

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΤΟΜΙΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μάιος 2026

Ατομική Διπλωματική Εργασία

**ΒΙΒΛΙΟΜΕΤΡΙΚΗ ΑΝΑΛΥΣΗ ΔΗΜΟΣΙΕΥΣΕΩΝ ΑΝΑΦΟΡΙΚΑ
ΜΕ ΕΠΑΝΑΧΡΗΣΙΜΟΠΟΙΗΣΗ ΛΟΓΙΣΜΙΚΟΥ**

Σοφοκλής Κωνσταντίνου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2026

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Βιβλιομετρική Ανάλυση Δημοσιεύσεων Αναφορικά Με Επαναχρησιμοποίηση

Λογισμικού

Σοφοκλής Κωνσταντίνου

Επιβλέπουσα Καθηγήτρια

Δρ. Γεωργία Καπιτσάκη

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου

Κύπρου

Μάιος 2026

Ευχαριστίες

Η ολοκλήρωση της συγκεκριμένης Ατομικής Διπλωματικής Εργασίας δεν θα ήταν εφικτή χωρίς τη βοήθεια των παρακάτω προσώπων, τα οποία θα ήθελα να ευχαριστήσω.

Αρχικά, θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια της εργασίας μου, Δρ. Γεωργία Καπιτσάκη, για τη συνολική καθοδήγηση κατά την εκπόνηση της εργασίας, καθώς και για την πολύτιμη και άμεση ανατροφοδότηση που μου παρείχε. Η βοήθειά της κατά τη διάρκεια αυτών των δύο εξαμήνων ήταν καθοριστική για το τελικό αποτέλεσμα της εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, οι οποίοι ήταν δίπλα μου καθ' όλη τη διάρκεια του πτυχίου μου και ειδικότερα κατά την ολοκλήρωση της Διπλωματικής Εργασίας.

Περίληψη

Η παρούσα διπλωματική εργασία με τίτλο «Βιβλιομετρική Ανάλυση Δημοσιεύσεων Αναφορικά με την Επαναχρησιμοποίηση Λογισμικού» εξετάζει την επιστημονική βιβλιογραφία του πεδίου, με στόχο την αποτύπωση της εξέλιξης, της δομής και των βασικών ερευνητικών κατευθύνσεων του. Η εργασία βασίζεται σε δημοσιεύσεις που συλλέχθηκαν από τη βιβλιογραφική βάση Scopus και αξιοποιεί βιβλιομετρικές τεχνικές και τεχνικές ανάλυσης κειμένου για τη μελέτη της επιστημονικής παραγωγής.

Η ανάλυση επικεντρώνεται σε διαφορετικές πτυχές της βιβλιογραφίας, όπως η χρονική εξέλιξη των δημοσιεύσεων, οι κύριοι ερευνητικοί συντελεστές, η γεωγραφική κατανομή της έρευνας και οι βασικές θεματικές περιοχές που διαμορφώνουν το πεδίο. Παράλληλα, η χρήση τεχνικών θεματικής ανάλυσης και οπτικοποίησης βιβλιομετρικών δικτύων επιτρέπει την καλύτερη κατανόηση των σχέσεων μεταξύ των κυρίαρχων εννοιών της βιβλιογραφίας.

Τα αποτελέσματα της εργασίας αναδεικνύουν ότι η επαναχρησιμοποίηση λογισμικού αποτελεί ένα πολυδιάστατο και εξελισσόμενο ερευνητικό πεδίο, το οποίο συνδέεται με αρκετές θεματικές ενότητες. Τέλος, στο πλαίσιο της εργασίας αναπτύχθηκε διαδραστική διαδικτυακή εφαρμογή για την παρουσίαση και εξερεύνηση των αποτελεσμάτων της ανάλυσης.

Δηλώνω ότι η παρούσα διατριβή και οποιοδήποτε συνοδευτικό λογισμικό αποτελούν δικό μου πρωτότυπο έργο, εκπονημένο σε πλήρη συμμόρφωση με τον Κώδικα Δεοντολογίας για την Έρευνα του Πανεπιστημίου Κύπρου

(<https://www.ucy.ac.cy/legislation/kodikas-deontologias-kai-politikes/>), και με πλήρη ευθύνη εκ μέρους μου για την ακρίβεια, την ακεραιότητα και την εγκυρότητα όλου του περιεχομένου.

Περιεχόμενα

Κεφάλαιο 1	9
Εισαγωγή	9
1.1 Γενικό Πλαίσιο	9
1.2 Διατύπωση Προβλήματος	9
1.3 Στόχοι Εργασίας	10
1.4 Δομή της Εργασίας	10
Κεφάλαιο 2	12
Θεωρητικό Υπόβαθρο	12
2.1 Επαναχρησιμοποίηση Λογισμικού	12
2.2 Βιβλιομετρική Ανάλυση	12
2.2.1 Μεταδεδομένα Δημοσιεύσεων	13
2.3 Επεξεργασία Φυσικής Γλώσσας και Τεχνικές Εξόρυξης Κειμένου	13
Κεφάλαιο 3	16
Μεθοδολογία	16
3.1 Προκαταρκτική Μελέτη	16
3.2 Συλλογή Δεδομένων	17
3.3 Εργαλεία Ανάλυσης Δεδομένων	19
3.3.1 Visual Studio Code	19
3.3.2 R	20
3.3.3 VOSviewer	20
3.3.4 Python & Streamlit	21
3.4 Καθαρισμός Δεδομένων	21
3.4.1 Αυτοματοποιημένο Φιλτράρισμα	21
3.4.2 Λεμματοποίηση	23
3.4.3 Λέξεις Διακοπής	23

3.5 Τεχνικές Ανάλυσης	24
3.5.1 N-Grams	24
3.5.2 TF-IDF	25
3.5.3 LDA	25
3.6 Ερευνητικές Ερωτήσεις	26
Κεφάλαιο 4	28
Αποτελέσματα	28
4.1 Χρονικές Τάσεις Δημοσιεύσεων	28
4.2 Κορυφαία Περιοδικά και Συνέδρια	30
4.3 Κατανομή Τύπων Δημοσιεύσεων	31
4.4 Κορυφαίες Χώρες και Ιδρύματα	32
4.5 Κορυφαίοι Συγγραφείς	35
4.6 Ανάλυση Λέξεων-Κλειδιών	36
4.6.1 Πιο Συχνοί Όροι (N-Grams)	36
4.6.2 Ανάλυση TF-IDF	40
4.6.3 Χάρτης Συν-Εμφάνισης Λέξεων-Κλειδιών	41
4.7 Αποτελέσματα Μοντελοποίησης Θεμάτων	43
4.7.1 Αναγνωρισμένα Θέματα (LDA)	43
4.7.2 Κατανομή Κυρίαρχων Θεμάτων	45
4.7.3 Εξέλιξη Θεμάτων στον Χρόνο	46
4.8 Εξέλιξη Συνεργασίας Συγγραφέων στον Χρόνο	48
4.9 Διαδραστική Διαδικτυακή Εφαρμογή	49
Κεφάλαιο 5	54
Σχολιασμός Αποτελεσμάτων	54
5.1 Επισκόπηση Ευρημάτων	54
5.2 Απαντήσεις στις Ερευνητικές Ερωτήσεις	56
5.3 Ερμηνεία Αποτελεσμάτων	62
5.3.1 Software Product Lines (SPL)	62
5.3.2 Component-Based Architecture	63
5.3.3 Software Process & Services	64
5.4 Πιθανή Επίδραση των LLMs στην Εξέλιξη της Επαναχρησιμοποίησης Λογισμικού	65

5.5 Περιορισμοί	67
Κεφάλαιο 6	68
Συμπεράσματα	68
6.1 Σύνοψη της Εργασίας	68
6.2 Κύρια Συμπεράσματα	69
6.3 Συνεισφορές της Εργασίας	69
6.4 Μελλοντική Έρευνα	70
Βιβλιογραφία.....	72
Παράρτημα Α.....	74
A1. Λίστα έγκριτων περιοδικών και συνεδρίων της Πληροφορικής.....	74
Παράρτημα Β	78
B1. Πίνακας με πρόσθετες λέξεις διακοπής και χειροκίνητες κανονικοποιήσεις όρων	78

