδεδομένων [9].

Στην RDF, η βασική ιδέα είναι ότι τα άτομα που πρόκειται να περιγραφούν έχουν κάποιες ιδιότητες για τις οποίες παρουσιάζουν κάποιες συγκεκριμένες τιμές. Η αναπαράσταση γίνεται με τη μορφή τριάδων που αποτελούνται από ένα *υποκείμενο (subject)*, μία *ιδιότητα (property)* και ένα *αντικείμενο (object)*. Μία τέτοια τριάδα ονομάζεται πρόταση.

Για την αναπαράσταση της ιδιότητας ενός ατόμου, το άτομο παίζει το ρόλο του υποκειμένου, η ιδιότητα παίζει το ρόλο της ιδιότητας και η τιμή που έχει το άτομο για τη συγκεκριμένη ιδιότητα παίζει το ρόλο του αντικειμένου.

Τα υποκείμενα, οι ιδιότητες και τα αντικείμενα καταγράφονται με τη μορφή URIs. Η γλώσσα που χρησιμοποιείται για την καταγραφή προτάσεων είναι η XML. Η ακριβής σύνταξη περιγράφεται με το πρότυπο RDF/XML.

Η RDF είναι μία σχετικά απλή γλώσσα αναπαράστασης γνώσεις και λειτουργεί ως βάση για την ανάπτυξη άλλων, πιο πολύπλοκων, αυξημένης λειτουργικότητας, γλωσσών, όπως η OWL που εξετάζεται παρακάτω.

3.2.2 OWL

Τα αρχικά OWL προκύπτουν με αναγραμματισμό από τον όρο *Web Ontology Language*. Όπως υποδηλώνει και το όνομά της, η OWL είναι μία γλώσσα συγγραφής οντολογιών που μπορεί να χρησιμοποιηθεί στο διαδίκτυο [10].

Η OWL αποτελεί μία επέκταση της RDF, σε σχέση με την οποία παρουσιάζει βελτιώσεις τόσο ως προς τη λειτουργικότητα όσο και ως προς την ευχρηστία. Ειδικότερα, η OWL

παρέχει επιπλέον της RDF τη δυνατότητα ορισμού κλάσεων μέσω της διάζευξης, σύζευξης ή άρνησης άλλων κλάσεων. Επίσης, λόγω του συντακτικού της, γίνεται εύκολα κατανοητή από τον άνθρωπο και επομένως είναι εύκολη στη χρήση. Ταυτόχρονα, ως επέκταση της RDF, είναι συμβατή με αυτήν και φυσικά με το πρότυπο XML.

Η OWL ορίζει τρεις υπογλώσσες διαφορετικού επιπέδου εκφραστικότητας:

- **OWL Lite:** Πρόκειται για την απλούστερη έκδοση της OWL. Απευθύνεται σε εφαρμογές που δεν έχουν υψηλές απαιτήσεις ως προς την εκφραστικότητα, το οποίο επιτρέπει την ανάπτυξη αποδοτικότερων εργαλείων που λειτουργούν ταχύτερα σε σχέση με τα εργαλεία που αφορούν σε πιο πλήρεις εκδοχές της OWL.
- **OWL DL:** Σε αυτή την έκδοση της OWL παρέχεται η μέγιστη εκφραστικότητα, ενώ παράλληλα διατηρούνται σε ικανοποιητικό επίπεδο οι υπολογιστικές δυνατότητες των εργαλείων της γλώσσας.
- **OWL Full:** Εδώ παρέχεται το ίδιο εκφραστικό επίπεδο με την OWL DL, χωρίς όμως να υπάρχουν συντακτικοί περιορισμοί, ενώ ταυτόχρονα παρέχεται η δυνατότητα της μεταμοντελοποίησης. Αυτά τα χαρακτηριστικά κάνουν τη γλώσσα να είναι μη-αποφασίσιμη (undecidable).

3.2.3 SKOS

Τα αρχικά SKOS προκύπτουν από την ονομασία Simple Knowledge Organization System (SKOS) [11]. Πρόκειται για ένα μοντέλο δεδομένων που αφορά στο διαμοιρασμό και τη σύνδεση συστημάτων οργάνωσης γνώσης μέσω του διαδικτύου.

Το SKOS έχει δημιουργηθεί σύμφωνα με το πρότυπο RDF και μπορεί να χρησιμοποιηθεί σε συνδυασμό με την OWL. Αφορά στην κωδικοποίηση υπαρχόντων θησαυρών, συστημάτων ταξινόμησης, ταξονομιών κ.α. και στην ενσωμάτωσή τους στο Σημασιολογικό Ιστό με τη μορφή συνδεδεμένων δεδομένων.

Το μοντέλο SKOS διαθέτει απλό συντακτικό το οποίο γίνεται εύκολα κατανοητό από τον άνθρωπο. Οι κατηγορίες ατόμων θεωρούνται ως *έννοιες (concepts)*. Το SKOS επιτρέπει τη μοντελοποίηση της ιεραρχίας των εννοιών αλλά και τη δήλωση εναλλακτικών ονομασιών για κάθε έννοια.

3.3 Συσταδοποίηση

Με τον όρο Συσταδοποίηση ή Clustering αναφέρεται η διαδικασία τμηματοποίησης ενός συνόλου αντικειμένων σε επιμέρους υποσύνολα υψηλής νοηματικής συνάφειας [12].

Αξίζει να σημειωθεί ότι ο όρος Συσταδοποίηση αναφέρεται γενικά στη διαδικασία τμηματοποίησης ενός συνόλου δεδομένων και όχι σε μία συγκεκριμένη μέθοδο ή ένα συγκεκριμένο αλγόριθμο τμηματοποίησης.

Η Συσταδοποίηση είναι μία διαδικασία με εφαρμογές σε πληθώρα επιστημών, όπως η Βιολογία, η Ιατρική, στις Κοινωνικές Επιστήμες και φυσικά στην Επιστήμη της Πληροφορικής. Αποτελεί μία από τις κυριότερες μεθόδους ανάλυσης Μαζικών Δεδομένων.

3.4 Μέθοδοι συσταδοποίησης

Για την επίτευξη της Συσταδοποίησης ενός συνόλου δεδομένων έχουν σχεδιαστεί πολλοί διαφορετικοί αλγόριθμοι, ο καθένας με διαφορετικά πλεονεκτήματα αλλά και αδυναμίες.

Η Συσταδοποίηση επιτυγχάνεται κυρίως με τρεις βασικές μεθόδους:

- Συσταδοποίηση k-μέσων (k-means clustering)
- Ιεραρχική Συσταδοποίηση (hierarchical clustering)
- Συσταδοποίηση πυκνότητας (density-based clustering)

Οι περισσότεροι αλγόριθμοι Συσταδοποίησης αποτελούν παραλλαγές των τριών παραπάνω βασικών μεθόδων. Στις επόμενες παραγράφους γίνεται προσπάθεια επεξήγησης κάθε μίας από τις βασικές μεθόδους Συσταδοποίησης, και επίσης παρουσιάζονται τα πλεονεκτήματα και οι δυσκολίες που παρουσιάζει η εφαρμογή της κάθε μεθόδου.

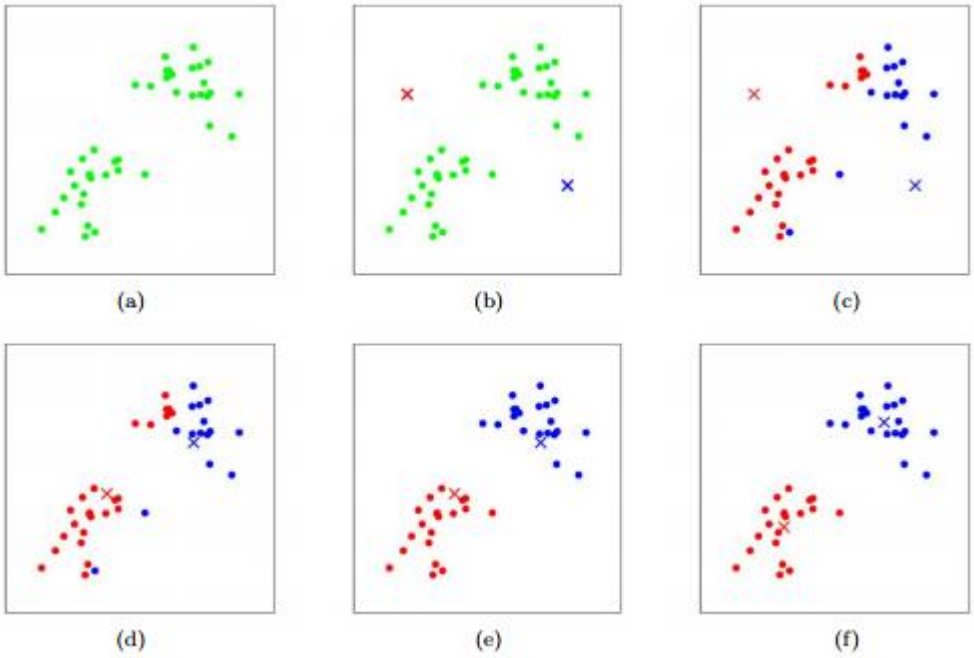
3.4.1 Συσταδοποίηση k-μέσων

Στην τεχνική αυτή γίνεται προσπάθεια συσταδοποίησης των στοιχείων ενός συνόλου διανυσμάτων γύρω από κεντρικά διανύσματα αναπαράστασης. Ο αριθμός των συστάδων (υποσυνόλων) στις οποίες πρόκειται να χωριστούν τα στοιχεία του συνόλου πρέπει να έχει προαποφασιστεί.

Έστω ότι το σύνολο των στοιχείων πρόκειται να χωριστεί σε k συστάδες. Η Συσταδοποίηση επιτυγχάνεται με τον αλγόριθμο *k-means* [13], ο οποίος περιγράφεται παρακάτω:

1. Ορίζονται αυθαίρετα k τυχαία διανύσματα αναπαράστασης v_i, που λειτουργούν ως αντιπρόσωποι των k συστάδων
2. Για κάθε στοιχείο e του συνόλου δεδομένων, υπολογίζεται η απόστασή του από το τον αντιπρόσωπο – κέντρο όλων των συστάδων. Επιλέγεται ο αντιπρόσωπος v_i με την ελάχιστη απόσταση από το e και το στοιχείο e τοποθετείται στη συστάδα i.
3. Για κάθε συστάδα i, υπολογίζεται ο νέος αντιπρόσωπος – κέντρο v_i ως ο μέσος όρος όλων των στοιχείων – διανυσμάτων που έχουν τοποθετηθεί στη συστάδα i.

4. Τα βήματα 2, 3 επαναλαμβάνονται εναλλάξ έως ότου να μην προκύψουν αλλαγές στην ταξινόμηση των στοιχείων του συνόλου ή/και στους υπολογισμένους αντιπροσώπους – κέντρα των συστάδων, ή μέχρι να ξεπεραστεί ένας προκαθορισμένος μέγιστος αριθμός επαναλήψεων.



Σχήμα 3.1: Παράδειγμα εκτέλεσης αλγορίθμου k-means

Ως μέτρο της απόστασης μεταξύ των διανυσμάτων χρησιμοποιείται η Ευκλείδεια απόσταση:

Έστω δύο διανύσματα $\boldsymbol{p}, \boldsymbol{q} \epsilon R^n$. Τότε η Ευκλείδεια απόσταση μεταξύ των $\boldsymbol{p}, \boldsymbol{q}$ είναι ίση με το μήκος του ευθύγραμμου τμήματος που συνδέει τα $\boldsymbol{p}, \boldsymbol{q}$:

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Ο αλγόριθμος k-means είναι πολύ συνηθισμένος και εύκολος στην υλοποίηση, ωστόσο παρουσιάζει ορισμένα μειονεκτήματα:

- Ο αριθμός των συστάδων k πρέπει να δοθεί ως παράμετρος εισόδου.
- Ο προσδιορισμός των αρχικών k αντιπροσώπων πρέπει να γίνει χειροκίνητα.
- Μία όχι καλή αρχικοποίηση αντιπροσώπων μπορεί να κάνει τον αλγόριθμο να συγκλίνει σε κάποιο τοπικό ελάχιστο που αντιπροσωπεύει λανθασμένη συσταδοποίηση.

Στο Σχήμα 3.1 παρουσιάζεται ένα παράδειγμα εκτέλεσης του αλγορίθμου *k-means*. Αρχικά (a) τα στοιχεία του συνόλου που παρουσιάζονται με πράσινο χρώμα δεν έχουν

τοποθετηθεί σε καμία συστάδα. Στη συνέχεια (b) τοποθετούνται δύο τυχαίοι αντιπρόσωποι των δύο συστάδων (με κόκκινο και μπλε χρώμα) στις οποίες πρόκειται να χωριστεί το σύνολο των στοιχείων. Στο επόμενο βήμα (c) το κάθε στοιχείο τοποθετείται στη συστάδα του πλησιέστερου σε αυτό αντιπροσώπου. Έπειτα (d) υπολογίζεται ο νέος αντιπρόσωπος – μέσος όρος κάθε συστάδας. Η διαδικασία επαναλαμβάνεται έως ότου ο αλγόριθμος συγκλίνει (f).

3.4.2 Ιεραρχική Συσταδοποίηση

Όπως υποδεικνύεται από το όνομά της, η Ιεραρχική Συσταδοποίηση [12] είναι μία τεχνική Συσταδοποίησης που αποσκοπεί στη δημιουργία μίας ιεραρχίας συστάδων.

Οι μέθοδοι Ιεραρχικής Συσταδοποίησης ταξινομούνται σε δύο κύριες κατηγορίες:

- Διαιρετικές μέθοδοι (divisive)
- Συσσωρευτικές μέθοδοι (agglomerative)

Στις **διαιρετικές μεθόδους Ιεραρχικής Συσταδοποίησης**, όλα τα στοιχεία αρχικά βρίσκονται σε μία και μοναδική συστάδα. Σε κάθε βήμα, η συστάδα που παρουσιάζει τη μεγαλύτερη διαφοροποίηση διαιρείται σε δύο επιμέρους συστάδες, έως ότου το σύνολο των n στοιχείων χωριστεί σε n συστάδες αποτελούμενες από ένα στοιχείο η κάθε μία.

Στις **συσσωρευτικές μεθόδους Ιεραρχικής Συσταδοποίησης**, το κάθε στοιχείο αρχικά βρίσκεται μόνο του σε μία δική του συστάδα. Σε κάθε βήμα, ο αριθμός των συστάδων μειώνεται κατά 1, πράγμα που επιτυγχάνεται συνενώνοντας (συσσωρεύοντας) τις δύο πλησιέστερες συστάδες.