Κεφάλαιο 1

Εισαγωγή

1.1 Γενικό Πλαίσιο	9
1.2 Διατύπωση Προβλήματος	9
1.3 Στόχοι Εργασίας	10
1.4 Δομή της Εργασίας	10

1.1 Γενικό Πλαίσιο

Η επαναχρησιμοποίηση λογισμικού αναφέρεται στη διαδικασία αξιοποίησης ήδη υπάρχοντων στοιχείων λογισμικού, όπως κώδικας, συνιστώσες, σχεδιαστικά πρότυπα και πλαίσια ανάπτυξης, για την κατασκευή νέων συστημάτων αντί για την ανάπτυξή τους από το μηδέν. Σύμφωνα με τον Kueger [1], η επαναχρησιμοποίηση λογισμικού αποτελεί μία από τις βασικές πρακτικές της μηχανικής λογισμικού, καθώς στοχεύει στη συστηματική αξιοποίηση υφιστάμενων τεχνουργημάτων λογισμικού για την ανάπτυξη νέων εφαρμογών. Η πρακτική αυτή συμβάλλει στη μείωση του κόστους ανάπτυξης, στη βελτίωση της ποιότητας και στην επιτάχυνση της παραγωγικής διαδικασίας, καθώς επιτρέπει την αξιοποίηση λύσεων που έχουν ήδη σχεδιαστεί, υλοποιηθεί και δοκιμαστεί. Η ευρεία διάδοση του λογισμικού ανοικτού κώδικα έχει ενισχύσει περαιτέρω τη σημασία της επαναχρησιμοποίησης, προσφέροντας πρόσβαση σε μεγάλο όγκο διαθέσιμου κώδικα και εργαλείων μέσω πλατφορμών που χρησιμοποιούνται καθημερινά από προγραμματιστές και ερευνητές.

1.2 Διατύπωση Προβλήματος

Παρά τη μακρά ιστορία και τη σημασία της επαναχρησιμοποίησης λογισμικού, η σχετική επιστημονική βιβλιογραφία είναι εκτεταμένη, πολυδιάστατη και συνεχώς αυξανόμενη. Αυτό καθιστά δύσκολη τη συνολική κατανόηση των βασικών ερευνητικών τάσεων, των κυρίαρχων θεματικών περιοχών και των σημαντικότερων συντελεστών της ερευνητικής περιοχής. Επιπλέον, δεν είναι πάντα σαφές πώς εξελίσσεται η έρευνα με την πάροδο του χρόνου, ποιες χώρες και ιδρύματα ηγούνται στην περιοχή, και ποιες θεματικές περιοχές αναδύονται ή υποχωρούν. Προηγούμενες βιβλιομετρικές μελέτες στη μηχανική λογισμικού έχουν

προσφέρει σημαντική εικόνα για τη γενικότερη εξέλιξη του κλάδου, όπως η μελέτη των Wong et al. [16], η οποία εξετάζει θεματικές, συγγραφείς και ιδρύματα στη μηχανική λογισμικού για την περίοδο 2013–2020. Ωστόσο, τέτοιες μελέτες επικεντρώνονται συχνά σε ευρύτερες περιοχές της μηχανικής λογισμικού και δεν αποτυπώνουν ειδικά και σε βάθος την εξέλιξη της βιβλιογραφίας της επαναχρησιμοποίησης λογισμικού. Συνεπώς, κρίνεται απαραίτητη η διεξαγωγή μιας σύγχρονης, συστηματικής και ποσοτικής βιβλιομετρικής ανάλυσης που να αποτυπώνει την τρέχουσα κατάσταση της έρευνας στη συγκεκριμένη ερευνητική περιοχή.

1.3 Στόχοι Εργασίας

Η παρούσα διπλωματική εργασία αποσκοπεί στη συστηματική διερεύνηση της επιστημονικής βιβλιογραφίας στον τομέα της επαναχρησιμοποίησης λογισμικού, με στόχο την κατανόηση της δομής και της εξέλιξης της ερευνητικής περιοχής. Πιο συγκεκριμένα, επιδιώκεται η αποτύπωση των βασικών χαρακτηριστικών της ερευνητικής δραστηριότητας, όπως οι κύριες τάσεις δημοσίευσης, οι σημαντικότεροι συντελεστές και τα κυρίαρχα μέσα διάχυσης της γνώσης.

Παράλληλα, η εργασία στοχεύει στην ανάδειξη των θεματικών κατευθύνσεων που διαμορφώνουν το πεδίο της επαναχρησιμοποίησης λογισμικού, καθώς και στη διερεύνηση του τρόπου με τον οποίο αυτές εξελίσσονται με την πάροδο του χρόνου. Μέσα από αυτή τη συνολική προσέγγιση, επιδιώκεται η παροχή μιας ολοκληρωμένης εικόνας της επιστημονικής δραστηριότητας, συμβάλλοντας στην καλύτερη κατανόηση της δυναμικής και των μελλοντικών προοπτικών του πεδίου.

1.4 Δομή της Εργασίας

Η παρούσα εργασία οργανώνεται σε έξι κεφάλαια, καθένα από τα οποία καλύπτει ένα συγκεκριμένο μέρος της ερευνητικής διαδικασίας. Στο Κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο, με έμφαση στις βασικές έννοιες της επαναχρησιμοποίησης λογισμικού και της βιβλιομετρικής ανάλυσης, καθώς και στις τεχνικές επεξεργασίας φυσικής γλώσσας και εξόρυξης κειμένου που αξιοποιούνται για την ανάλυση της βιβλιογραφίας.

Στο Κεφάλαιο 3 περιγράφεται η μεθοδολογία της έρευνας, όπου παρουσιάζονται η διαδικασία συλλογής και επεξεργασίας των δεδομένων, καθώς και οι ερευνητικές ερωτήσεις που καθοδηγούν την ανάλυση. Στο Κεφάλαιο 4 παρουσιάζονται τα αποτελέσματα, τα οποία

περιλαμβάνουν την αποτύπωση της ερευνητικής δραστηριότητας, καθώς και την ανάδειξη των βασικών θεματικών κατευθύνσεων του πεδίου.

Στο Κεφάλαιο 5 πραγματοποιείται η συζήτηση των ευρημάτων, όπου τα αποτελέσματα ερμηνεύονται σε σχέση με τις ερευνητικές ερωτήσεις και το ευρύτερο ερευνητικό πλαίσιο. Τέλος, στο Κεφάλαιο 6 παρουσιάζονται τα συμπεράσματα της εργασίας, αναδεικνύονται οι περιορισμοί της μελέτης και διατυπώνονται προτάσεις για μελλοντική έρευνα στον τομέα της επαναχρησιμοποίησης λογισμικού.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Επαναχρησιμοποίηση Λογισμικού	12
2.2 Βιβλιομετρική Ανάλυση	12
2.2.1 Μεταδεδομένα Δημοσιεύσεων	13
2.3 Επεξεργασία Φυσικής Γλώσσας και Τεχνικές Εξόρυξης Κειμένου	13

2.1 Επαναχρησιμοποίηση Λογισμικού

Η επαναχρησιμοποίηση λογισμικού αποτελεί μία θεμελιώδη έννοια στη μηχανική λογισμικού, η οποία αναφέρεται στη διαδικασία αξιοποίησης υφιστάμενων στοιχείων λογισμικού για την ανάπτυξη νέων συστημάτων. Περιλαμβάνει τη χρήση ήδη αναπτυγμένων τεχνουργημάτων (artefacts), όπως κώδικα, σχεδιαστικών προτύπων και αρχιτεκτονικών στοιχείων, με στόχο τη βελτίωση της παραγωγικότητας και της ποιότητας του λογισμικού.

Η πρακτική αυτή συμβάλλει στη μείωση του κόστους ανάπτυξης, στον περιορισμό των σφαλμάτων και στην επιτάχυνση της διαδικασίας υλοποίησης, καθώς βασίζεται σε δοκιμασμένα και αξιόπιστα στοιχεία. Επιπλέον, η επαναχρησιμοποίηση μπορεί να εφαρμοστεί σε διαφορετικά επίπεδα, από μεμονωμένα συστατικά μέχρι ολοκληρωμένα συστήματα και πλατφόρμες, ενισχύοντας τη συστηματική ανάπτυξη λογισμικού και την επεκτασιμότητα των εφαρμογών.

2.2 Βιβλιομετρική Ανάλυση

Η βιβλιομετρική ανάλυση αποτελεί μια ποσοτική προσέγγιση για τη μελέτη της επιστημονικής βιβλιογραφίας, η οποία βασίζεται στην εφαρμογή στατιστικών και μαθηματικών μεθόδων για την αξιολόγηση της ερευνητικής δραστηριότητας. Η βιβλιομετρία επιτρέπει την ανάλυση μεγάλων όγκων επιστημονικών δεδομένων με στόχο την κατανόηση της δομής, της εξέλιξης και της δυναμικής ενός ερευνητικού πεδίου.

Μέσω της βιβλιομετρικής ανάλυσης καθίσταται δυνατός ο εντοπισμός των σημαντικότερων συντελεστών της επιστημονικής παραγωγής, όπως συγγραφείς, ιδρύματα και πηγές δημοσίευσης, καθώς και η αξιολόγηση της επιρροής συγκεκριμένων εργασιών. Επιπλέον, τεχνικές όπως η ανάλυση συν-εμφάνισης και η χαρτογράφηση δικτύων γνώσης συμβάλλουν στην αποκάλυψη των σχέσεων μεταξύ εννοιών και στη διερεύνηση των θεματικών κατευθύνσεων ενός πεδίου.

Η βιβλιομετρική προσέγγιση χρησιμοποιείται ευρέως για την αποτύπωση της εξέλιξης της επιστημονικής γνώσης, καθώς επιτρέπει τη συστηματική και αναπαραγώγιμη ανάλυση της βιβλιογραφίας, αναδεικνύοντας τόσο καθιερωμένα όσο και αναδυόμενα ερευνητικά θέματα.

2.2.1 Μεταδεδομένα Δημοσιεύσεων

Ο όρος μεταδεδομένα (metadata) αναφέρεται σε δεδομένα που περιγράφουν άλλα δεδομένα —με άλλα λόγια, είναι πληροφορίες που χαρακτηρίζουν και περιγράφουν το περιεχόμενο ενός αρχείου, εγγράφου ή δημοσίευσης χωρίς να αποτελούν το ίδιο το περιεχόμενο. Στο πλαίσιο της επιστημονικής βιβλιογραφίας, τα μεταδεδομένα μιας δημοσίευσης περιλαμβάνουν πληροφορίες όπως ο τίτλος, οι συγγραφείς, η περίληψη, το έτος δημοσίευσης, το περιοδικό ή συνέδριο δημοσίευσης, οι λέξεις-κλειδιά, τα ιδρύματα των συγγραφέων και ο αριθμός αναφορών.

Στη βιβλιομετρική ανάλυση, τα μεταδεδομένα αποτελούν την κύρια πηγή πληροφοριών, καθώς επιτρέπουν την ποσοτική ανάλυση της επιστημονικής παραγωγής χωρίς να απαιτείται η ανάγνωση του πλήρους κειμένου κάθε δημοσίευσης. Μέσω της ανάλυσης μεταδεδομένων είναι δυνατός ο εντοπισμός των πιο παραγωγικών συγγραφέων και ιδρυμάτων, η χαρτογράφηση της γεωγραφικής κατανομής της έρευνας, η ανάλυση των χρονικών τάσεων δημοσίευσης και η αναγνώριση των πιο σημαντικών πηγών δημοσίευσης. Στην παρούσα εργασία, τα μεταδεδομένα που συλλέχθηκαν από τη βάση Scopus για κάθε δημοσίευση περιλαμβάνουν τον τίτλο, τους συγγραφείς, την περίληψη, το έτος δημοσίευσης, τον τίτλο πηγής, τον τύπο δημοσίευσης, τα ιδρύματα των συγγραφέων, τις λέξεις-κλειδιά των συγγραφέων, τις ευρετηριασμένες λέξεις-κλειδιά και ο κωδικός (DOI).

2.3 Επεξεργασία Φυσικής Γλώσσας και Τεχνικές Εξόρυξης Κειμένου

Η Επεξεργασία Φυσικής Γλώσσας (NLP-Natural Language Processing) είναι ένας κλάδος της τεχνητής νοημοσύνης που ασχολείται με την αλληλεπίδραση μεταξύ υπολογιστών και

ανθρώπινης γλώσσας, με στόχο την αυτοματοποιημένη κατανόηση, ανάλυση και επεξεργασία φυσικής γλώσσας από υπολογιστικά συστήματα. Στο πλαίσιο της ανάλυσης επιστημονικής βιβλιογραφίας, οι τεχνικές επεξεργασίας φυσικής γλώσσας χρησιμοποιούνται για την επεξεργασία μεγάλων συλλογών κειμένων, επιτρέποντας την αναγνώριση συχνών όρων, θεματικών μοτίβων και κρυφών δομών μέσα στα δεδομένα. Σύμφωνα με τους Sandu et al. [12], οι τεχνικές εξόρυξης κειμένου και επεξεργασίας φυσικής γλώσσας μπορούν να αξιοποιηθούν για τη συστηματική ανάλυση μεγάλου όγκου κειμενικών δεδομένων και την εξαγωγή χρήσιμων πληροφοριών από αυτά. Στην παρούσα εργασία χρησιμοποιήθηκαν τρεις βασικές τεχνικές ανάλυσης κειμένου: η ανάλυση N-Grams, η μέθοδος TF-IDF και η θεματική μοντελοποίηση μέσω LDA.

Τα N-Grams είναι συνεχόμενες ακολουθίες N λέξεων ή χαρακτήρων από ένα κείμενο και αποτελούν μία από τις πιο θεμελιώδεις τεχνικές εξόρυξης κειμένου και το NLP. Ειδικότερα, τα unigrams αντιστοιχούν σε μεμονωμένες λέξεις, τα bigrams σε ζεύγη δύο διαδοχικών λέξεων και τα trigrams σε ομάδες τριών λέξεων. Τα N-Grams είναι μια τεχνική που είναι ανεξάρτητη σε θέμα γλώσσας, γεγονός που τα καθιστά εφαρμόσιμα σε συλλογές κειμένων σε οποιαδήποτε γλώσσα, ενώ παράλληλα παρουσιάζουν ανθεκτικότητα σε ορθογραφικά λάθη και παραλλαγές κειμένου. Χρησιμοποιούνται ευρέως για ανάκτηση πληροφοριών, αυτόματη περίληψη κειμένου και αναγνώριση θεματικής ομοιότητας μεταξύ εγγράφων.

Το TF-IDF είναι μια στατιστική μέθοδος αξιολόγησης της σημασίας ενός όρου μέσα σε ένα έγγραφο, σε σχέση με ένα σύνολο εγγράφων. Οι θεωρητικές βάσεις της μεθόδου συνδέονται με την εργασία της Spärck Jones [3], η οποία υποστήριξε ότι η σημασία ενός όρου στην ανάκτηση πληροφοριών μπορεί να εκτιμηθεί στατιστικά με βάση τη συχνότητα εμφάνισής του στη συλλογή εγγράφων. Συγκεκριμένα, οι όροι που εμφανίζονται συχνά σε ένα συγκεκριμένο έγγραφο αλλά σπάνια στο σύνολο της συλλογής θεωρούνται πιο διακριτικοί και άρα πιο αντιπροσωπευτικοί για το περιεχόμενο του εγγράφου. Η μέθοδος βασίζεται σε δύο συνιστώσες: τη συχνότητα όρου, η οποία μετρά πόσο συχνά εμφανίζεται ένας όρος σε ένα έγγραφο, και την αντίστροφη συχνότητα εγγράφων, η οποία μειώνει τη βαρύτητα όρων που εμφανίζονται σε μεγάλο αριθμό εγγράφων. Ο συνδυασμός των δύο αυτών συνιστωσών οδηγεί σε μια βαθμολογία που αναδεικνύει όρους με υψηλή διακριτική αξία μέσα στη συλλογή κειμένων.