Για την εύρεση των δύο πλησιέστερων συστάδων σε κάθε βήμα μίας συσσωρευτικής μεθόδου Ιεραρχικής Συσταδοποίησης, είναι ανάγκη να οριστεί μία μετρική απόστασης συστάδων. Συνήθως χρησιμοποιούνται δύο είδη ορισμού απόστασης συστάδων, με την τεχνική Ιεραρχικής Συσταδοποίησης να χαρακτηρίζεται αντίστοιχα ως:

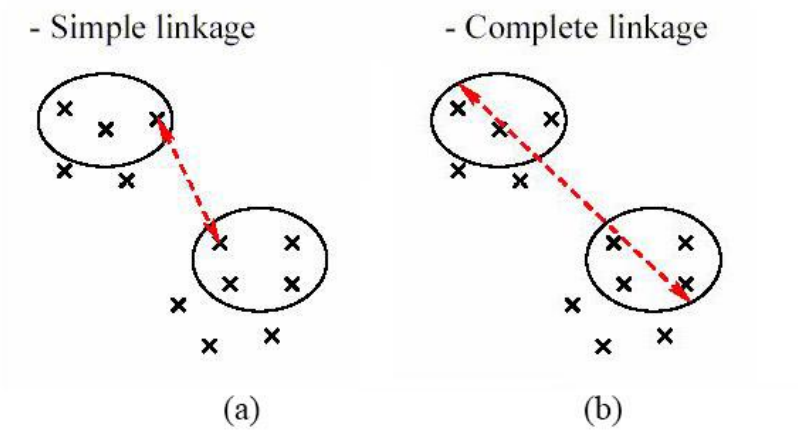
- Συσταδοποίηση μοναδικής διασύνδεσης (single-linkage clustering)
- Συσταδοποίηση πλήρους διασύνδεσης (complete-linkage clustering)

Στη **συσταδοποίηση μοναδικής διασύνδεσης**, η απόσταση δύο συστάδων s_1, s_2 ορίζεται ως η ελάχιστη απόσταση των στοιχείων της συστάδας s_1 από τα στοιχεία της συστάδας s_2 (Σχήμα 3.2.a):

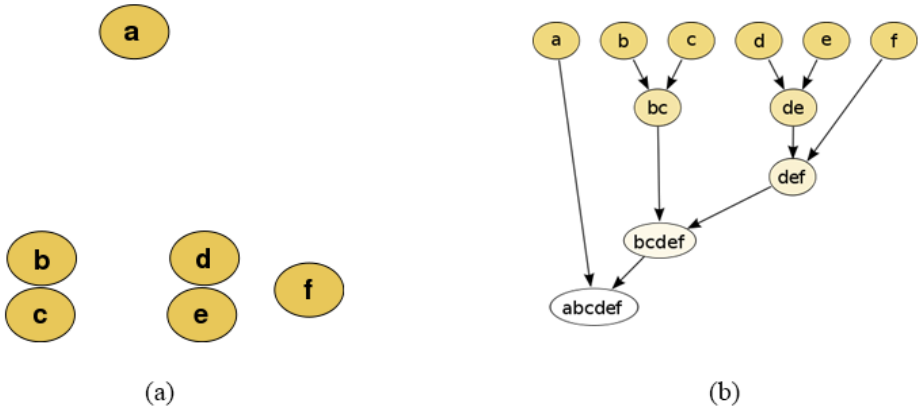
$$d(s_1, s_2) = min\{\, d(x, y) : x \in s_1, y \in s_2 \,\}$$

Στη **συσταδοποίηση πλήρους διασύνδεσης**, η απόσταση δύο συστάδων s_1, s_2 ορίζεται ως η μέγιστη απόσταση των στοιχείων της συστάδας s_1 από τα στοιχεία της συστάδας s_2 (Σχήμα 3.2.b):

$$d(s_1, s_2) = max\{\, d(x, y) : x \in s_1, y \in s_2 \,\}$$



Σχήμα 3.2: Ορισμοί απόστασης συστάδων στην Ιεραρχική Συσταδοποίηση



Σχήμα 3.3: Παράδειγμα Ιεραρχικής Συσταδοποίησης

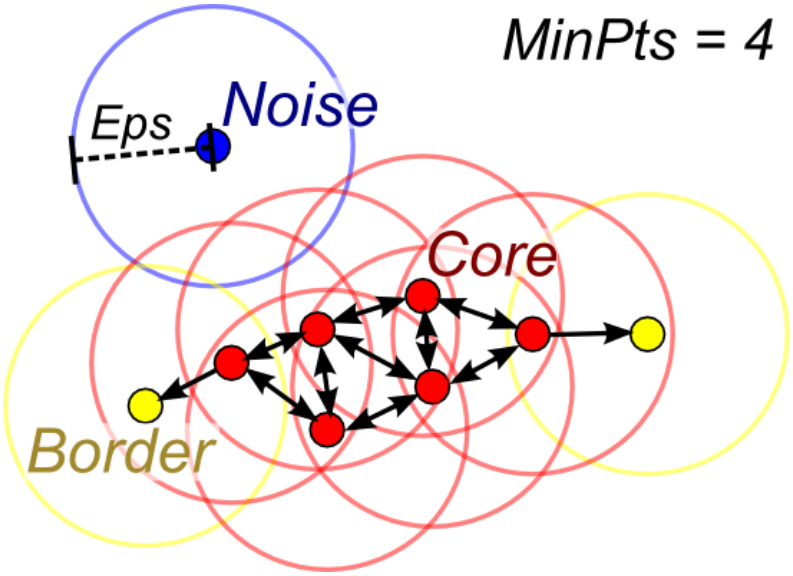
(a) Δεδομένα *(b) Δενδρόγραμμα*

Τα αποτελέσματα εκτέλεσης Συσσωρευτικής Ιεραρχικής Συσταδοποίησης παρουσιάζονται συνήθως σε ένα δενδρόγραμμα.

Στο Σχήμα 3.3 παρουσιάζεται ένα παράδειγμα εκτέλεσης Ιεραρχικής Συσταδοποίησης.

3.4.3 Συσταδοποίηση πυκνότητας

Η μέθοδος της Συσταδοποίησης πυκνότητας οδηγεί στη δημιουργία συστάδων που περιέχουν στοιχεία που βρίσκονται κοντά το ένα στο άλλο σε περιοχές υψηλής πυκνότητας του χώρου του συνόλου δεδομένων. Ταυτόχρονα, τα στοιχεία που βρίσκονται σε περιοχές χαμηλής πυκνότητας χαρακτηρίζονται ως θόρυβος (outliers) [14].



Σχήμα 3.4: Παράδειγμα διαχωρισμού σημείων κατά την εκτέλεση αλγορίθμου DBSCAN

Για την κατανόηση της μεθόδου Συσταδοποίησης πυκνότητας, κρίνεται σκόπιμη η μελέτη του αλγορίθμου Density-based spatial clustering of applications with noise (DBSCAN) [15], ο οποίος αποτελεί τον πιο συνηθισμένο αλγόριθμο Συσταδοποίησης πυκνότητας και τη βάση πολλών άλλων αλγορίθμων της ίδιας κατηγορίας.

Ο αλγόριθμος DBSCAN [16], εκτός από το σύνολο των στοιχείων προς συσταδοποίηση, δέχεται ως είσοδο δύο επιπλέον παραμέτρους:

- την ακτίνα αναζήτησης ε (ή Eps)
- τον ελάχιστο αριθμό γειτόνων $minPts$

Κατά την εκτέλεση του αλγορίθμου DBSCAN τα στοιχεία (σημεία – points) του συνόλου δεδομένων χωρίζονται σε τρεις κατηγορίες:

- σημεία πυρήνα (core points)
- σημεία περιθωρίου (edge points)
- θόρυβος (noise points)

Ο διαχωρισμός γίνεται ως εξής: Για κάθε σημείο p του συνόλου δεδομένων, γίνεται αναζήτηση γύρω από αυτό σε ακτίνα ε και καταμετρούνται τα σημεία – γείτονες p_n του p.

- Αν ο αριθμός των γειτόνων του p εντός της ακτίνας ε είναι τουλάχιστον $minPts$, τότε το σημείο p χαρακτηρίζεται ως **σημείο πυρήνα**.
- Αν το σημείο p έχει λιγότερους από $minPts$ γείτονες στη δικιά του ε-περιοχή, αλλά βρίσκεται εντός της ε-περιοχής ενός άλλου σημείου – πυρήνα, τότε το p χαρακτηρίζεται ως **σημείο περιθωρίου**.
- Όλα τα υπόλοιπα σημεία χαρακτηρίζονται ως **θόρυβος**.

Στο Σχήμα 3.4 παρουσιάζεται ένα παράδειγμα διαχωρισμού σημείων κατά την εκτέλεση αλγορίθμου DBSCAN με παράμετρο $minPts = 4$. Τα σημεία με κόκκινο χρώμα περιέχουν στην ε-περιοχή τους τουλάχιστον 4 σημεία (συμπεριλαμβανομένου του εαυτού τους) επομένως χαρακτηρίζονται ως σημεία πυρήνα. Τα σημεία με κίτρινο χρώμα περιλαμβάνουν λιγότερα από 4 σημεία εντός της ε-περιοχής τους, βρίσκονται όμως μέσα στην ε-περιοχή κάποιου άλλου σημείου που έχει χαρακτηριστεί ως σημείο πυρήνα. Τέλος, το σημείο με μπλε χρώμα δεν εμπίπτει σε καμία από τις δύο προηγούμενες περιπτώσεις, επομένως χαρακτηρίζεται ως θόρυβος.

Για δεδομένη τιμή των παραμέτρων ε και $minPts$, ο αλγόριθμος DBSCAN συνοψίζεται σε μία απλουστευμένη μορφή ως εξής:

1. Προσδιορισμός όλων των **σημείων πυρήνα** με τουλάχιστον $minPts$ γείτονες εντός ακτίνας ε.
2. Ομαδοποίηση των ε-γειτονικών σημείων πυρήνα σε **συστάδες.**
3. Τοποθέτηση καθενός από τα εναπομείναντα σημεία σε μία από τις κοντινές (που απέχουν το πολύ απόσταση ε) συστάδες και χαρακτηρισμός του ως **σημείου περιθωρίου**. Διαφορετικά, αν δεν υπάρχουν ε-κοντινές συστάδες, χαρακτηρισμός του σημείου ως **θόρυβος.**

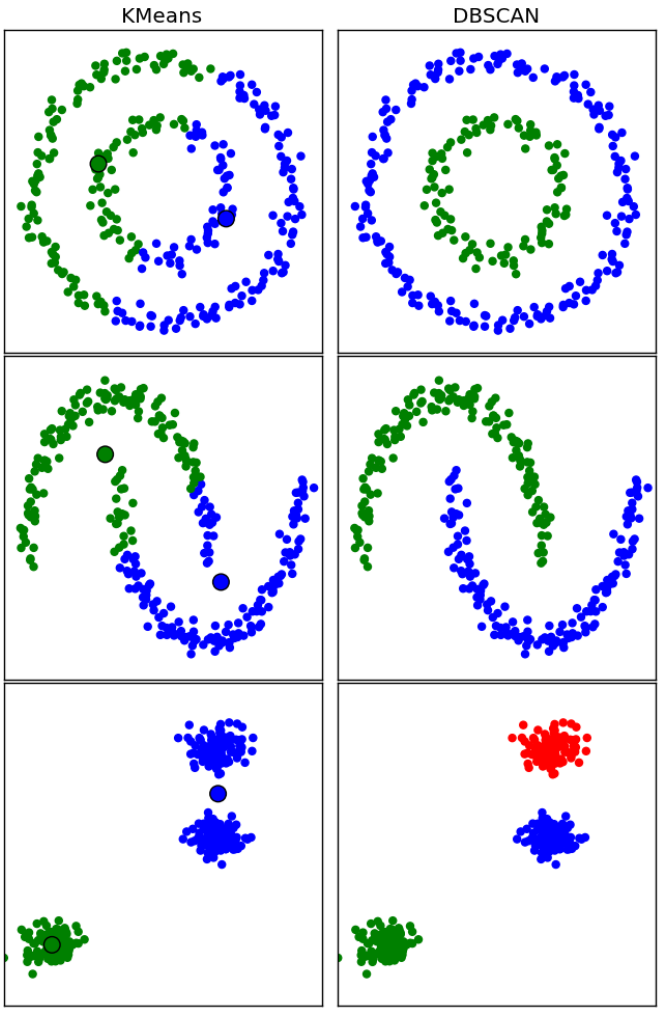
Στο Σχήμα 3.5 συγκρίνονται τα αποτελέσματα Συσταδοποίησης με εφαρμογή των αλγορίθμων *k-means* και DBSCAN σε τρία διαφορετικά παραδείγματα συνόλου στοιχείων.

Τα πλεονεκτήματα που παρουσιάζει ο αλγόριθμος DBSCAN είναι ποικίλα, μεταξύ των οποίων τα κυριότερα είναι:

- Δεν απαιτείται εξ αρχής ο προσδιορισμός των συστάδων στις οποίες πρόκειται να διαχωριστούν τα στοιχεία του συνόλου δεδομένων.
- Είναι δυνατός ο εντοπισμός συστάδων μη τετριμμένου σχήματος.
- Γίνεται αποτελεσματική διαχείριση του θορύβου.

Παράλληλα, όμως, παρουσιάζει και κάποια μειονεκτήματα, κυρίως τα εξής:

- Δεν είναι δυνατή η αποτελεσματική διαχείριση συνόλων δεδομένων με μεγάλες διαφοροποιήσεις στην πυκνότητα των στοιχείων, καθώς δεν είναι δυνατόν να προσδιοριστεί ένα ζεύγος παραμέτρων ε και $minPts$ που να είναι κατάλληλο για όλες τις συστάδες.
- Ο αλγόριθμος δεν είναι εντελώς ντετερμινιστικός, καθώς η σειρά προσπέλασης των δεδομένων μπορεί να επηρεάσει την τοποθέτηση των σημείων περιθωρίου σε συστάδες.