Το Topic Modelling είναι μια τεχνική μηχανικής μάθησης που επιτρέπει την αυτόματη ανακάλυψη κρυφών θεματικών δομών σε μεγάλες συλλογές κειμένων. Ο πιο διαδεδομένος αλγόριθμος topic modelling είναι το Latent Dirichlet Allocation (LDA), που αναπτύχθηκε από τους Blei, Ng και Jordan (2003) [4]. Όπως οι ίδιοι περιγράφουν, το LDA είναι ένα γενετικό πιθανοτικό μοντέλο (generative probabilistic model) για συλλογές διακριτών

δεδομένων, όπως τα κείμενα. Πιο συγκεκριμένα, αποτελεί ένα τριεπίπεδο ιεραρχικό Bayesian μοντέλο, στο οποίο κάθε έγγραφο μοντελοποιείται ως ένα πεπερασμένο μείγμα θεμάτων, και κάθε θέμα χαρακτηρίζεται από μια κατανομή πιθανοτήτων πάνω σε λέξεις. Η βασική αρχή λειτουργίας του LDA στηρίζεται στην υπόθεση ότι κάθε λέξη ενός εγγράφου παράγεται από ένα θέμα που επιλέγεται τυχαία, και ότι κάθε έγγραφο μπορεί να συνδέεται με πολλαπλά θέματα σε διαφορετικές αναλογίες. Αυτό διαφοροποιεί το LDA από απλούστερα μοντέλα, όπως το "mixture of unigrams", το οποίο περιορίζει κάθε έγγραφο σε ένα μόνο θέμα. Ο αλγόριθμος εκτελεί την παρακάτω γενετική διαδικασία για κάθε έγγραφο: αρχικά επιλέγεται μια κατανομή θεμάτων από μια Dirichlet κατανομή (κατανομή πάνω σε πιθανότητες), και στη συνέχεια για κάθε λέξη του εγγράφου επιλέγεται ένα θέμα από αυτή την κατανομή και η λέξη παράγεται από την αντίστοιχη κατανομή λέξεων του θέματος. Το LDA παρουσιάζει σημαντικά πλεονεκτήματα έναντι προγενέστερων μεθόδων, καθώς δεν εμφανίζει πρόβλημα υπερπροσαρμογής και μπορεί να αποδώσει πιθανότητα σε νέα έγγραφα που δεν συμπεριλαμβάνονταν στο σύνολο εκπαίδευσης.

Κεφάλαιο 3

Μεθοδολογία

3.1 Προκαταρκτική Μελέτη	16
3.2 Συλλογή Δεδομένων	17
3.3 Εργαλεία Ανάλυσης Δεδομένων	19
3.3.1 Visual Studio Code	19
3.3.2 R	20
3.3.3 VOSviewer	20
3.3.4 Python & Streamlit	21
3.4 Καθαρισμός Δεδομένων	21
3.4.1 Αυτοματοποιημένο Φιλτράρισμα	21
3.4.2 Λεμματοποίηση	22
3.4.3 Λέξεις Διακοπής	23
3.5 Τεχνικές Ανάλυσης	24
3.5.1 N-Grams	24
3.5.2 TF-IDF	24
3.5.3 LDA	25
3.6 Ερευνητικές Ερωτήσεις	26

3.1 Προκαταρκτική Μελέτη

Πριν από την έναρξη της κύριας ερευνητικής διαδικασίας, πραγματοποιήθηκε μια εκτεταμένη προκαταρκτική μελέτη με στόχο την εξοικείωση με το ερευνητικό πεδίο της επαναχρησιμοποίησης λογισμικού και με τη μεθοδολογία της βιβλιομετρικής ανάλυσης. Η φάση αυτή αποτέλεσε απαραίτητο θεμέλιο για τον ορισμό των ερευνητικών ερωτήσεων, την επιλογή των κατάλληλων εργαλείων και τη διαμόρφωση της μεθοδολογίας που ακολουθήθηκε στα επόμενα βήματα.

Αρχικά, μελετήθηκαν προηγούμενες βιβλιομετρικές αναλύσεις σε συναφείς περιοχές της μηχανικής λογισμικού, με στόχο την κατανόηση της δομής και της μεθοδολογίας τέτοιων μελετών. Συγκεκριμένα, μελετήθηκαν εργασίες που πραγματεύονται τον σχεδιασμό και την εφαρμογή βιβλιομετρικών μελετών, όπως η εργασία των Öztürk et al. [11], καθώς και

εργασίες που αξιολογούν βιβλιομετρικά εργαλεία, όπως η μελέτη του Tomaszewski [15]. Επιπλέον, εξετάστηκαν εφαρμογές βιβλιομετρικής ανάλυσης σε συναφείς περιοχές της μηχανικής λογισμικού, όπως η μελέτη των Wong et al. [16]. Η μελέτη αυτών των εργασιών παρείχε χρήσιμες κατευθύνσεις για τον ορισμό των ερευνητικών ερωτήσεων, την επιλογή των κατάλληλων εργαλείων ανάλυσης και τη δομή της παρούσας μελέτης.

Παράλληλα, πραγματοποιήθηκε εξοικείωση με το ερευνητικό τοπίο της επαναχρησιμοποίησης λογισμικού μέσω της μελέτης των κύριων συνεδρίων και περιοδικών του πεδίου. Συγκεκριμένα, εξετάστηκαν τα θέματα και οι κατευθύνσεις του International Conference on Software and Systems Reuse (ICSR), το οποίο αποτελεί το κύριο αφιερωμένο συνέδριο στο πεδίο, καθώς και έγκριτα περιοδικά όπως το Elsevier Journal of Systems and Software (JSS) και το IEEE Transactions on Software Engineering. Η εξέταση αυτών των πηγών παρείχε μια ολοκληρωμένη εικόνα των τρεχουσών ερευνητικών κατευθύνσεων και των ανοιχτών προβλημάτων στο πεδίο.

Τέλος, πραγματοποιήθηκε εξοικείωση με τη βιβλιογραφική βάση Scopus, εκτελώντας δοκιμαστικές αναζητήσεις και εξερευνώντας τις δυνατότητες φιλτραρίσματος και εξαγωγής δεδομένων που παρέχει. Η φάση αυτή ήταν καθοριστική για τον ορισμό του τελικού όρου αναζήτησης και των κριτηρίων φιλτραρίσματος που χρησιμοποιήθηκαν στη συνέχεια.

3.2 Συλλογή Δεδομένων

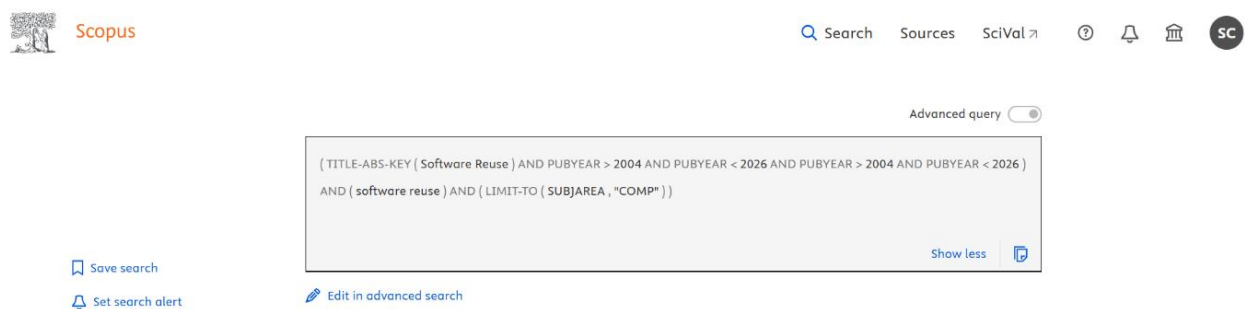
Για τη συλλογή των δεδομένων χρησιμοποιήθηκε η βιβλιογραφική βάση Scopus, η οποία επιλέχθηκε λόγω της εκτεταμένης κάλυψής της σε επιστημονικές δημοσιεύσεις στον τομέα της Πληροφορικής. Το Scopus είναι η μεγαλύτερη βάση δεδομένων peer-reviewed (αξιολόγηση από επιστήμονες του ίδιου πεδίου) βιβλιογραφίας, αναπτυγμένη από τον εκδοτικό οίκο Elsevier, και παρέχει πρόσβαση σε εκατομμύρια επιστημονικές δημοσιεύσεις από περιοδικά, συνέδρια και βιβλία. Διαθέτει προηγμένες δυνατότητες αναζήτησης και φιλτραρίσματος, καθώς και εργαλεία για την εξαγωγή δομημένων δεδομένων σε διάφορες μορφές, όπως CSV και BibTeX, που διευκολύνουν την περαιτέρω επεξεργασία και ανάλυση.



Εικόνα 3.2α: Scopus

Η επιλογή της Scopus βασίστηκε επίσης στην ευρεία χρήση της σε βιβλιομετρικές μελέτες, καθώς αποτελεί μία από τις βασικές πηγές άντλησης επιστημονικών μεταδεδομένων για την ποσοτική ανάλυση της βιβλιογραφίας. Οι Donthu et al. [2] αναφέρουν ότι η ανάπτυξη της βιβλιομετρικής ανάλυσης συνδέεται με τη διαθεσιμότητα οργανωμένων επιστημονικών βάσεων δεδομένων, όπως το Scopus και το Web of Science, οι οποίες επιτρέπουν τη συλλογή και ανάλυση μεγάλου αριθμού δημοσιεύσεων. Επιπλέον, στη μηχανική λογισμικού, οι Wong et al. [16] πραγματοποίησαν βιβλιομετρική ανάλυση για τη χαρτογράφηση θεμάτων, ερευνητών και ιδρυμάτων του πεδίου. Για τον λόγο αυτό, η Scopus κρίθηκε κατάλληλη για τη συλλογή των δεδομένων της παρούσας εργασίας. Η αναζήτηση πραγματοποιήθηκε μέσω του advanced search της Scopus, χρησιμοποιώντας το ακόλουθο όρο αναζήτησης:

```
( TITLE-ABS-KEY ( "Software reuse" ) AND PUBYEAR > 2004 AND PUBYEAR < 2026 ) AND ( LIMIT-TO ( SUBJAREA , "COMP" ) )
```



Εικόνα 3.2β: Ο όρος αναζήτησης

Ο όρος (query) εφαρμόστηκε στα πεδία τίτλου, περίληψης και λέξεων-κλειδίων (TITLE-ABS-KEY), ενώ παράλληλα ορίστηκε χρονικό εύρος από το 2005 έως το 2025 και φίλτρο θεματικής περιοχής αποκλειστικά στον τομέα της Πληροφορικής, ώστε να αποκλειστούν δημοσιεύσεις από άσχετα επιστημονικά πεδία. Η επιλογή της περιόδου 2005–2025 έγινε με στόχο την αποτύπωση της πιο πρόσφατης και σύγχρονης ερευνητικής δραστηριότητας στο πεδίο της επαναχρησιμοποίησης λογισμικού. Παρόλο που η έννοια της επαναχρησιμοποίησης λογισμικού έχει παλαιότερες ρίζες, η παρούσα εργασία επικεντρώνεται στις δύο τελευταίες δεκαετίες, κατά τις οποίες το πεδίο συνδέεται περισσότερο με σύγχρονες πρακτικές μηχανικής λογισμικού. Με αυτόν τον τρόπο, η ανάλυση διατηρείται εστιασμένη στην τρέχουσα εξέλιξη της ερευνητικής περιοχής και αποφεύγεται η υπερβολική διεύρυνση του συνόλου δεδομένων με παλαιότερες δημοσιεύσεις που ενδέχεται να αφορούν διαφορετικό τεχνολογικό πλαίσιο.

Για κάθε δημοσίευση που επέστρεψε η αναζήτηση εξήχθησαν οι ακόλουθες κατηγορίες πληροφοριών:

- τίτλος δημοσίευσης (Title),
- συγγραφείς (Authors),
- περίληψη (Abstract),
- έτος δημοσίευσης (Year),
- τίτλος πηγής δημοσίευσης (Source Title),
- τύπος δημοσίευσης (Source & Document Type),
- ιδρύματα των συγγραφέων (Affiliations),
- λέξεις-κλειδιά των συγγραφέων (Author Keywords),
- ευρετηριασμένες λέξεις-κλειδιά (Indexed Keywords),
- μοναδικός κωδικός δημοσίευσης (DOI).

Η αναζήτηση επέστρεψε συνολικά 12.673 δημοσιεύσεις, οι οποίες εξήχθησαν από τη βάση Scopus σε μορφή αρχείου CSV για περαιτέρω επεξεργασία. Στο εξαγόμενο αρχείο, κάθε γραμμή αντιπροσώπευε μία μοναδική δημοσίευση, ενώ οι στήλες περιλάμβαναν τις παραπάνω κατηγορίες. Στη συνέχεια, το αρχικό σύνολο δεδομένων υποβλήθηκε σε διαδικασία φιλτραρίσματος με στόχο τη διασφάλιση της ποιότητας και της συνέπειας του τελικού συνόλου δεδομένων.

3.3 Εργαλεία Ανάλυσης Δεδομένων

Για την ανάλυση του συνόλου δεδομένων χρησιμοποιήθηκαν τέσσερα κύρια εργαλεία: το Visual Studio Code, η γλώσσα προγραμματισμού R, το λογισμικό VOSviewer και η γλώσσα προγραμματισμού Python με το framework Streamlit.

3.3.1 Visual Studio Code

Το Visual Studio Code (VS Code) είναι ένα ελαφρύ αλλά ισχυρό περιβάλλον ανάπτυξης κώδικα (IDE) ανοικτού κώδικα, που αναπτύχθηκε από τη Microsoft και κυκλοφόρησε για πρώτη φορά το 2015. Υποστηρίζει πληθώρα γλωσσών προγραμματισμού μέσω ενός εκτεταμένου συστήματος επεκτάσεων, ενώ προσφέρει λειτουργίες όπως αυτόματη συμπλήρωση κώδικα, εντοπισμό σφαλμάτων και ενσωματωμένο terminal. Χρησιμοποιείται ευρέως τόσο από επαγγελματίες προγραμματιστές όσο και από ερευνητές, λόγω της

ευελιξίας και της ευκολίας χρήσης του. Στην παρούσα εργασία χρησιμοποιήθηκε ως το κύριο περιβάλλον συγγραφής και εκτέλεσης του κώδικα, αξιοποιώντας την R για την ανάπτυξη και εκτέλεση των σεναρίων ανάλυσης καθώς και την Python για την υλοποίηση της διαδικτυακής εφαρμογής της παρούσας εργασίας.



Εικόνα 3.3.1: Visual Studio Code

3.3.2 R

Η R είναι μια γλώσσα προγραμματισμού ανοικτού κώδικα, σχεδιασμένη ειδικά για στατιστική ανάλυση, επεξεργασία δεδομένων και οπτικοποίηση. Χρησιμοποιείται ευρέως στην ακαδημαϊκή κοινότητα και στην έρευνα, καθώς προσφέρει ένα πλούσιο οικοσύστημα πακέτων που καλύπτουν ένα ευρύ φάσμα αναλυτικών εργασιών, από την εξόρυξη δεδομένων και την ανάλυση κειμένου έως την εφαρμογή αλγορίθμων μηχανικής μάθησης. Στην παρούσα εργασία η R αποτέλεσε τη βασική γλώσσα προγραμματισμού που χρησιμοποιήθηκε τόσο για το φιλτράρισμα των δεδομένων όσο και για την εφαρμογή των τεχνικών ανάλυσης κειμένου. Συγκεκριμένα, χρησιμοποιήθηκαν οι ακόλουθες βιβλιοθήκες: η tidyverse για τη γενική επεξεργασία και οπτικοποίηση δεδομένων, η tidytext για τη διακριτοποίηση (tokenization) και την ανάλυση κειμένου, η tm για τη δημιουργία του πίνακα εγγράφων-όρων (document-term matrix) και τη διαχείριση συλλογών κειμένων, η topicmodels για την εφαρμογή του αλγορίθμου Λανθάνουσας Κατανομής (LDA), η textstem για τη λεμματοποίηση (lemmatization) και η scales για τη μορφοποίηση των κλιμάκων στις οπτικοποιήσεις.