Σχήμα 3.5: Σύγκριση εφαρμογής αλγορίθμων k-means και DBSCAN σε κοινά σύνολα δεδομένων

3.5 Συσταδοποίηση κειμένων

Η Συσταδοποίηση κειμένων είναι η εφαρμογή τεχνικών συσταδοποίησης σε ένα σύνολο εγγράφων, συνήθως με σκοπό την αυτόματη ταξινόμησή τους σε κατηγορίες, την εξαγωγή θέματος ή τη γρήγορη ανεύρεση πληροφοριών [4].

Σε αυτή την παράγραφο περιγράφονται συνηθισμένες τεχνικές που ακολουθούνται για τη Συσταδοποίηση κειμένων [17].

3.5.1 Αναπαράσταση κειμένων στο μοντέλο διανυσματικού χώρου

Προκειμένου να είναι δυνατή η Συσταδοποίηση κειμένων, είναι ανάγκη να αναπτυχθούν τεχνικές αναπαράστασης κειμένων από διανύσματα. Για αυτό το λόγο αναπτύχθηκε η έννοια του **μοντέλου διανυσματικού χώρου (vector space model)** [18].

Το μοντέλο διανυσματικού χώρου είναι ένα αλγεβρικό μοντέλο αναπαράστασης κειμένων με τη μορφή διανυσμάτων από χαρακτηριστικά. Τέτοια χαρακτηριστικά, για παράδειγμα, μπορεί να είναι:

- το μέγεθος του κειμένου (σε σελίδες / λέξεις / χαρακτήρες)
- το είδος του κειμένου (για παράδειγμα άρθρο, επιστολή)
- η ύπαρξη ή μη συγκεκριμένων λέξεων – κλειδιών
- η απόλυτη ή σχετική συχνότητα εμφάνισης συγκεκριμένων όρων

Παρακάτω παρουσιάζονται δύο συνηθισμένες τεχνικές αναπαράστασης εγγράφων στο μοντέλο διανυσματικού χώρου: το **μοντέλο bag-of-words** και το **μοντέλο βαρών tf-idf**.

3.5.1.1 Μοντέλο bag-of-words

Στο μοντέλο bag-of-words, ένα έγγραφο αναπαρίσταται ως ένα σύνολο όρων / λέξεων, μαζί με τη συχνότητα εμφάνισης των όρων εντός του εγγράφου [19].

Τα διανύσματα αναπαράστασης των εγγράφων έχουν τόσες διαστάσεις όσοι είναι και οι διαφορετικοί όροι που εμφανίζονται στο σύνολο των εγγράφων. Κάθε διάσταση αντιστοιχεί σε έναν από τους όρους αυτούς. Το διάνυσμα αναπαράστασης ενός εγγράφου παίρνει σε κάθε διάσταση τιμή ίση με τη συχνότητα εμφάνισης του αντίστοιχου όρου στο έγγραφο αυτό.

Παρακάτω παρουσιάζεται ένα παράδειγμα εφαρμογής αναπαράστασης δύο αποσπασμάτων στο μοντέλο bag-of-words.

Έστω τα παρακάτω αποσπάσματα:

1. John likes to watch movies. Mary likes movies too.
2. John also likes to watch football games.

Σε αυτό το παράδειγμα, στα αποσπάσματα μπορούν να εντοπιστούν 10 διαφορετικές λέξεις. Επομένως, το διάνυσμα αναπαράστασης των αποσπασμάτων – εγγράφων αποτελείται από 10 διαστάσεις, κάθε μία από τις οποίες αντιστοιχεί στις λέξεις:

1. John
2. likes
3. to
4. watch
5. movies
6. Mary
7. too
8. also
9. football
10. games

Λέξη	Απόσπασμα 1	Απόσπασμα 2
John	1	1
likes	2	1
to	1	1
watch	1	1
movies	2	0
Mary	1	0
too	1	0
also	0	1
football	0	1
games	0	1

Πίνακας 3.1: Παράδειγμα αναπαράστασης κειμένου στο μοντέλο bag-of-words

Ο Πίνακας 3.1 παρουσιάζει την αναπαράσταση των δύο αποσπασμάτων στο μοντέλο bag-of-words.

3.5.1.2 Μοντέλο βαρών tf-idf

Το μοντέλο bag-of-words είναι αρκετά απλό και εύκολο στην κατανόηση και υλοποίηση, ωστόσο παρουσιάζει ένα πρόβλημα: λέξεις που είναι πολύ συνηθισμένες παρουσιάζουν πολύ υψηλές τιμές στα διανύσματα αναπαράστασης, χωρίς όμως αυτές οι υψηλές τιμές να αντιστοιχούν σε κάποιου είδους πληροφορία ή θεματική ενότητα.

Με το μοντέλο βαρών tf-idf [18] γίνεται προσπάθεια αντιμετώπισης αυτού ακριβώς του προβλήματος. Ο όρος tf-idf προκύπτει από τα εξής:

- **t**erm **f**requency: η συχνότητα εμφάνισης του όρου στο εκάστοτε κείμενο
- **i**nverse **d**ocument **f**requency: «αντίστροφη συχνότητα εγγράφου», δηλαδή το πλήθος των εγγράφων του συνόλου δεδομένων στα οποία εμφανίζεται ο όρος

Στο μοντέλο tf-idf, το διάνυσμα αναπαράστασης είναι όμοιο με αυτό του μοντέλου bag-of-words, με τη διαφορά ότι οι τιμές των διανυσμάτων κανονικοποιούνται σύμφωνα με την αντίστροφη συχνότητα εγγράφου του αντίστοιχου όρου.

Το μοντέλο tf-idf περιγράφεται αναλυτικότερα στην παράγραφο 5.1.

3.5.2 Απόσταση κειμένων στο μοντέλο διανυσματικού χώρου

Έχοντας αναπαραστήσει τα έγγραφα στο μοντέλο διανυσματικού χώρου, είναι δυνατόν να ορίσουμε τη μεταξύ τους απόσταση χρησιμοποιώντας αλγεβρικές εκφράσεις μεταξύ των διανυσμάτων αναπαράστασης [20].

Ένας απλός και συνηθισμένος τρόπος υπολογισμού απόστασης εγγράφων στο μοντέλο διανυσματικού χώρου είναι η απόσταση συνημιτόνου, η οποία περιγράφεται στην παράγραφο 5.1.

4 Dataset

Για τη δημιουργία του dataset του πειράματος επιλέχθηκαν ~1700 δημοσιεύσεις από την Ψηφιακή Βιβλιοθήκη της ACM επισημειωμένες τόσο με ελεύθερες ετικέτες ή λέξεις – κλειδιά (keywords) όσο και με ετικέτες από το Σύστημα Ταξινόμησης της ACM.

4.1 Σύστημα Ταξινόμησης ACM

Το Σύστημα Ταξινόμησης της ACM (2012 ACM Computing Classification System) αναπτύχθηκε με τη μορφή μιας πολυ-ιεραρχικής οντολογίας που μπορεί να χρησιμοποιηθεί σε εφαρμογές σημασιολογικού ιστού [8]. Βασίζεται σε ένα ψηφιακό λεξιλόγιο και λειτουργεί ως μία μοναδική για την ACM πηγή κατηγοριών και εννοιών της σύγχρονης Επιστήμης των Υπολογιστών.

Κάθε ετικέτα του Συστήματος Ταξινόμησης της ACM αναφέρεται σε ένα πεδίο επιστημονικού ενδιαφέροντος της Επιστήμης της Πληροφορικής. Κάθε ετικέτα διαθέτει ένα μοναδικό αριθμητικό αναγνωριστικό, έναν προτεινόμενο τίτλο και προαιρετικά έναν ή περισσότερους εναλλακτικούς τίτλους. Επίσης, για κάθε ετικέτα προσδιορίζονται οι γονικές ετικέτες – υπερκατηγορίες αυτής καθώς και οι υποκατηγορίες αυτής, εφόσον υπάρχουν.

Από τεχνικής πλευράς, η οντολογία παρέχεται σε μορφή SKOS [21], [11].

4.2 Ψηφιακή Βιβλιοθήκη ACM

Η Ψηφιακή Βιβλιοθήκη της ACM (ACM Digital Library) είναι μία από τις πιο ολοκληρωμένες βάσεις δεδομένων παγκοσμίως για άρθρα πλήρους κειμένου και βιβλιογραφία που αφορά σε υπολογιστικές τεχνολογίες και τεχνολογίες πληροφορικής. Αυτή η βιβλιοθήκη περιλαμβάνει την πλήρη συλλογή των δημοσιεύσεων της ACM καθώς και μια εκτεταμένη βιβλιογραφική βάση δεδομένων σημαντικών έργων στον τομέα της πληροφορικής από ακαδημαϊκούς συγγραφείς.

Οι ενσωματωμένες λειτουργίες αλλά και το περιεχόμενο της Ψηφιακής Βιβλιοθήκης της ACM αποτελούν κρίσιμη πηγή για τους επαγγελματίες αλλά και τους ερευνητές του χώρου.

Η Ψηφιακή Βιβλιοθήκη έχει σχεδιαστεί για να διευκολύνει τη διάδοση και ανταλλαγή πληροφοριών, τη διαλειτουργικότητα, το σχεδιασμό με επίκεντρο το χρήστη και τη συνεργασία μεταξύ επαγγελματιών, ερευνητών και εκπαιδευτικών. Προσφέρει πρόσβαση σε ιδέες και καινοτομίες που συνεχίζουν να καλλιεργούν και να διαμορφώνουν την εποχή της πληροφορίας.

Η δημιουργία κρίσιμου περιεχομένου που μπορεί να ανακαλυφθεί και είναι ευρέως προσβάσιμη υπήρξε πρωταρχικός στόχος της Ψηφιακής Βιβλιοθήκης της ACM από την ίδρυσή της. Η επέκταση του πεδίου δράσης της Ψηφιακής Βιβλιοθήκης πέρα από τις δημοσιεύσεις συνεδρίων, περιοδικών, ενημερωτικών περιοδικών και ενημερωτικών δελτίων της ACM, ώστε να συμπεριλάβει πλήρως ολοκληρωμένα βιβλιογραφικά δεδομένα όλων των υπολογιστικών βιβλίων, έχει αποδειχθεί εξαιρετικά σημαντικό μέρος αυτού του πρωταρχικού στόχου. Η Ψηφιακή Βιβλιοθήκη στοχεύει να είναι ένας προορισμός όπου η παρουσίαση και η συνεργασία επιτρέπουν τη διαμόρφωση σχέσεων, την επέκταση των ορίων του παρελθόντος και την οραματισμό του μέλλοντος.

4.3 Δημιουργία του dataset

4.3.1 Αναζήτηση και λήψη εγγράφων

Για τη δημιουργία του dataset αρχικά επιλέχθηκαν 13 όροι αναζήτησης, και συγκεκριμένα:

- computational complexity
- artificial intelligence
- semantic web
- human computer interaction
- machine learning
- computer graphics
- cryptography
- network security
- network algorithms
- parallel computing
- genetic algorithm
- clustering algorithms
- deep learning

Οι παραπάνω όροι επιλέχθηκαν έτσι ώστε να αναφέρονται σε συγγενικούς τομείς της Επιστήμης της Πληροφορικής, οι οποίοι όμως να παρουσιάζουν και επαρκείς διαφοροποιήσεις ως προς τις έννοιες αλλά και το σχετικό λεξιλόγιο.

Για κάθε έναν από τους όρους αναζήτησης πραγματοποιήθηκε αναζήτηση στην Ψηφιακή Βιβλιοθήκη της ACM και αποθηκεύθηκαν κατά μέσο όρο 310 έγγραφα σε μορφή PDF, δημιουργώντας συνολικά μία βάση από 4040 έγγραφα.

Για την αποθήκευση των μεταδεδομένων των εγγράφων του dataset δημιουργήθηκε μία βάση δεδομένων. Σε αυτήν διατηρήθηκαν όλα τα διαθέσιμα μεταδεδομένα κάθε εγγράφου από την Ψηφιακή Βιβλιοθήκη της ACM. Τα κυριότερα από αυτά είναι τα εξής:

- το μοναδικό id του εγγράφου στην Ψηφιακή Βιβλιοθήκη της ACM
- ο τίτλος
- ο συγγραφέας ή οι συγγραφείς
- το έτος δημοσίευσης
- ο αριθμός των σελίδων
- οι ελεύθερες λέξεις – κλειδιά
- οι ετικέτες από το Σύστημα Ταξινόμησης της ACM

4.3.2 Διαλογή εγγράφων

Εν συνεχεία, επιλέχθηκαν όσα έγγραφα διαθέτουν τα παρακάτω χαρακτηριστικά:

- περιέχουν τουλάχιστον 2 σελίδες
- περιέχουν μία από τις λέξεις Abstract, Introduction σε τίτλο ενότητας
- περιέχουν μία από τις λέξεις References, Bibliography σε τίτλο ενότητας
- έχουν καταχωρημένες ελεύθερες λέξεις – κλειδιά (τουλάχιστον μία)
- έχουν καταχωρημένες ετικέτες από το Σύστημα Ταξινόμησης της ACM (τουλάχιστον μία)

Τα παραπάνω κριτήρια οδήγησαν εν τέλει στην επιλογή 1757 εγγράφων.

4.3.3 Μετατροπή σε καθαρό κείμενο

Τέλος, το κάθε έγγραφο μετατράπηκε σε txt μορφή. Εκκαθαρίστηκαν οι «άκρες» του εγγράφου. Συγκεκριμένα, διατηρήθηκε το κείμενο ανάμεσα στους τίτλους των ενοτήτων με τις λέξεις Abstact / Introduction και References / Bibliography. Έτσι, κατά τη δημιουργία της αρχικής συσταδοποίησης με βάση το κείμενο, η οποία θεωρήθηκε ως ground truth ή συσταδοποίηση αναφοράς, δε λήφθηκαν υπόψη ο τίτλος του εγγράφου και η βιβλιογραφία.

5 *Αξιολόγηση Επισημειώσεων με τη Μέθοδο της Συσταδοποίησης*

Σε αυτό το κεφάλαιο περιγράφονται τα επιμέρους βήματα του πειράματος που πραγματοποιήθηκε για τη δημιουργία του προτεινόμενου συστήματος αξιολόγησης επισημειώσεων. Εν συντομία, αρχικά κατασκευάζεται μία αρχική συσταδοποίηση με βάση το καθαρό κείμενο των εγγράφων του dataset. Στη συνέχεια κατασκευάζονται νέες συσταδοποιήσεις με βάση διαφορετικές εκδοχές επισημειώσεων των εγγράφων και κάθε νέα συσταδοποίηση συγκρίνεται με την αρχική.