3.3.3 VOSviewer

Το VOSviewer είναι ένα εξειδικευμένο λογισμικό ανοικτού κώδικα για τη δημιουργία και οπτικοποίηση βιβλιομετρικών δικτύων, που αναπτύχθηκε από το Centre for Science and Technology Studies (CWTS) του Πανεπιστημίου Leiden. Χρησιμοποιείται ευρέως στην ακαδημαϊκή κοινότητα για την ανάλυση δικτύων συν-συγγραφίας, συν-εμφάνισης λέξεων-κλειδιών και αναφορών, επιτρέποντας την οπτική αναπαράσταση της δομής και των σχέσεων μέσα σε ένα ερευνητικό πεδίο. Στην παρούσα εργασία χρησιμοποιήθηκε για τη δημιουργία χάρτη συν-εμφάνισης λέξεων-κλειδιών (keyword co-occurrence map), ο οποίος αναδεικνύει συστάδες συσχετιζόμενων ερευνητικών θεμάτων, καθώς και για την ανάλυση των

κορυφαίων χωρών και ιδρυμάτων που συμβάλλουν στην έρευνα της επαναχρησιμοποίησης λογισμικού.

3.3.4 Python & Streamlit

Η Python είναι μια γλώσσα προγραμματισμού γενικής χρήσης και ανοικτού κώδικα, η οποία έχει καθιερωθεί ως μία από τις πιο δημοφιλείς γλώσσες στον τομέα της επιστήμης δεδομένων και της ανάπτυξης διαδικτυακών εφαρμογών. Χαρακτηρίζεται από απλή και ευανάγνωστη σύνταξη, ενώ διαθέτει ένα εκτεταμένο οικοσύστημα βιβλιοθηκών που καλύπτουν ένα ευρύ φάσμα εφαρμογών. Στην παρούσα εργασία η Python χρησιμοποιήθηκε για την ανάπτυξη μιας διαδικτυακής εφαρμογής παρουσίασης των αποτελεσμάτων, αξιοποιώντας το framework Streamlit. Το Streamlit είναι μια βιβλιοθήκη Python που επιτρέπει την ανάπτυξη διαδραστικών διαδικτυακών εφαρμογών με ελάχιστο κώδικα, χωρίς να απαιτείται γνώση τεχνολογιών front-end (διεπαφή χρήστη). Η εφαρμογή που αναπτύχθηκε επιτρέπει στον χρήστη να εξερευνά διαδραστικά τα δεδομένα και τα αποτελέσματα της ανάλυσης, παρέχοντας μια φιλική προς τον χρήστη διεπαφή για την οπτικοποίηση των ευρημάτων της εργασίας.

3.4 Καθαρισμός Δεδομένων

Μετά τη συλλογή των δεδομένων από το Scopus, ακολούθησε η διαδικασία καθαρισμού και προεπεξεργασίας των δεδομένων, με στόχο την εξασφάλιση της ποιότητας και της συνέπειας του τελικού συνόλου δεδομένων. Η διαδικασία αυτή περιλάμβανε τρία βασικά βήματα: το αυτοματοποιημένο φιλτράρισμα, τη λεμματοποίηση και την αφαίρεση των λέξεων διακοπής.

3.4.1 Αυτοματοποιημένο Φιλτράρισμα

Για το φιλτράρισμα των 12.673 δημοσιεύσεων που επέστρεψε η αναζήτηση στο Scopus, αναπτύχθηκε ένα αυτοματοποιημένο σύστημα φιλτραρίσματος σε γλώσσα R. Η διαδικασία αποτελούνταν από δύο κύρια στάδια.

Στο πρώτο στάδιο εφαρμόστηκε φίλτρο αξιόπιστων περιοδικών και συνεδρίων, διατηρώντας μόνο δημοσιεύσεις από προκαθορισμένη λίστα έγκριτων περιοδικών και συνεδρίων της

Πληροφορικής, αποκλείοντας έτσι δημοσιεύσεις από άσχετα ερευνητικά πεδία. Η πλήρης λίστα των αποδεκτών venues παρατίθεται στο Παράρτημα Α.

Στο δεύτερο στάδιο εφαρμόστηκε μηχανισμός βαθμολόγησης συνάφειας, όπου κάθε δημοσίευση έλαβε βαθμολογία ανάλογα με την παρουσία όρων σχετικών με την επαναχρησιμοποίηση λογισμικού στον τίτλο, τις λέξεις-κλειδιά και την περίληψη. Οι σχετικοί όροι αναζήτησης περιλάμβαναν κύριες φράσεις όπως "software reuse", "software reusability", "software product lines" και "SPL", καθώς και παραλλαγές όπως "code reuse", "component reuse", "library reuse", "framework reuse" και "asset reuse". Για παράδειγμα, μια δημοσίευση με τίτλο "A Framework for Software Reuse in Product Lines" θα λάμβανε υψηλή βαθμολογία λόγω της παρουσίας του όρου "software reuse" στον τίτλο και του "product lines" στις λέξεις-κλειδιά.

Η βαθμολόγηση υπολογίστηκε ως εξής: ακριβής αντιστοίχιση του όρου "software reuse" στον τίτλο → +6 βαθμοί, παραλλαγές στον τίτλο → +2 βαθμοί, αντιστοίχιση στις λέξεις-κλειδιά → +5 βαθμοί, παραλλαγές στις λέξεις-κλειδιά → +2 βαθμοί, αντιστοίχιση στο πλήρες κείμενο (τίτλος + περίληψη + λέξεις-κλειδιά) → +2 βαθμοί, παραλλαγές στο πλήρες κείμενο → +1 βαθμός.

Παράλληλα, εφαρμόστηκε φίλτρο αποκλεισμού για δημοσιεύσεις που περιείχαν όρους μη σχετικούς με την επαναχρησιμοποίηση λογισμικού. Συγκεκριμένα, αποκλείστηκαν δημοσιεύσεις που περιείχαν όρους από τις ακόλουθες κατηγορίες: επαναχρησιμοποίηση νερού ("water reuse", "wastewater", "greywater", "sewage", "desalination", "membrane bioreactor"), ιατρικές εφαρμογές ("medical reuse", "organ transplant", "kidney", "liver", "surgical", "clinical"), κατασκευαστικά υλικά ("construction", "concrete reuse", "material reuse", "textile reuse", "plastic reuse") και ηλεκτρονικά απόβλητα ("hardware reuse", "e-waste", "electronic waste").

Δημοσιεύσεις με βαθμολογία 7 ή άνω κατατάχθηκαν ως σχετικές (KEEP), βαθμολογία 4–6 ως οριακές (BORDERLINE) και οι υπόλοιπες απορρίφθηκαν (DROP). Για την επαλήθευση της ορθότητας του φιλτραρίσματος, πραγματοποιήθηκε χειροκίνητος έλεγχος 50 τυχαία επιλεγμένων δημοσιεύσεων από το τελικό σύνολο δεδομένων, επιβεβαιώνοντας ότι οι διατηρηθείσες δημοσιεύσεις ήταν πράγματι σχετικές με την επαναχρησιμοποίηση λογισμικού. Μετά την εφαρμογή αυτής της διαδικασίας, το τελικό σύνολο δεδομένων αποτελούνταν από 2.962 δημοσιεύσεις.

3.4.2 Λεμματοποίηση

Η λεμματοποίηση (lemmatization) είναι η διαδικασία αναγωγής των λέξεων στη βασική τους μορφή (λήμμα), ώστε διαφορετικές μορφές της ίδιας λέξης να αντιμετωπίζονται ως ένας ενιαίος όρος κατά την ανάλυση. Για παράδειγμα, οι λέξεις “running”, “runs” και “ran” ανάγονται στη βασική μορφή “run”.

Στην παρούσα εργασία, η λεμματοποίηση εφαρμόστηκε στο πλαίσιο των μεθόδων ανάλυσης κειμένου με χρήση της βιβλιοθήκης `textstem` της R, η οποία παρέχει συναρτήσεις για τη μετατροπή λέξεων στη λεμματική τους μορφή, αξιοποιώντας λεξικά και κανόνες μορφολογικής ανάλυσης της αγγλικής γλώσσας. Συγκεκριμένα, χρησιμοποιήθηκε η συνάρτηση `lemmatize_words()` για τη συστηματική κανονικοποίηση των όρων πριν από την εφαρμογή των τεχνικών ανάλυσης.

Επιπλέον, παρατηρήθηκε ότι ορισμένοι όροι δεν κανονικοποιήθηκαν πλήρως μέσω της αυτόματης λεμματοποίησης, με αποτέλεσμα να απαιτηθεί επιπλέον χειροκίνητη ενοποίησή τους. Οι συγκεκριμένες περιπτώσεις χειροκίνητης κανονικοποίησης που εφαρμόστηκαν ήταν οι εξής: οι όροι “spl” και “sple” ενοποιήθηκαν στη μορφή “spl”, ο όρος “engineer” αντιστοιχίστηκε στη μορφή “engineering”, ο όρος “orient” αντιστοιχίστηκε στη μορφή “oriented” και ο όρος “datum” αντιστοιχίστηκε στη μορφή “data”. Η διαδικασία αυτή συνέβαλε στη βελτίωση της συνέπειας των όρων και στη μείωση της διάσπασης εννοιών που αναφέρονταν στο ίδιο θεματικό περιεχόμενο.

3.4.3 Λέξεις Διακοπής

Οι λέξεις διακοπής (stop words) αναφέρονται σε συχνά εμφανιζόμενες λέξεις που συνήθως δεν προσφέρουν ουσιαστική θεματική πληροφορία κατά την ανάλυση κειμένου. Πρόκειται κυρίως για γραμματικές λέξεις υψηλής συχνότητας, όπως άρθρα, προθέσεις, αντωνυμίες, σύνδεσμοι και κοινά ρήματα, τα οποία εμφανίζονται συστηματικά σε μεγάλο αριθμό εγγράφων χωρίς να συμβάλλουν ουσιαστικά στην κατανόηση του περιεχομένου.

Στην παρούσα εργασία χρησιμοποιήθηκε το προκαθορισμένο σύνολο stop words της βιβλιοθήκης `tidytext` της R, το οποίο περιλαμβάνει περίπου 1.140 κοινές αγγλικές λέξεις. Ενδεικτικά, το σύνολο αυτό περιλαμβάνει άρθρα όπως “the”, “a” και “an”, προθέσεις όπως “of”, “in”, “to” και “for”, αντωνυμίες όπως “it”, “this” και “that”, συνδέσμους όπως “and”, “or” και “but”, καθώς και συχνά χρησιμοποιούμενα ρήματα όπως “is”, “are”, “was”, “be” και “have”.

Ωστόσο, λόγω της φύσης του συγκεκριμένου dataset, κρίθηκε απαραίτητη και η χειροκίνητη προσθήκη επιπλέον λέξεων διακοπής, οι οποίες δεν αφαιρούνταν από το προκαθορισμένο σύνολο. Συγκεκριμένα, αφαιρέθηκαν γενικοί όροι όπως “use”, “based”, “method” και “result”, οι οποίοι εμφανίζονταν πολύ συχνά χωρίς να προσφέρουν ουσιαστική θεματική πληροφορία. Επιπλέον, αφαιρέθηκαν όροι ειδικού περιεχομένου όπως “software” και “reuse”, καθώς εμφανίζονταν σχεδόν σε όλες τις δημοσιεύσεις λόγω της φύσης του πεδίου και επηρέαζαν δυσανάλογα τα αποτελέσματα. Τέλος, αφαιρέθηκαν ονόματα εκδοτών και βιβλιογραφικοί όροι όπως “elsevier”, “springer”, “ieee”, “acm”, “paper”, “article” και “journal”, καθώς προέρχονταν κυρίως από μεταδεδομένα ή τυποποιημένο κείμενο δημοσιεύσεων και δεν σχετίζονταν με το πραγματικό θεματικό περιεχόμενο.

Η διαδικασία αυτή συνέβαλε στη βελτίωση της ποιότητας των δεδομένων και στην παραγωγή πιο ακριβών αποτελεσμάτων στις επόμενες τεχνικές ανάλυσης κειμένου.

3.5 Τεχνικές Ανάλυσης

Πριν από την παρουσίαση των τεχνικών ανάλυσης που χρησιμοποιήθηκαν, είναι σημαντικό να διευκρινιστεί ότι το σύνολο της ανάλυσης βασίστηκε αποκλειστικά στα μεταδεδομένα και τις περιλήψεις (abstracts) των δημοσιεύσεων, όπως αυτά εξήχθησαν από τη βάση Scopus. Το πλήρες κείμενο των δημοσιεύσεων δεν χρησιμοποιήθηκε, καθώς η πρόσβαση σε αυτό δεν είναι πάντα ελεύθερα διαθέσιμο. Η επιλογή αυτή είναι συνηθισμένη πρακτική στη βιβλιομετρική ανάλυση, καθώς οι περιλήψεις περιέχουν επαρκή πληροφορία για την αναγνώριση των κυρίαρχων θεμάτων και τάσεων ενός ερευνητικού πεδίου.

3.5.1 N-Grams

Στο πλαίσιο της παρούσας βιβλιομετρικής ανάλυσης, η τεχνική των N-Grams εφαρμόστηκε στις περιλήψεις και στις λέξεις-κλειδιά των δημοσιεύσεων με στόχο την αναγνώριση των πιο συχνά εμφανιζόμενων όρων και συνδυασμών λέξεων. Πιο συγκεκριμένα, αναλύθηκαν μονολεκτικοί όροι (unigrams), συνδυασμοί δύο λέξεων (bigrams) και συνδυασμοί τριών λέξεων (trigrams), ώστε να εντοπιστούν επαναλαμβανόμενα μοτίβα στη βιβλιογραφία.

Η συγκεκριμένη τεχνική χρησιμοποιήθηκε για την αποτύπωση των βασικών θεματικών τάσεων του πεδίου και για την ανάδειξη σημαντικών εννοιών που εμφανίζονται συστηματικά στις δημοσιεύσεις. Τα αποτελέσματα της ανάλυσης αξιοποιούνται στο Κεφάλαιο 4 για την παρουσίαση των συχνότερων όρων και φράσεων, ενώ συμβάλλουν και στη συνολική ερμηνεία των θεματικών κατευθύνσεων στο Κεφάλαιο 5.

3.5.2 TF-IDF

Στο πλαίσιο της παρούσας βιβλιομετρικής ανάλυσης, η τεχνική TF-IDF εφαρμόστηκε στις περιλήψεις των δημοσιεύσεων με στόχο τον εντοπισμό όρων που παρουσιάζουν υψηλή σημασία σε συγκεκριμένα έγγραφα, χωρίς όμως να εμφανίζονται υπερβολικά συχνά σε ολόκληρο το σύνολο των δεδομένων.

Η χρήση της συγκεκριμένης τεχνικής κρίθηκε απαραίτητη, καθώς η απλή ανάλυση συχνότητας δεν είναι πάντοτε αρκετή για την ανάδειξη ουσιαστικών όρων. Μέσω του TF-IDF κατέστη δυνατός ο εντοπισμός πιο εξειδικευμένων εννοιών που χαρακτηρίζουν επιμέρους δημοσιεύσεις και ερευνητικές περιοχές. Τα αποτελέσματα χρησιμοποιούνται στο Κεφάλαιο 4 για την παρουσίαση των σημαντικότερων όρων και στη συνέχεια συγκρίνονται με τις υπόλοιπες τεχνικές ανάλυσης.