Προκειμένου να γίνει συσταδοποίηση σε ένα σύνολο δεδομένων, είναι ανάγκη να οριστεί η απόσταση μεταξύ δύο στοιχείων του συνόλου. Η κατασκευή διαφορετικών εκδοχών συσταδοποίησης του συνόλου δεδομένων πραγματοποιείται με:

- υλοποίηση εναλλακτικών ορισμών της απόστασης μεταξύ δύο εγγράφων
- χρήση διαφορετικών αλγορίθμων συσταδοποίησης

Στις επόμενες παραγράφους εξετάζονται οι διαφορετικοί ορισμοί απόστασης σε συνδυασμό με τη χρήση διαφορετικών αλγορίθμων συσταδοποίησης που οδήγησαν στις διάφορες συσταδοποιήσεις του συνόλου δεδομένων.

5.1 Δημιουργία αρχικής συσταδοποίησης με βάση το καθαρό κείμενο

Για τη δημιουργία της αρχικής συσταδοποίησης, η οποία χρησιμοποιήθηκε ως ground truth ή συσταδοποίηση αναφοράς, χρησιμοποιήθηκε ως απόσταση κειμένων το συνημίτονο της γωνίας των διανυσμάτων αναπαράστασης των κειμένων στο μοντέλο διανυσματικού χώρου (vector space model), όπως ορίστηκε από τους G. Salton, A. Wond και C. S. Yang [18].

Για κάθε κείμενο d κατασκευάζεται ένα διάνυσμα αναπαράστασης $\boldsymbol{v_d}$ που περιλαμβάνει τα βάρη *tf-idf* κάθε όρου από το dataset:

$$\boldsymbol{v_d} = \left[w_{1,d}, w_{2,d}, \dots, w_{N,d}\right]^T$$

όπου

$$w_{t,d} = tf_{t,d} \cdot \log \frac{|\mathrm{D}|}{|\{d' \in D \mid t \in d'\}|}$$

και

- $tf_{t,d}$ είναι η συχνότητα του όρου t στο έγγραφο d
- $\log \frac{|\mathrm{D}|}{|\{d' \in D \mid t \in d'\}|}$ είναι η αντίστροφη συχνότητα εγγράφου. $|\mathrm{D}|$ είναι ο συνολικός αριθμός των εγγράφων του dataset. $|\{d' \in D \mid t \in d'\}|$ είναι ο αριθμός των εγγράφων που περιέχουν τον όρο t.

Χρησιμοποιώντας το συνημίτονο ως μέτρο της ομοιότητας της διεύθυνσης δύο διανυσμάτων, η ομοιότητα δύο εγγράφων d_1, d_2 ορίζεται ως:

$$sim(d_1, d_2) = \frac{\boldsymbol{v_{d1}} \cdot \boldsymbol{v_{d2}}}{|\boldsymbol{v_{d1}}| \cdot |\boldsymbol{v_{d2}}|}$$

Η παραπάνω σχέση είναι δείκτης ομοιότητας. Για να μετατραπεί σε δείκτη απόστασης, τον αφαιρούμε από τη μέγιστη τιμή που μπορεί να πάρει, που είναι 1:

$$distance(d_1, d_2) = 1 - sim(d_1, d_2)$$

ή

$$distance(d_1, d_2) = \frac{\boldsymbol{v_{d1}} \cdot \boldsymbol{v_{d2}}}{|\boldsymbol{v_{d1}}| \cdot |\boldsymbol{v_{d2}}|}$$

Χρησιμοποιώντας τον παραπάνω ορισμό απόστασης δύο εγγράφων, πραγματοποιήθηκε συσταδοποίηση με τους αλγορίθμους Hierarchical Clustering και DBScan.

5.1.1 Hierarchical Clustering

Αρχικά δοκιμάστηκαν οι τεχνικές single-linkage agglomerative hierarchical clustering και complete-linkage agglomerative hierarchical clustering [4]. Συγκρίνοντας εποπτικά τις παραγόμενες συσταδοποιήσεις διαπιστώθηκε ότι το complete-linkage clustering παράγει υποσύνολα καλύτερης ποιότητας από πλευράς νοηματικής συνάφειας. Συνεπώς αυτή ήταν και η τεχνική που επιλέχθηκε για τη δημιουργία της αρχικής συσταδοποίησης αναφοράς.

Μέσω της τεχνικής complete-linkage agglomerative hierarchical clustering παράχθηκαν 240 εναλλακτικές συσταδοποιήσεις, με τον αριθμό των υποσυνόλων να ποικίλει από 11 ως 250 υποσύνολα αντιστοίχως.

5.1.2 DBScan

Έγινε προσπάθεια συσταδοποίησης με εφαρμογή του αλγορίθμου DBScan.

Αρχικά ο αλγόριθμος DBScan θεωρήθηκε καλή εναλλακτική, καθώς εντοπίζεται αυτόματα ο κατάλληλος αριθμός των υποσυνόλων και ταυτόχρονα γίνεται καλή διαχείριση του «θορύβου», δηλαδή – στο παρόν πείραμα – των εγγράφων που δε σχετίζονται επαρκώς με τα υπόλοιπα έγγραφα του dataset.

Ωστόσο, δεν ήταν δυνατόν να βρεθεί τιμή της παραμέτρου ε του αλγορίθμου που να οδηγεί σε ικανοποιητική συσταδοποίηση του συγκεκριμένου dataset. Για μικρές τιμές της παραμέτρου, η πλειοψηφία των εγγράφων θεωρούνται εσφαλμένα ως θόρυβος. Καθώς αυξάνεται η τιμή της παραμέτρου ε, ο αριθμός των ταξινομημένων εγγράφων παρουσιάζεται μεν αυξημένος, όμως όλα τα έγγραφα ταξινομούνται σε ένα μοναδικό υποσύνολο.

Τελικά, ο αλγόριθμος DBScan κρίθηκε ακατάλληλος για την παρούσα φάση του πειράματος.

5.2 Συσταδοποίηση με βάση τις ελεύθερες λέξεις – κλειδιά

Πριν γίνει προσπάθεια συσταδοποίησης με βάση τις ελεύθερες λέξεις – κλειδιά, πραγματοποιήθηκε ανάλυση των ελεύθερων λέξεων – κλειδιών με τις οποίες έχουν επισημειωθεί τα έγγραφα του dataset. Διαπιστώθηκε ότι στο σύνολο των 1757 εγγράφων οι συγγραφείς έχουν χρησιμοποιήσει για την επισήμανση των εγγράφων 4065 διαφορετικές μεταξύ τους ελεύθερες λέξεις – κλειδιά, από τις οποίες μόνο οι 741 (δηλαδή περίπου το 18,23%) έχουν χρησιμοποιηθεί σε τουλάχιστον δύο διαφορετικά έγγραφα.

Συνεπώς, το διάνυσμα αναπαράστασης των κειμένων με βάση τις ελεύθερες λέξεις κλειδιά προκύπτει στην πράξη υπερβολικά αραιό, γεγονός που δεν επέτρεψε την εφαρμογή συσταδοποίησης με τις ελεύθερες λέξεις – κλειδιά.

5.3 Συσταδοποίηση με βάση τις ετικέτες από το Σύστημα Ταξινόμησης ACM

Στα πλαίσια της επισήμανσης των 1757 εγγράφων με ετικέτες από το προκαθορισμένο λεξιλόγιο του Συστήματος Ταξινόμησης της ACM, οι συγγραφείς επέλεξαν μόλις 886 διαφορετικές ετικέτες, από τις οποίες οι 562 (64,9%) χρησιμοποιήθηκαν σε τουλάχιστον 2 έγγραφα. Επιπλέον, αξιοποιώντας τις ιεραρχικές σχέσεις που προσδιορίζει η οντολογία του Συστήματος Ταξινόμησης της ACM, είναι δυνατόν να οριστεί η απόσταση δύο ετικετών του

Συστήματος και κατ' επέκταση να οριστεί η απόσταση δύο εγγράφων, ακόμα και αν αυτά δεν έχουν επισημειωθεί με καμία κοινή ετικέτα.

5.3.1 Απόσταση ετικετών

Για την απόσταση δύο ετικετών του Συστήματος Ταξινόμησης της ACM προτείνονται δύο ορισμοί:

- Απλή απόσταση ετικετών (Naive Tag Distance)
- Απόσταση ετικετών με προσαρμοσμένα βάρη (Adjusted Tag Distance)

5.3.1.1 Απλή απόσταση ετικετών (Naive Tag Distance)

Για τον υπολογισμό της απόστασης δύο ετικετών του Συστήματος Ταξινόμησης της ACM, κατασκευάζεται ένας μη κατευθυνόμενος γράφος $G(V, E)$ με βάρη. Για κάθε ετικέτα t του Συστήματος Ταξινόμησης εισάγεται στο γράφο μία κορυφή v_t με id το id της ετικέτας. Στη συνέχεια, για κάθε ετικέτα t εντοπίζονται όλες οι ετικέτες $T_c = \{t_1, \dots, t_n\}$ οι οποίες έχουν ως γονέα την ετικέτα t και οι αντίστοιχες κορυφές $V_c = \{v_1, \dots, v_n\}$. Για κάθε κορυφή v_i του συνόλου V_c εισάγεται στο γράφο η ακμή $e(v_t, v_i)$ με βάρος $w_e = 1$.

Σε αυτήν την περίπτωση, η απόσταση δύο ετικετών t_1, t_2 είναι ίση με το μήκος του ελάχιστου μονοπατιού [22], [23] που συνδέει τις αντίστοιχες κορυφές v_1, v_2.

Για παράδειγμα, ας εξεταστεί η παρακάτω ενδεικτική κατηγοριοποίηση:

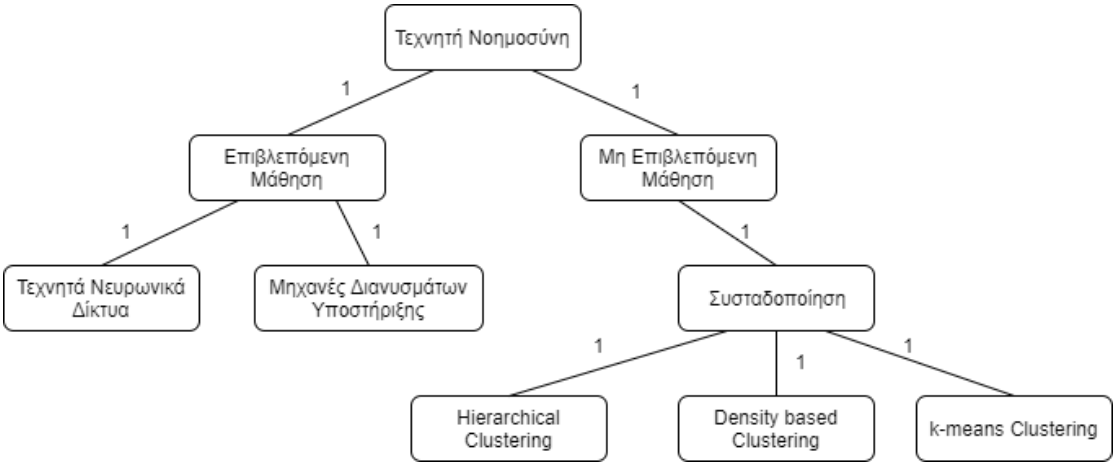
- Τεχνητή Νοημοσύνη
    - Επιβλεπόμενη Μάθηση
        - Τεχνητά Νευρωνικά Δίκτυα
        - Μηχανές Διανυσμάτων Υποστήριξης
    - Μη Επιβλεπόμενη Μάθηση
        - Συσταδοποίηση
            - Hierarchical Clustering
            - Density based Clustering
            - k-means Clustering

Σε αυτήν την περίπτωση, ο γράφος που προκύπτει παρουσιάζεται στο Σχήμα 5.1.

Η απόσταση των ετικετών *Επιβλεπόμενη Μάθηση* και *Τεχνητά Νευρωνικά Δίκτυα* είναι ίση με 1.

Η απόσταση των ετικετών *Τεχνητά Νευρωνικά Δίκτυα* και *Μηχανές Διανυσμάτων Υποστήριξης* είναι ίση με 2.

Τέλος, η απόσταση των ετικετών *Επιβλεπόμενη Μάθηση* και *Συσταδοποίηση* είναι ίση με 3.



Σχήμα 5.1: Γράφος απλής απόστασης ετικετών

5.3.1.2 Απόσταση ετικετών με προσαρμοσμένα βάρη (Adjusted Tag Distance)

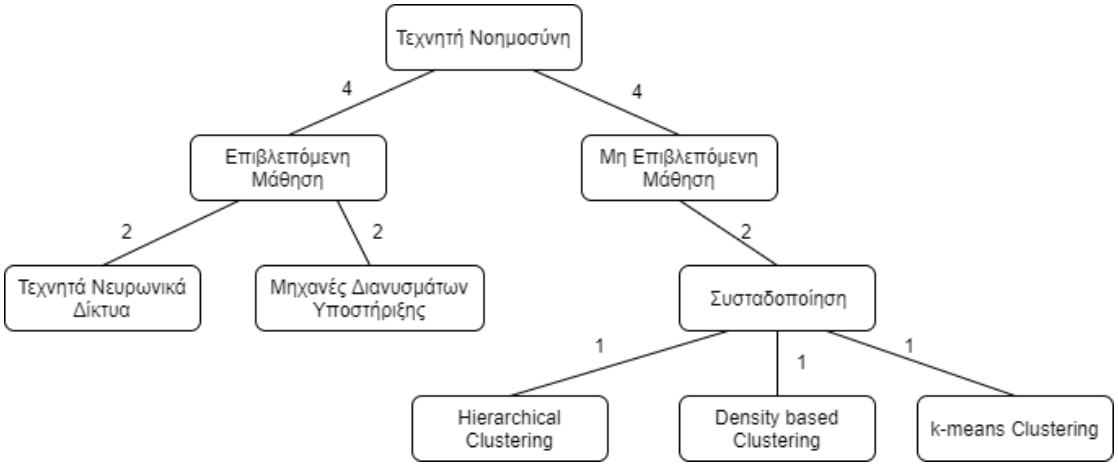
Η απλή απόσταση ετικετών ορίζεται με σαφή και ευνόητο τρόπο, ωστόσο παρουσιάζει ένα πρόβλημα: η απόσταση δύο ετικετών προκύπτει ανεξάρτητη του βάθους στο οποίο βρίσκονται οι ετικέτες. Με άλλα λόγια, η απόσταση δύο αρκετά εξειδικευμένων κατηγοριών κοινού γονέα είναι η ίδια με την απόσταση δύο γενικότερων κατηγοριών κοινού γονέα.

Για παράδειγμα, στο προηγούμενο παράδειγμα, οι ετικέτες *Επιβλεπόμενη Μάθηση* και *Μη Επιβλεπόμενη Μάθηση* έχουν την ίδια απόσταση με τις ετικέτες *Hierarchical Clustering* και *Density based Clustering*, παρόλο που το δεύτερο ζευγάρι ετικετών παρουσιάζει μεγαλύτερη εξειδίκευση και άρα συνάφεια.

Για να ξεπεραστεί αυτό το πρόβλημα, προτείνεται ο ορισμός της απόστασης ετικετών με προσαρμοσμένα βάρη.

Αρχικά υπολογίζεται το βάθος κάθε ετικέτας. Στη συνέχεια, οι ακμές του γράφου προστίθενται με βάρος που μειώνεται εκθετικά ως προς το βάθος της γονικής ετικέτας. Στο Σχήμα 5.2 παρουσιάζεται ο γράφος της ενδεικτικής κατηγοριοποίησης του προηγούμενου παραδείγματος με προσαρμοσμένα βάρη.

Χρησιμοποιώντας τον ορισμό της απόστασης ετικετών με προσαρμοσμένα βάρη, μοντελοποιείται καλύτερα η εξειδίκευση των κατηγοριών του Συστήματος Ταξινόμησης της ACM. Για παράδειγμα, οι γενικές κατηγορίες *Επιβλεπόμενη Μάθηση* και *Μη Επιβλεπόμενη Μάθηση* του παραδείγματος έχουν τώρα απόσταση ίση με 8, ενώ οι ειδικότερες κατηγορίες *Hierarchical Clustering* και *Density based Clustering*, που παρουσιάζουν μεγαλύτερη συσχέτιση, έχουν απόσταση ίση με 2.



Σχήμα 5.2: Γράφος απόστασης ετικετών με προσαρμοσμένα βάρη

5.3.2 Απόσταση εγγράφων

Έχοντας ορίσει την απόσταση δύο ετικετών από το Σύστημα Ταξινόμησης της ACM, είναι δυνατόν να οριστεί η απόσταση δύο εγγράφων με βάση τις ετικέτες από το Σύστημα Ταξινόμησης με τις οποίες τα έγγραφα αυτά έχουν επισημειωθεί.

Στην περίπτωση όπου και τα δύο έγγραφα έχουν ακριβώς μία ετικέτα, ο ορισμός της απόστασης των εγγράφων είναι απλός: η απόσταση των εγγράφων ορίζεται ως ίση με την απόσταση των δύο ετικετών.

Στην περίπτωση, όμως, όπου κάποιο από τα δύο έγγραφα ή και τα δύο έγγραφα είναι επισημειωμένα με δύο ή περισσότερες ετικέτες, υπάρχουν πολλοί εναλλακτικοί ορισμοί της απόστασής τους με βάση την απόσταση των ετικετών τους. Στην παρούσα εργασία προτείνονται τρεις διαφορετικοί ορισμοί απόστασης εγγράφων με βάση τις ετικέτες τους:

- Ελάχιστη απόσταση ετικετών
- Μέγιστη απόσταση ετικετών
- Μέση απόσταση ετικετών

5.3.2.1 Ελάχιστη απόσταση ετικετών

Η απόσταση δύο εγγράφων d_1, d_2 που είναι επισημειωμένα με ετικέτες $T_1 = \{t_{1,1}, \dots, t_{1,m}\}$ και $T_2 = \{t_{2,1}, \dots, t_{2,n}\}$ αντίστοιχα ορίζεται ως η ελάχιστη απόσταση όλων των συνδυασμών των ετικετών των δύο εγγράφων:

$$distance(d_1, d_2) = \min_{i=1..m, j=1..n} distance(t_{1,i}, t_{2,j})$$

5.3.2.2 Μέγιστη απόσταση ετικετών

Η απόσταση δύο εγγράφων d_1, d_2 που είναι επισημειωμένα με ετικέτες $T_1 = \{t_{1,1}, \dots, t_{1,m}\}$ και $T_2 = \{t_{2,1}, \dots, t_{2,n}\}$ αντίστοιχα ορίζεται ως η μέγιστη απόσταση όλων των συνδυασμών των ετικετών των δύο εγγράφων:

$$distance(d_1, d_2) = \max_{i=1..m, j=1..n} distance(t_{1,i}, t_{2,j})$$

5.3.2.3 Μέση απόσταση ετικετών

Στους δύο προηγούμενους ορισμούς, δοκιμάζονται όλα τα πιθανά ζεύγη ετικετών ανάμεσα στα δύο έγγραφα, όμως τελικά ένα μόνο ζεύγος καθορίζει την τελική απόσταση των δύο εγγράφων.

Με αυτόν τον ορισμό απόστασης δύο εγγράφων με βάση τις ετικέτες με τις οποίες έχουν επισημειωθεί, γίνεται προσπάθεια να ληφθούν υπόψη ταυτόχρονα όλες οι ετικέτες των δύο εγγράφων για τον υπολογισμό της μεταξύ τους απόστασης.

Έστω δύο έγγραφα d_1, d_2 που είναι επισημειωμένα με ετικέτες $T_1 = \{t_{1,1}, \dots, t_{1,m}\}$ και $T_2 = \{t_{2,1}, \dots, t_{2,n}\}$ αντίστοιχα. Αρχικά, για κάθε ετικέτα του εγγράφου d_1 εντοπίζεται η πλησιέστερη ετικέτα του εγγράφου d_2 και υπολογίζεται η μεταξύ τους απόσταση. Ακολουθεί η αντίστροφη διαδικασία, δηλαδή για κάθε ετικέτα του εγγράφου d_2 εντοπίζεται η πλησιέστερη ετικέτα του εγγράφου d_1 και υπολογίζεται η μεταξύ τους απόσταση. Τέλος, υπολογίζεται η απόσταση των δύο εγγράφων ως ο μέσος όρος όλων των αποστάσεων ετικετών που έχουν υπολογισθεί προηγουμένως.

$$distance(d_1, d_2) = \frac{\sum_{i=1}^{m} \min_{j=1..n} distance(t_{1,i}, t_{2,j}) + \sum_{j=1}^{n} \min_{i=1..m} distance(t_{2,j}, t_{1,i})}{m + n}$$

5.3.3 Συσταδοποίηση

Σύμφωνα με τους παραπάνω ορισμούς, έγινε προσπάθεια συσταδοποίησης των εγγράφων του dataset σύμφωνα με την απόσταση με βάση τις ετικέτες με τις οποίες είναι επισημειωμένα.

Για τη συσταδοποίηση δοκιμάστηκαν οι αλγόριθμοι Hierarchical Clustering και DBScan.

Συνδυάζοντας τους εναλλακτικούς ορισμούς απόστασης ετικετών και απόστασης εγγράφων με βάση τις ετικέτες τους, προκύπτουν 6 διαφορετικές μέθοδοι συσταδοποίησης των εγγράφων με βάση τις ετικέτες του Συστήματος Ταξινόμησης της ACM.

Ο Πίνακας 5.1 παρουσιάζει τους αλγορίθμους clustering που δοκιμάστηκαν σε κάθε μία από τις 6 εναλλακτικές μεθόδους συσταδοποίησης.

	Ελάχιστη απόσταση ετικετών	**Μέγιστη απόσταση ετικετών**	**Μέση απόσταση ετικετών**
Απλή απόσταση ετικετών	Hierarchical clustering	Hierarchical clustering	Hierarchical clustering, DBScan
Απόσταση ετικετών με προσαρμοσμένα βάρη	Hierarchical clustering, DBScan	Hierarchical clustering	Hierarchical clustering, DBScan

Πίνακας 5.1: Αλγόριθμοι συσταδοποίησης με βάση τις ετικέτες του Συστήματος Ταξινόμησης της ACM

5.4 Εμπλουτισμός επισημειώσεων και νέα συσταδοποίηση

Σε αυτό το στάδιο του πειράματος έγινε εμπλουτισμός των ετικετών του Συστήματος Ταξινόμησης της ACM με τις οποίες είναι επισημειωμένα τα έγγραφα του dataset και στη συνέχεια έγινε εκ νέου συσταδοποίηση με όλους τους τρόπους που περιγράφονται στην παράγραφο 5.3.3.

Για τον εμπλουτισμό των ετικετών χρησιμοποιήθηκαν οι ελεύθερες λέξεις – κλειδιά που έχουν προστεθεί από τους συγγραφείς στα κείμενα του dataset. Για κάθε μία από τις ελεύθερες λέξεις – κλειδιά ενός εγγράφου, έγινε προσπάθεια αντιστοίχισης σε μία ετικέτα από το Σύστημα Ταξινόμησης της ACM.

5.4.1 Αντιστοίχιση ελεύθερων λέξεων – κλειδιών σε ετικέτες του Συστήματος Ταξινόμησης ACM

1. Αρχικά υπολογίζεται η απόσταση Levenshtein [24] της υποψήφιας ετικέτας t_c από τους τίτλους όλων των ετικετών του Συστήματος Ταξινόμησης της ACM, συμπεριλαμβανομένων των προτεινόμενων εναλλακτικών τίτλων.
2. Στη συνέχεια επιλέγεται η ετικέτα t_{orig} του Συστήματος Ταξινόμησης της ACM με την ελάχιστη απόσταση Levenshtein από την υποψήφια ετικέτα.
3. Τέλος, υπολογίζεται το ποσοστιαίο σφάλμα της υποψήφιας ετικέτας από την ετικέτα του Συστήματος Ταξινόμησης της ACM:

$$error = \frac{Levenshtein(\mathrm{t_c}, \mathrm{t_{orig}})}{\frac{length(t_c) + length(t_{orig})}{2}}$$

4. Αν το ποσοστιαίο σφάλμα είναι μικρότερο από 15% και εφόσον το έγγραφο δεν είναι ήδη επισημειωμένο με την ετικέτα t_{orig}, τότε η ετικέτα t_{orig} προστίθεται στο έγγραφο.

Με αυτή τη διαδικασία, τα έγγραφα του dataset επισημειώθηκαν με 1446 νέες ετικέτες από το Σύστημα Ταξινόμησης της ACM, αυξάνοντας το συνολικό αριθμό των ετικετών από 4435 σε 5881 (αύξηση 32,6%).

5.4.2 Συσταδοποίηση

Στο τελευταίο μέρος του πειράματος έγινε εκ νέου συσταδοποίηση των εγγράφων του dataset, χρησιμοποιώντας το σύνολο των εμπλουτισμένων ετικετών του Συστήματος Ταξινόμησης της ACM με τις οποίες επισημειώθηκαν τα έγγραφα.

Αρχικά έγινε συσταδοποίηση με εφαρμογή αλγορίθμου Hierarchical Clustering, με όλες τις παραλλαγές των ορισμών απόστασης ετικετών και απόστασης εγγράφων, όπως περιγράφονται στην παράγραφο 5.3.3.

Στη συνέχεια έγινε προσπάθεια εφαρμογής του αλγορίθμου DBScan. Ωστόσο, δεν ήταν δυνατό να βρεθεί κατάλληλη τιμή της παραμέτρου ε που να οδηγεί σε συσταδοποίηση ικανοποιητικής ποιότητας, καθώς προέκυψε το ίδιο πρόβλημα που περιγράφεται στην παράγραφο 5.1.2.

6 *Αποτελέσματα*

Σκοπός αυτής της ενότητας είναι η αξιολόγηση των αποτελεσμάτων του πειράματος. Για τους σκοπούς της παρούσας εργασίας θεωρήθηκε ότι η σωστότερη ή/και λεπτομερέστερη επισημείωση ενός συνόλου εγγράφων οδηγεί σε ακριβέστερη συσταδοποίηση των εγγράφων αυτών, δηλαδή σε δημιουργία συστάδων υψηλότερης νοηματικής συνάφειας.

Η αρχική συσταδοποίηση με βάση τις λέξεις-κλειδιά λήφθηκε ως ground truth ή συσταδοποίηση αναφοράς. Στη συνέχεια, οι εναλλακτικές συσταδοποιήσεις με βάση τις επισημειώσεις συγκρίθηκαν με την αρχική συσταδοποίηση. Υψηλότερος βαθμός ομοιότητας με την αρχική συσταδοποίηση συνεπάγεται καλύτερη επισημείωση των εγγράφων.

6.1 *Μέτρο σύγκρισης συσταδοποιήσεων*

Ως μέτρο σύγκρισης δύο συσταδοποιήσεων του ίδιου συνόλου δεδομένων χρησιμοποιήθηκε ο δείκτης Rand Index [25].

6.1.1 Rand Index

Ο δείκτης Rand Index είναι ένα μέτρο σύγκρισης δύο συσταδοποιήσεων [26].

Έστω ένα σύνολο n στοιχείων $S = \{o_1, \dots, o_n\}$ και δύο διαμερίσεις του συνόλου προς σύγκριση: $X = \{X_1, \dots, X_r\}$ μία διαμέριση του S σε r υποσύνολα και $Y = \{Y_1, \dots, Y_s\}$ μία διαμέριση του S σε s υποσύνολα. Επίσης, ορίζουμε:

- a το πλήθος των ζευγαριών των στοιχείων του S που βρίσκονται στο ίδιο υποσύνολο τόσο στο X όσο και στο Y
- b το πλήθος των ζευγαριών των στοιχείων του S που βρίσκονται σε διαφορετικό υποσύνολο τόσο στο X όσο και στο Y
- c το πλήθος των ζευγαριών των στοιχείων του S που βρίσκονται στο ίδιο υποσύνολο στο X αλλά σε διαφορετικό υποσύνολο στο Y

- d το πλήθος των ζευγαριών των στοιχείων του S που βρίσκονται σε διαφορετικό υποσύνολο στο X αλλά σε ίδιο υποσύνολο στο Y

Τότε ο δείκτης Rand Index ορίζεται ως:

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Εναλλακτικά, το άθροισμα $a+b$ μπορεί να θεωρηθεί το πλήθος των συμφωνιών μεταξύ των X και Y ενώ το άθροισμα $c+d$ το πλήθος των διαφωνιών μεταξύ των δύο. Καθώς ο παρονομαστής είναι το πλήθος όλων των πιθανών ζευγαριών του S, ο δείκτης Rand Index εκφράζει την πιθανότητα τα X και Y να συμφωνούν σε ένα τυχαία επιλεγμένο ζευγάρι των στοιχείων του S.

6.2 Αξιολόγηση ελεύθερων λέξεων – κλειδιών

Όπως αναφέρθηκε, δεν ήταν δυνατόν να δημιουργηθεί συσταδοποίηση των εγγράφων του dataset με βάση τις ελεύθερες λέξεις – κλειδιά που επέλεξαν οι συγγραφείς. Επομένως, δεν ήταν δυνατή η αξιολόγηση της επισήμανσης των εγγράφων με ελεύθερες λέξεις – κλειδιά.

6.3 Αξιολόγηση ετικετών μέσω ιεραρχικής συσταδοποίησης

Μετά την ολοκλήρωση της ιεραρχικής συσταδοποίησης με βάση τις αρχικές ετικέτες του Συστήματος Ταξινόμησης της ACM, πραγματοποιήθηκε σύγκριση των συσταδοποιήσεων με τις αρχικές συσταδοποιήσεις αναφοράς που δημιουργήθηκαν με βάση το καθαρό κείμενο [27].

Έγινε ιεραρχική συσταδοποίηση χρησιμοποιώντας τους έξι διαφορετικούς ορισμούς απόστασης εγγράφων, όπως τους περιγράφει ο Πίνακας 5.1 της παραγράφου 5.3. Όπως και στη συσταδοποίηση καθαρού κειμένου, για κάθε διαφορετικό ορισμό απόστασης εγγράφων δημιουργήθηκαν 240 συσταδοποιήσεις, με τον αριθμό των συστάδων να ποικίλλει από 11 έως 250 συστάδες.

Στη συνέχεια, κάθε συσταδοποίηση με βάση τις ετικέτες συγκρίθηκε με την αντίστοιχη (με τον ίδιο αριθμό συστάδων) συσταδοποίηση καθαρού κειμένου.

Στη συνέχεια, πραγματοποιήθηκε εμπλουτισμός των ετικετών, όπως περιγράφεται στην παράγραφο 5.4.1, και έγινε εκ νέου συσταδοποίηση, ακολουθώντας και πάλι όλους τους διαφορετικούς ορισμούς απόστασης εγγράφων της παραγράφου 5.3.

Τέλος, έγινε σύγκριση όλων των νέων συσταδοποιήσεων με τις αντίστοιχες (με τον ίδιο αριθμό ετικετών) συσταδοποιήσεις αναφοράς.

Σε αυτό το σημείο ήταν δυνατόν να ελεγχθεί κατά πόσο βελτιώθηκε η ομοιότητα κάθε συσταδοποίησης μέσω του εμπλουτισμού των ετικετών. Στα παρακάτω σχήματα (Σχήμα 6.1 έως Σχήμα 6.6) παρουσιάζεται η τιμή του δείκτη Rand Index ως προς τον αριθμό των συστάδων για τους διαφορετικούς ορισμούς απόστασης εγγράφων που δοκιμάστηκαν.

Ακολουθεί σύντομος σχολιασμός για την απόδοση κάθε ορισμού απόστασης εγγράφων ως μέτρο συσταδοποίησης με σκοπό την αξιολόγηση της ποιότητας των επισημειώσεων των εγγράφων.

6.3.1 Απλή απόσταση ετικετών

Με τη χρήση αυτού του ορισμού απόστασης εγγράφων, η τιμή του δείκτη Rand Index μετά τον εμπλουτισμό των ετικετών παρουσίασε μείωση, αντί να παρουσιάσει αύξηση, όπως αναμενόταν (Σχήμα 6.1). Επομένως, αυτός ο τρόπος συσταδοποίησης κρίνεται ακατάλληλος για τους στόχους του παρόντος πειράματος.

6.3.2 Μέγιστη απόσταση ετικετών

Η τιμή του δείκτη Rand Index σε αυτήν την περίπτωση παρουσιάζει σημαντική αύξηση μετά τον εμπλουτισμό των ετικετών, συγκεκριμένα από 6% έως και περίπου 30%, για όλες σχεδόν τις συσταδοποιήσεις που περιέχουν τουλάχιστον 50 συστάδες (Σχήμα 6.2). Δεδομένου ότι η συσταδοποίηση σε μικρότερο αριθμό συστάδων δεν έχει νόημα για το μέγεθος του συγκεκριμένου συνόλου δεδομένων (1757 έγγραφα), ο ορισμός αυτός κρίνεται κατάλληλος για συσταδοποίηση με στόχο την αξιολόγηση των επισημειώσεων των εγγράφων.

6.3.3 Μέση απόσταση ετικετών

Σε αυτήν την περίπτωση, ο δείκτης Rand Index είναι άλλοτε μικρότερος και άλλοτε μεγαλύτερος μετά τον εμπλουτισμό των ετικετών (Σχήμα 6.3), ενώ δεν ήταν δυνατόν να προσδιορισθεί κάποιο μοτίβο που να ερμηνεύει αυτήν τη συμπεριφορά. Επομένως, η μέση απόσταση ετικετών κρίνεται ακατάλληλος ορισμός απόστασης για το παρόν πείραμα.

6.3.4 Απλή απόσταση ετικετών με προσαρμοσμένα βάρη

Όπως συνέβη και με την απλή απόσταση ετικετών στην παράγραφο 6.3.1, έτσι και στην απλή απόσταση ετικετών με προσαρμοσμένα βάρη ο δείκτης Rand Index παρουσίασε μείωση αντί για αύξηση (Σχήμα 6.4). Συνεπώς, η απλή απόσταση ετικετών με προσαρμοσμένα βάρη είναι ένας από τους ορισμούς απόστασης εγγράφων που απορρίπτονται.

6.3.5 *Μέγιστη απόσταση ετικετών με προσαρμοσμένα βάρη*

Σε αντίθεση με την μέγιστη απόσταση ετικετών (παράγραφος 6.3.2), η μέγιστη απόσταση ετικετών με προσαρμοσμένα βάρη δε στάθηκε ικανή να διακρίνει τις εμπλουτισμένες ετικέτες από τις αρχικές, καθώς ο δείκτης Rand Index παρουσιάζει άλλοτε άνοδο και άλλοτε πτώση (Σχήμα 6.5).

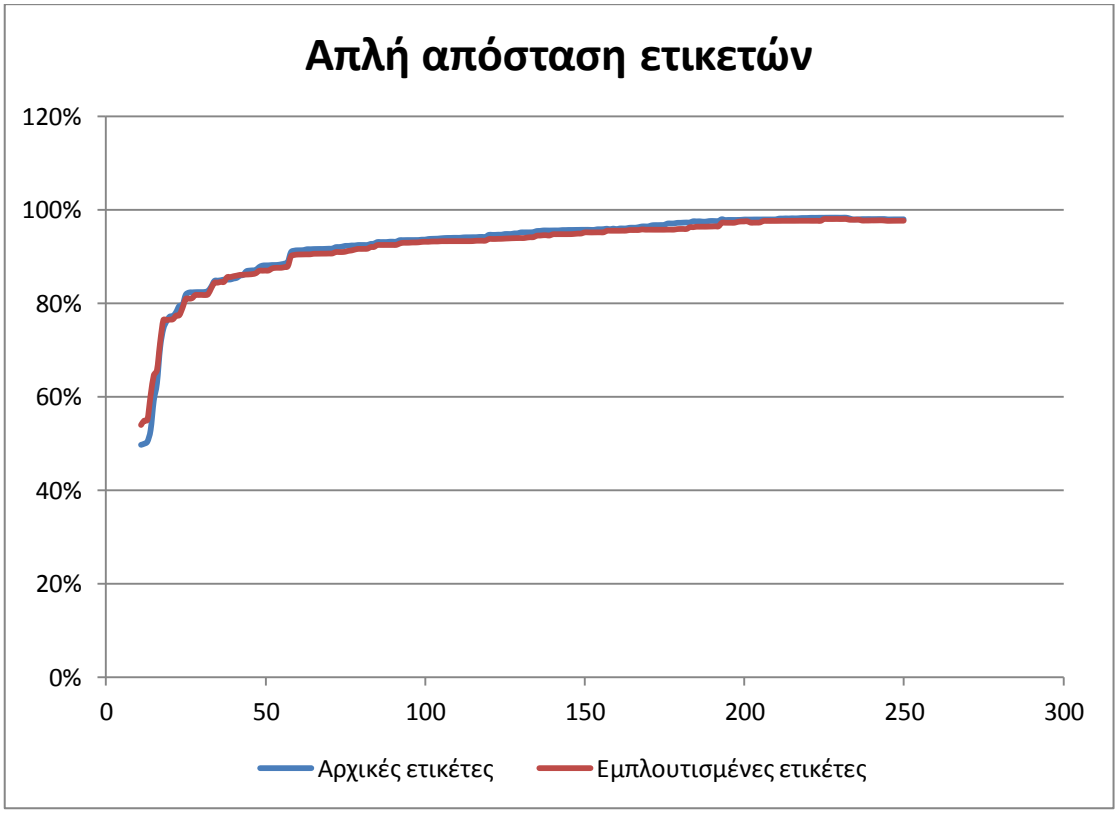
6.3.6 *Μέση απόσταση ετικετών με προσαρμοσμένα βάρη*

Χρησιμοποιώντας τη μέση απόσταση ετικετών με προσαρμοσμένα βάρη, ο δείκτης Rand Index παρουσιάζει αύξηση της τάξεως του 1% σε όλες τις συσταδοποιήσεις με αριθμό συστάδων μεγαλύτερο από 50 (Σχήμα 6.6). Επομένως, πρόκειται για έναν επιτυχημένο ορισμό απόστασης εγγράφων για τους σκοπούς του παρόντος πειράματος.

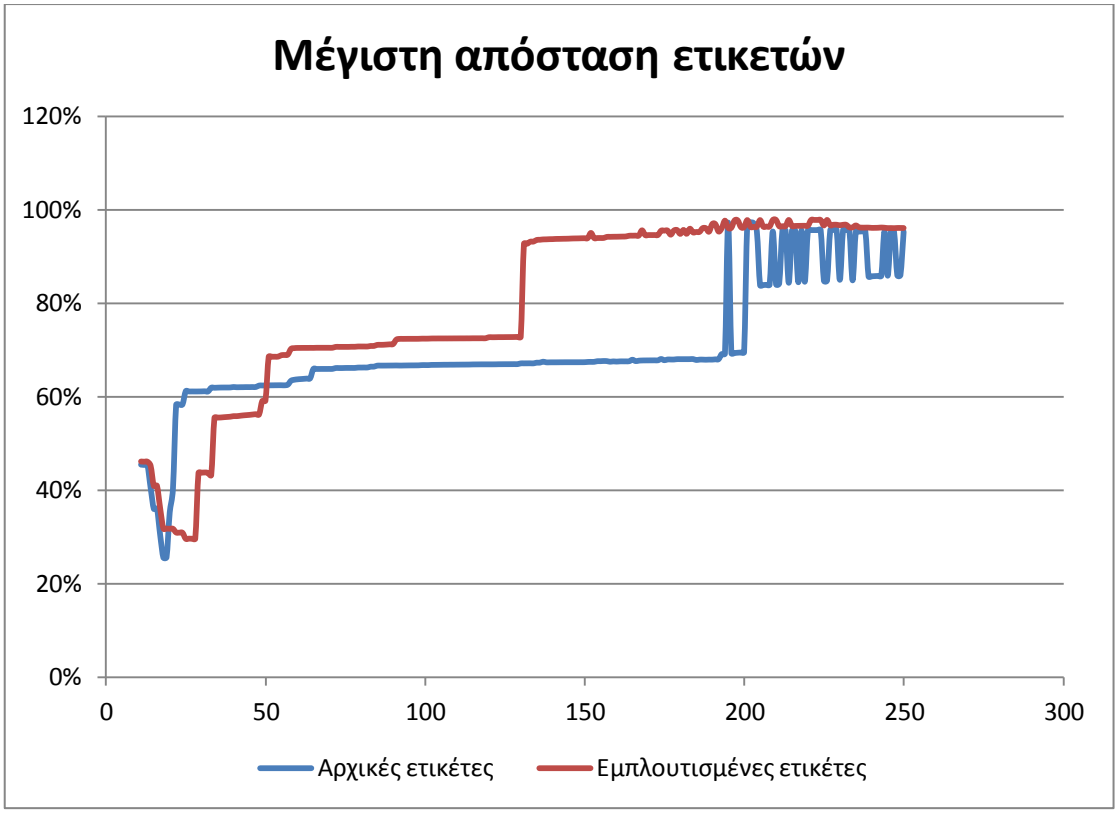
6.3.7 *Σύνοψη*

Οι ορισμοί απόστασης εγγράφων που στάθηκαν ικανοί να διακρίνουν τις εμπλουτισμένες ετικέτες από τις αρχικές ήταν οι ακόλουθοι:

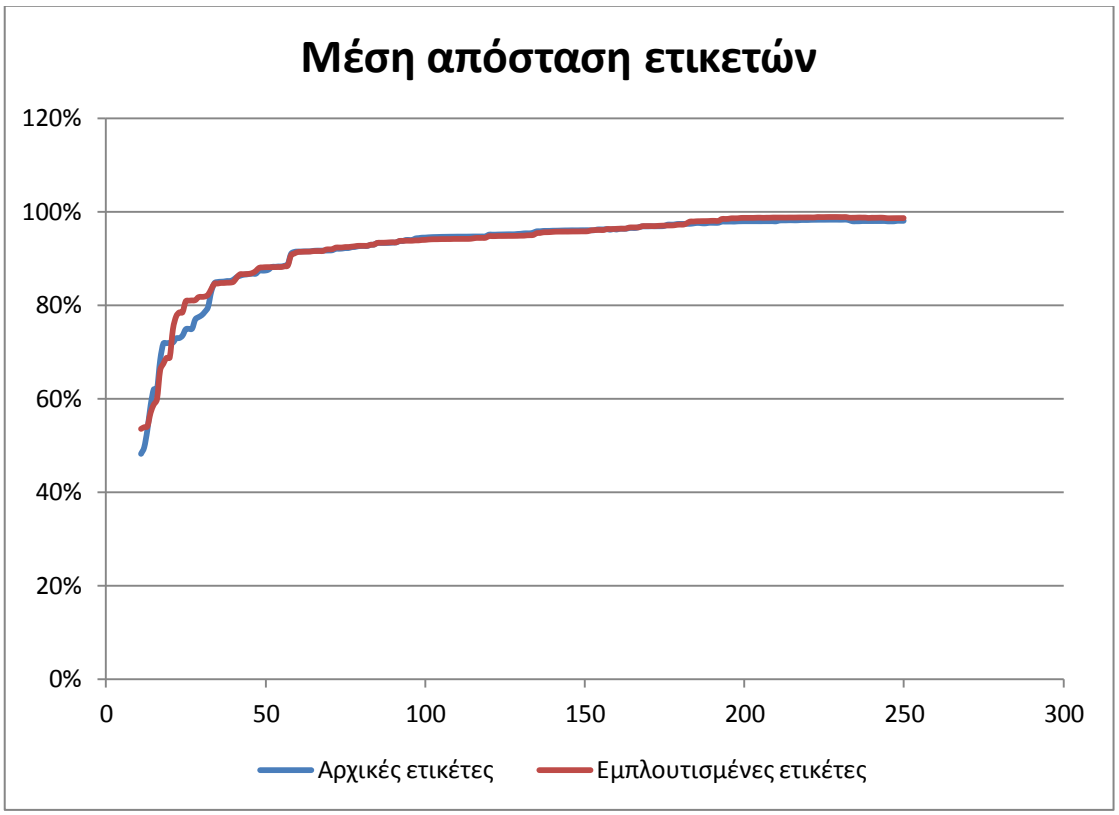
- Μέγιστη απόσταση ετικετών
- Μέση απόσταση ετικετών με προσαρμοσμένα βάρη



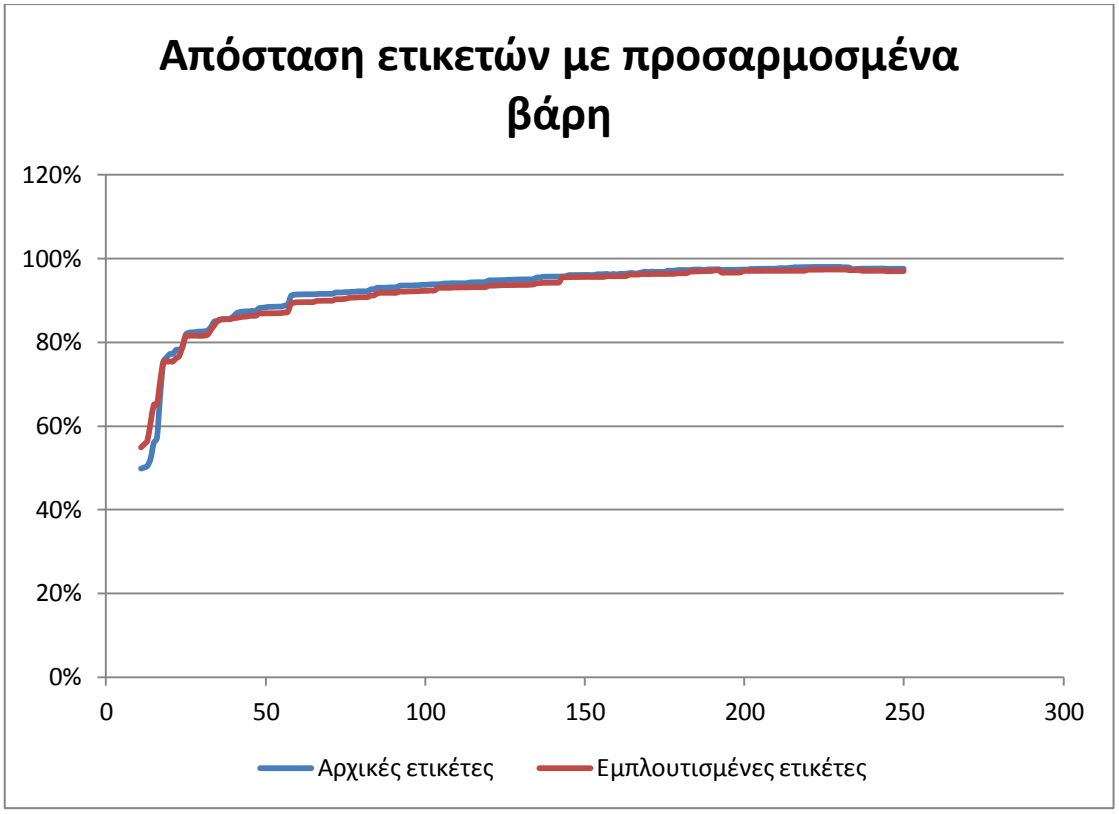
Σχήμα 6.1: Σύγκριση συσταδοποιήσεων με απλή απόσταση ετικετών



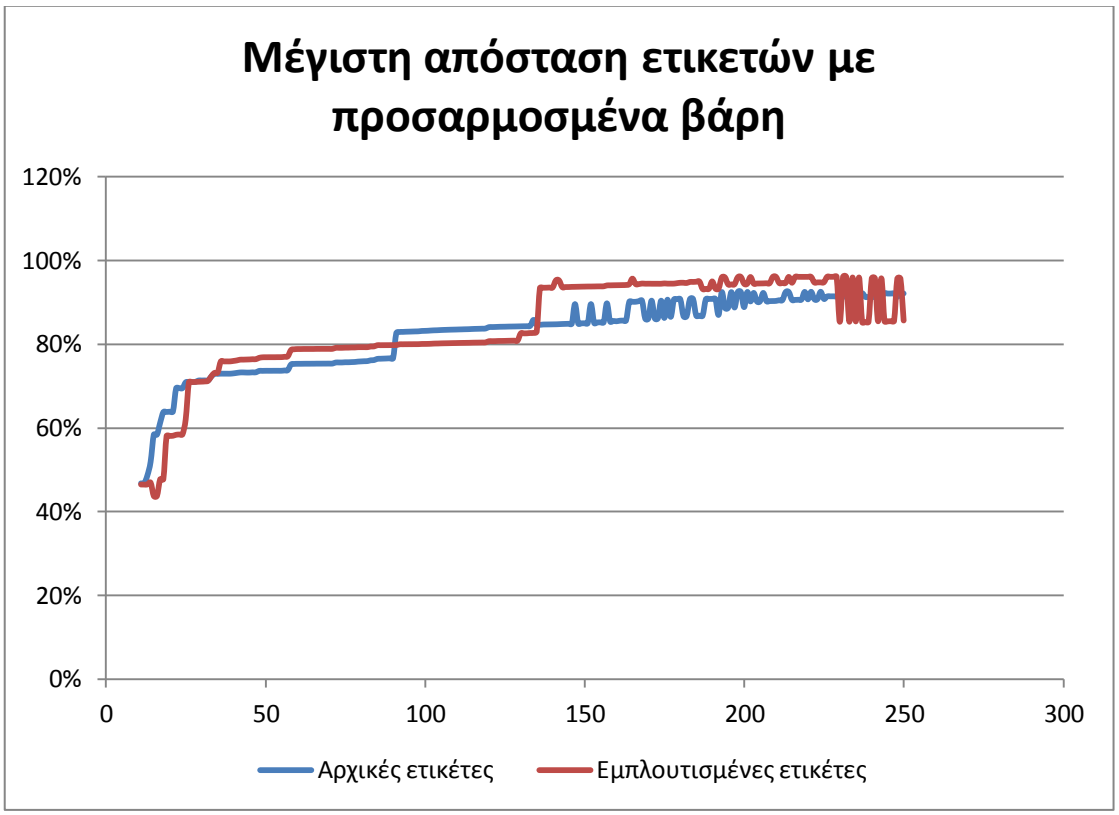
Σχήμα 6.2: Σύγκριση συσταδοποιήσεων με μέγιστη απόσταση ετικετών



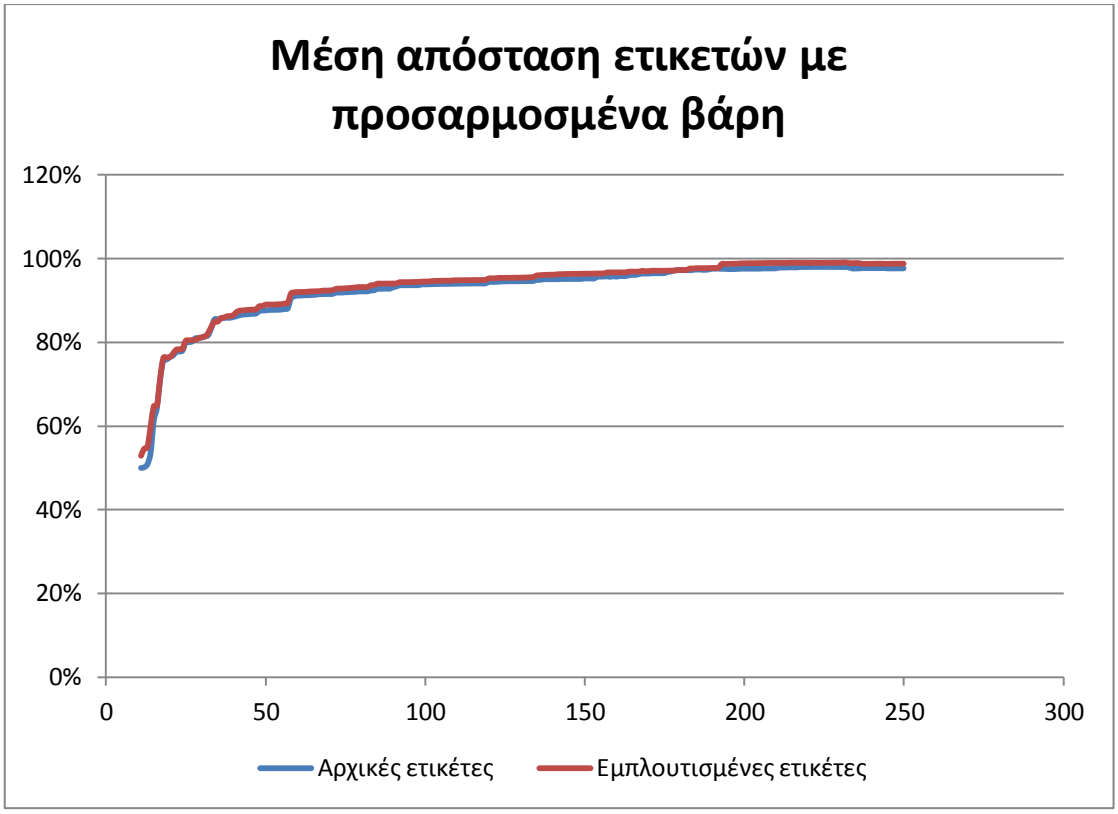
Σχήμα 6.3: Σύγκριση συσταδοποιήσεων με μέση απόσταση ετικετών



Σχήμα 6.4: Σύγκριση συσταδοποιήσεων με απόσταση ετικετών με προσαρμοσμένα βάρη



Σχήμα 6.5: Σύγκριση συσταδοποιήσεων με μέγιστη απόσταση ετικετών με προσαρμοσμένα βάρη



Σχήμα 6.6: Σύγκριση συσταδοποιήσεων με μέση απόσταση ετικετών με προσαρμοσμένα βάρη

6.4 Αξιολόγηση ετικετών μέσω συσταδοποίησης πυκνότητας

Όπως φαίνεται στις παραγράφους 5.1.2 και 5.4.2, δεν ήταν δυνατόν να πραγματοποιηθεί συσταδοποίηση πυκνότητας σε όλα τα στάδια του πειράματος με τρόπο ώστε να προκύπτουν κατάλληλες συστάδες. Επομένως, οι επισημειώσεις των εγγράφων δεν ήταν δυνατόν να αξιολογηθούν μέσω συσταδοποίησης πυκνότητας.

Ωστόσο, αξίζει να γίνει αναφορά στα αποτελέσματα της συσταδοποίησης πυκνότητας που έγινε με τις αρχικές, μη εμπλουτισμένες ετικέτες των εγγράφων.

Και σε αυτήν την περίπτωση έγινε συσταδοποίηση χρησιμοποιώντας τρεις διαφορετικούς ορισμούς απόστασης εγγράφων, όπως δείχνει ο Πίνακας 5.1. Για κάθε ορισμό απόστασης, έγινε δοκιμή συσταδοποίησης με διάφορες τιμές της παραμέτρου $MinPts$ και για κάθε ορισμό επιλέχθηκε τελικά η τιμή της παραμέτρου με τα καλύτερα αποτελέσματα.

Επιλέγοντας τη βέλτιστη τιμή της παραμέτρου $MinPts$ προέκυψαν συσταδοποιήσεις με περίπου 80 συστάδες ισορροπημένου μεγέθους και υψηλής νοηματικής συνάφειας.

7 Τεχνικές λεπτομέρειες

Σε αυτό το κεφάλαιο περιγράφονται οι βιβλιοθήκες που χρησιμοποιήθηκαν για την υλοποίηση του πειράματος, καθώς και άλλες αξιοσημείωτες τεχνικές λεπτομέρειες ή τρόποι επίλυσης ζητημάτων που προέκυψαν κατά τη διαδικασία του πειράματος.

7.1 SQLite

Για την καταγραφή των μεταδεδομένων των εγγράφων του dataset που δημιουργήθηκε, ήταν ανάγκη να κατασκευαστεί μία βάση δεδομένων. Η τεχνολογία που επιλέχθηκε ήταν η SQLite [28], καθώς είναι εύκολη στη χρήση ανεξαρτήτως λειτουργικού συστήματος, και επίσης διαθέτει μία έτοιμη και εύκολη στη χρήση βιβλιοθήκη αλληλεπίδρασης με τη γλώσσα προγραμματισμού Java, στην οποία είναι γραμμένος όλος ο κώδικας του πειράματος.

Για την αποθήκευση των μεταδεδομένων των εγγράφων δημιουργήθηκε στη βάση δεδομένων ένας πίνακας με το όνομα *papers*, αποτελούμενος από 31 πεδία, μεταξύ των οποίων τα κυριότερα είναι τα εξής:

- το **id** του εγγράφου στη βιβλιοθήκη της ACM
- ο συγγραφέας (**author**) ή οι συγγραφείς του εγγράφου
- ο τίτλος (**title**)
- ο αριθμός των σελίδων (**num_pages**)
- οι ελεύθερες λέξεις – κλειδιά που επέλεξαν οι συγγραφείς (**keywords**)
- το έτος δημοσίευσης (**year**)
- οι ετικέτες από το σύστημα ταξινόμησης της ACM (**acm_index_terms**)
- εάν το έγγραφο ήταν δυνατό να μετατραπεί σε καθαρό κείμενο (**plain_text**)
- εάν το έγγραφο επιλέχθηκε για τη δημιουργία του τελικού dataset (**clean_text**)
- οι επιπλέον ετικέτες από το σύστημα ταξινόμησης της ACM (**acm_additional_terms**) με τις οποίες έγινε ο εμπλουτισμός των αρχικών ετικετών

Η βάση δεδομένων που κατασκευάστηκε, εκτός από την αποθήκευση των μεταδεδομένων των εγγράφων, χρησίμευσε επίσης ως βοηθητικό μέσο στατιστικής ανάλυσης σε διάφορα στάδια του πειράματος. Συγκεκριμένα:

- Δημιουργήθηκε ο πίνακας *keywords* {**id**, **keyword**, **usage_count**) στον οποίο καταχωρήθηκαν όλες οι διαφορετικές ελεύθερες λέξεις – κλειδιά που επέλεξαν οι συγγραφείς των εγγράφων, καθώς και το πλήθος χρήσεων της κάθε λέξης – κλειδιού.
- Δημιουργήθηκε ο πίνακας *acm_tags* {**id**, **usage_count**} στον οποίο καταχωρήθηκαν όλες οι διαφορετικές ετικέτες από το σύστημα ταξινόμησης της ACM με τις οποίες ήταν επισημειωμένα τα έγγραφα, καθώς και το πλήθος χρήσεων της κάθε ετικέτας.
- Δημιουργήθηκε ο πίνακας *clustering_scores* {**file1**, **file2**, **score**} στον οποίο καταχωρήθηκε η τιμή του δείκτη rand_index που προέκυψε κατά τη σύγκριση όλων των συσταδοποιήσεων που προέκυψαν με dbscan με όλες τις υπόλοιπες συσταδοποιήσεις.

7.2 Tika

Το Tika [29] είναι μία βιβλιοθήκη ανοικτού κώδικα γραμμένη σε Java. Διατίθεται από τον οργανισμό Apache. Στην παρούσα εργασία, χρησιμοποιήθηκε για την εξαγωγή των μεταδεδομένων αλλά και του καθαρού κειμένου από τα έγγραφα του dataset.

Ως είσοδος στη βιβλιοθήκη δόθηκαν τα αρχεία PDF του dataset, ένα για κάθε έγγραφο – δημοσίευση.

Η βιβλιοθήκη έδωσε ως έξοδο για κάθε αρχείο τα αντίστοιχα μεταδεδομένα, τα οποία αποθηκεύθηκαν στη βάση δεδομένων που περιγράφεται στην παράγραφο 7.1, καθώς και το καθαρό κείμενο του εγγράφου, το οποίο στη συνέχεια αποθηκεύθηκε ως TXT αρχείο.

7.3 LingPipe

Το LingPipe [30] είναι μία βιβλιοθήκη Java ανοικτού κώδικα του οργανισμού alias-i. Η βιβλιοθήκη αυτή υλοποιεί, μεταξύ άλλων, αλγορίθμους ιεραρχικής συσταδοποίησης. Αυτός ήταν και ο λόγος που χρησιμοποιήθηκε στην παρούσα εργασία.

Επεκτείνοντας την αφηρημένη κλάση *Distance* ήταν δυνατόν να κατασκευασθούν νέες κλάσεις που υλοποιούν τους διάφορους ορισμούς απόστασης εγγράφων, όπως περιγράφονται στις παραγράφους 5.1 και 5.3.2.

Στη συνέχεια έγινε ιεραρχική συσταδοποίηση χρησιμοποιώντας την κλάση *HierarchicalClusterer*.

```
▼<clustering count="N">
  ▼<cluster count="M" id="I">
      <Paper id="ID" title="TITLE" year="YEAR"/>
      ...
    </cluster>
  </clustering>
```

Σχήμα 7.1: Δομή αποθήκευσης αποτελεσμάτων συσταδοποίησης

7.4 psjava

Η βιβλιοθήκη psjava παρέχει υλοποίηση διαφόρων αλγορίθμων για γράφους σε Java.

Στην παρούσα εργασία, χρησιμοποιήθηκε ο αλγόριθμος ελάχιστου μονοπατιού, με σκοπό την εύρεση της απόστασης μεταξύ δύο ετικετών του συστήματος ταξινόμησης της ACM, όπως περιγράφεται στην παράγραφο 5.3.1.

7.5 Apache Commons: Math

Πρόκειται για άλλη μία βιβλιοθήκη ανοικτού κώδικα που παρέχεται από την Apache, η οποία υλοποιεί διάφορους μαθηματικούς αλγορίθμους σε Java [31].

Στην παρούσα εργασία, χρησιμοποιήθηκε η υλοποίηση του αλγορίθμου DBScan για την πραγματοποίηση συσταδοποίησης πυκνότητας, όπως περιγράφεται στις παραγράφους 5.1.2, 5.3.3 και 5.4.

7.6 Maven

Για τη διαχείριση και την ενσωμάτωση όλων των παραπάνω βιβλιοθηκών, αλλά και για τη γενικότερη διαχείριση της ανάπτυξης του κώδικα που δημιουργήθηκε για το πείραμα, χρησιμοποιήθηκε η βιβλιοθήκη Maven της Apache.

Το Apache Maven [32] είναι ένα εργαλείο διαχείρισης project λογισμικού. Το Maven μπορεί να διαχειριστεί τη δημιουργία, την αναφορά και την τεκμηρίωση ενός έργου από μία κεντρική βάση πληροφοριών.

7.7 XML

Η XML (eXtensible Markup Language) [33] είναι μία γλώσσα καταγραφής πληροφοριών. Δημιουργήθηκε για την δομημένη αποθήκευση και τη μεταφορά πληροφοριών, με στόχο να μπορεί να γίνει κατανοητή τόσο από ανθρώπους όσο και από μηχανές.

Στην παρούσα εργασία, η XML χρησιμοποιήθηκε σε 2 διαφορετικές πτυχές του πειράματος:

- Οι ετικέτες του Συστήματος Ταξινόμησης της ACM είναι καταγεγραμμένες σε μορφή SKOS, που αποτελεί μία επέκταση της XML. Έτσι, χρειάστηκαν τεχνικές διαχείρισης XML για την ανάγνωση των ετικετών και των μεταξύ τους σχέσεων από τον κώδικα του πειράματος.
- Τα αποτελέσματα της εκάστοτε συσταδοποίησης αποθηκεύτηκαν σε αρχεία XML με τη δομή που παρουσιάζεται στο Σχήμα 7.1. Στο παράδειγμα του σχήματος, η παράμετρος *count* με τιμή *N* είναι ο αριθμός των συστάδων που δημιουργήθηκαν στη συγκεκριμένη συσταδοποίηση. Η παράμετρος *count* με τιμή *M* είναι ο αριθμός των εγγράφων που περιέχει η συγκεκριμένη συστάδα. Το κάθε έγγραφο *Paper* περιέχει τις εξής τρεις παραμέτρους:
    - **id**: το id του εγγράφου στην Ψηφιακή Βιβλιοθήκη της ACM
    - **title**: ο τίτλος του εγγράφου
    - **year**: το έτος δημοσίευσης του εγγράφου

8 Επίλογος

Στο κεφάλαιο αυτό συνοψίζονται τα αποτελέσματα του πειράματος που έγινε σε αυτή τη διπλωματική εργασία. Επίσης, γίνονται προτάσεις για τη μελλοντική βελτίωση του εργαλείου αξιολόγησης ετικετών που προτείνεται.

8.1 Σύνοψη και συμπεράσματα

Όπως φαίνεται στο κεφάλαιο 6, η συσταδοποίηση των εγγράφων βελτιώθηκε μετά τον εμπλουτισμό των ετικετών. Επομένως, το προτεινόμενο σύστημα αποτελεί έναν αποτελεσματικό τρόπο αξιολόγησης των ετικετών ενός συνόλου εγγράφων.

Η προτεινόμενη διαδικασία αξιολόγησης των ετικετών ενός συνόλου εγγράφων συνοψίζεται στα εξής βήματα:

1. Δημιουργία **συσταδοποίησης** με βάση το **κείμενο** των εγγράφων
2. Δημιουργία **συσταδοποίησης** με βάση τις **ετικέτες** των εγγράφων
3. **Σύγκριση** της συσταδοποίησης ετικετών με τη συσταδοποίηση κειμένου σύμφωνα με κάποια τιμή κατωφλίου
4. **Εμπλουτισμός** ή βελτίωση των ετικετών
5. Εκ νέου **συσταδοποίηση** με βάση τις **ετικέτες**
6. **Σύγκριση** της νέας συσταδοποίησης ετικετών με την αρχική συσταδοποίηση κειμένου σύμφωνα με την προηγούμενη τιμή κατωφλίου αλλά και σύμφωνα με το μέτρο ομοιότητας όπως προέκυψε από τη σύγκριση των συσταδοποιήσεων των βημάτων 1 και 2

8.2 Μελλοντικές επεκτάσεις

Το προτεινόμενο σύστημα αξιολόγησης ετικετών θα μπορούσε ίσως να βελτιωθεί ερευνώντας τα παρακάτω ζητήματα:

- Χρήση εναλλακτικών μεθόδων αναπαράστασης κειμένων στο μοντέλο χώρου διανυσμάτων

- Ορισμός διαφορετικών τρόπων υπολογισμού απόστασης ετικετών
- Χρήση εναλλακτικών αλγορίθμων συσταδοποίησης
- Χρήση εναλλακτικών μέτρων σύγκρισης συσταδοποιήσεων
- Κατασκευή ευφυούς συστήματος εμπλουτισμού ετικετών

9 *Βιβλιογραφία*

1. Stratogiannis, G., G. Siolas, and A. Stafylopatis, *Semantic Question Answering Using Wikipedia Categories Clustering.* International Journal on Artificial Intelligence Tools, 2014. **23**(04): p. 1460014.
2. Agirre, E. and G. Rigau. *Word sense disambiguation using conceptual density*. in *Proceedings of the 16th conference on Computational linguistics-Volume 1*. 1996. Association for Computational Linguistics.
3. Hotho, A., S. Staab, and G. Stumme, *Text clustering based on background knowledge.* Technical report 425, 2003: p. 36.
4. Steinbach, M., G. Karypis, and V. Kumar. *A comparison of document clustering techniques*. in *KDD workshop on text mining*. 2000. Boston.
5. Tan, P.-N., M. Steinbach, and V. Kumar, *Introduction to data mining. 1st*. 2005, Boston: Pearson Addison Wesley. xxi.
6. Rose, S., et al., *Automatic keyword extraction from individual documents.* Text Mining: Applications and Theory, 2010: p. 1-20.
7. Hotho, A., S. Staab, and G. Stumme. *Ontologies improve text document clustering*. in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. 2003. IEEE.
8. Stamou, G. and A. Chortaras. *Ontological Query Answering over Semantic Data*. in *Reasoning Web International Summer School*. 2017. Springer.
9. Pan, J.Z., *Resource description framework*, in *Handbook on ontologies*. 2009, Springer. p. 71-90.
10. Bechhofer, S., *OWL: Web ontology language*, in *Encyclopedia of database systems*. 2009, Springer. p. 2008-2009.
11. Isaac, A. and E. Summers, *SKOS Simple Knowledge Organization System.* Primer, World Wide Web Consortium (W3C), 2009.
12. Johnson, S.C., *Hierarchical clustering schemes.* Psychometrika, 1967. **32**(3): p. 241-254.
13. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. **28**(1): p. 100-108.
14. Kriegel, H.P., et al., *Density-based clustering.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011. **1**(3): p. 231-240.
15. Ester, M., et al. *A density-based algorithm for discovering clusters in large spatial databases with noise*. in *Kdd*. 1996.
16. Schubert, E., et al., *DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN.* ACM Transactions on Database Systems (TODS), 2017. **42**(3): p. 19.
17. Willett, P., *Recent trends in hierarchic document clustering: a critical review.* Information Processing & Management, 1988. **24**(5): p. 577-597.

18. Salton, G., A. Wong, and C.-S. Yang, *A vector space model for automatic indexing.* Communications of the ACM, 1975. **18**(11): p. 613-620.
19. Zhang, Y., R. Jin, and Z.-H. Zhou, *Understanding bag-of-words model: a statistical framework.* International Journal of Machine Learning and Cybernetics, 2010. **1**(1-4): p. 43-52.
20. Zamir, O. and O. Etzioni. *Web document clustering: A feasibility demonstration*. in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998. ACM.
21. Pastor-Sánchez, J.-A., F.J. Martínez Méndez, and J.V. Rodríguez-Muñoz, *Advantages of Thesaurus Representation Using the Simple Knowledge Organization System (SKOS) Compared with Proposed Alternatives.* Information Research: An International Electronic Journal, 2009. **14**(4): p. n4.
22. Ahuja, R.K., et al., *Faster algorithms for the shortest path problem.* Journal of the ACM (JACM), 1990. **37**(2): p. 213-223.
23. Cherkassky, B.V., A.V. Goldberg, and T. Radzik, *Shortest paths algorithms: Theory and experimental evaluation.* Mathematical programming, 1996. **73**(2): p. 129-174.
24. Yujian, L. and L. Bo, *A normalized Levenshtein distance metric.* IEEE transactions on pattern analysis and machine intelligence, 2007. **29**(6): p. 1091-1095.
25. Recasens, M. and E. Hovy, *BLANC: Implementing the Rand index for coreference evaluation.* Natural Language Engineering, 2011. **17**(4): p. 485-510.
26. Rand, W.M., *Objective criteria for the evaluation of clustering methods.* Journal of the American Statistical association, 1971. **66**(336): p. 846-850.
27. Vinh, N.X., J. Epps, and J. Bailey. *Information theoretic measures for clusterings comparison: is a correction for chance necessary?* in *Proceedings of the 26th annual international conference on machine learning*. 2009. ACM.
28. Owens, M. and G. Allen, *SQLite*. 2010: Springer.
29. Mattmann, C. and J. Zitting, *Tika in action*. 2011: Manning Publications Co.
30. Kano, Y., et al., *U-Compare: share and compare text mining tools with UIMA.* Bioinformatics, 2009. **25**(15): p. 1997-1998.
31. Math, C., *The apache commons mathematics library.* Np, nd Web, 2016. **9**.
32. McIntosh, S., B. Adams, and A.E. Hassan, *The evolution of Java build systems.* Empirical Software Engineering, 2012. **17**(4-5): p. 578-608.
33. Bray, T., et al., *Extensible markup language (XML).* World Wide Web Journal, 1997. **2**(4): p. 27-66.