3.5.3 LDA

Στην παρούσα εργασία, η τεχνική LDA εφαρμόστηκε με στόχο την αναγνώριση των κυρίαρχων θεματικών περιοχών στον τομέα της επαναχρησιμοποίησης λογισμικού και την καλύτερη κατανόηση της θεματικής δομής της βιβλιογραφίας.

Η ανάλυση πραγματοποιήθηκε στις περιλήψεις των δημοσιεύσεων μετά την ολοκλήρωση των σταδίων προεπεξεργασίας, ενώ εξετάστηκαν πολλαπλές τιμές θεμάτων πριν από την επιλογή του τελικού αριθμού. Ο τελικός αριθμός θεμάτων επιλέχθηκε με βάση τον δείκτη συνοχής UMass Coherence Score. Σύμφωνα με τους Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012) [13], η συνοχή θεμάτων χρησιμοποιείται για την αξιολόγηση του βαθμού στον οποίο οι λέξεις ενός θέματος εμφανίζονται συχνά μαζί και σχηματίζουν νοηματικά ερμηνεύσιμες ομάδες. Ο δείκτης ορίζεται ως:

$$score(v_i, v_j, \epsilon) = \log \frac{D(v_i, v_j) + \epsilon}{D(v_j)}$$

όπου $D(v_i, v_j)$ μετρά τον αριθμό εγγράφων που περιέχουν ταυτόχρονα τις λέξεις v_i και v_j , και $D(v_j)$ μετρά τον αριθμό εγγράφων που περιέχουν τη λέξη v_j . Ο δείκτης αυτός αξιολογεί την εσωτερική συνοχή κάθε θέματος εξετάζοντας κατά πόσο οι λέξεις που το αποτελούν εμφανίζονται μαζί στο ίδιο σύνολο κειμενικών δεδομένων που χρησιμοποιήθηκε για την

εκπαίδευση του μοντέλου, χωρίς να απαιτείται εξωτερική πηγή δεδομένων, καθιστώντας τον ιδιαίτερα κατάλληλο για βιβλιομετρικές αναλύσεις. Υψηλότερες (λιγότερο αρνητικές) τιμές υποδηλώνουν καλύτερη θεματική συνοχή. Παράλληλα, πραγματοποιήθηκε και χειροκίνητη αξιολόγηση των όρων κάθε θέματος ώστε να διασφαλιστεί η ερμηνευσιμότητα των αποτελεσμάτων.

Τα αποτελέσματα του LDA αξιοποιούνται στο Κεφάλαιο 4 για την παρουσίαση των βασικών θεματικών περιοχών και της εξέλιξής τους στον χρόνο, ενώ στο Κεφάλαιο 5 χρησιμοποιούνται για την ερμηνεία των κυρίαρχων ερευνητικών τάσεων του πεδίου.

3.6 Ερευνητικές Ερωτήσεις

Οι ερευνητικές ερωτήσεις (Research Questions — RQ) αποτελούν θεμελιώδες στοιχείο κάθε επιστημονικής μελέτης, καθώς ορίζουν με σαφήνεια τους στόχους της έρευνας και κατευθύνουν τη διαδικασία συλλογής και ανάλυσης δεδομένων. Στο πλαίσιο μιας βιβλιομετρικής ανάλυσης, οι ερευνητικές ερωτήσεις χρησιμοποιούνται για να οριοθετήσουν το πεδίο της μελέτης και να διασφαλίσουν ότι η ανάλυση είναι συστηματική και εστιασμένη. Κάθε ερευνητική ερώτηση αντιστοιχεί σε μια συγκεκριμένη πτυχή της βιβλιογραφίας που επιθυμούμε να διερευνήσουμε, και τα αποτελέσματα της ανάλυσης αξιολογούνται με βάση το κατά πόσο απαντώνται επαρκώς. Η διατύπωση ερευνητικών ερωτήσεων πριν από την έναρξη της ανάλυσης αποτρέπει την αυθαίρετη επιλογή αποτελεσμάτων και ενισχύει την αντικειμενικότητα της μελέτης.

Με στόχο την οργανωμένη και συστηματική ανάλυση της βιβλιογραφίας, ορίστηκαν οκτώ ερευνητικές ερωτήσεις, οι οποίες καλύπτουν χρονικές τάσεις, γεωγραφική κατανομή, θεματική ανάλυση, ανάλυση κειμένου και χαρακτηριστικά της επιστημονικής κοινότητας. Οι ερωτήσεις αυτές παρουσιάζονται στον παρακάτω πίνακα μαζί με τη σημασία τους.

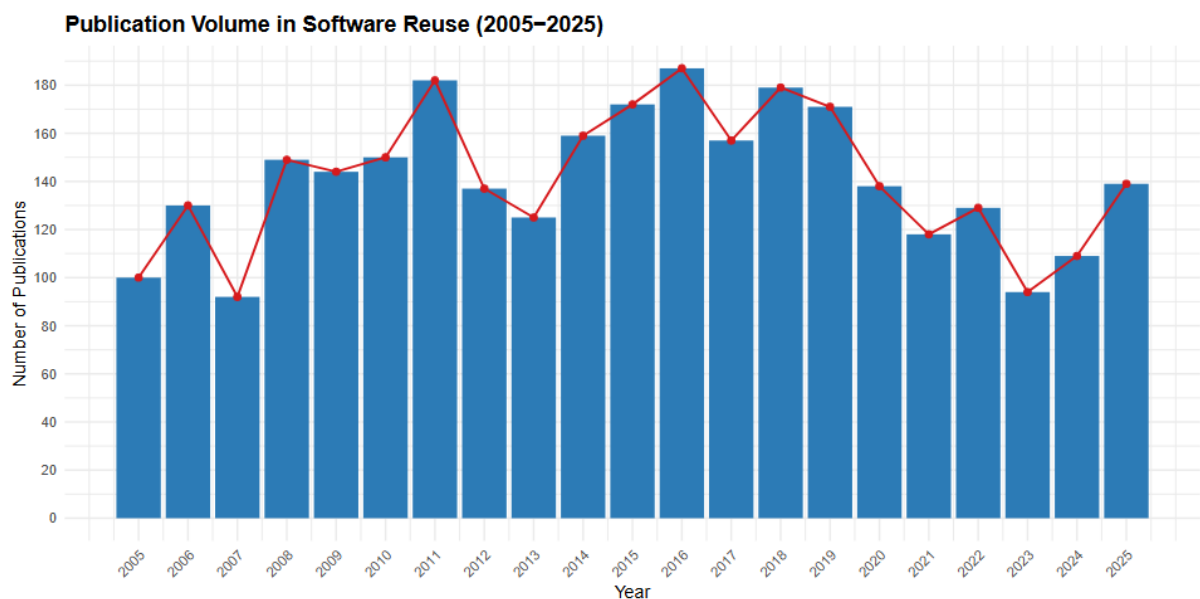
Research Question	Σημασία
RQ1: Ποια είναι η χρονική εξέλιξη του όγκου των δημοσιεύσεων στον τομέα της επαναχρησιμοποίησης λογισμικού κατά την περίοδο 2005–2025;	Βοηθά στην κατανόηση της ανάπτυξης και εξέλιξης του ερευνητικού ενδιαφέροντος στο πεδίο, καθώς και στην πρόβλεψη μελλοντικών τάσεων.
RQ2: Ποια είναι η κατανομή των τύπων δημοσιεύσεων στον τομέα της επαναχρησιμοποίησης λογισμικού;	Αποκαλύπτει τον τρόπο με τον οποίο διαχέεται η γνώση στο πεδίο, δείχνοντας αν η έρευνα δημοσιεύεται κυρίως σε συνέδρια ή περιοδικά, γεγονός που αντικατοπτρίζει τη φύση και την ωριμότητα του ερευνητικού πεδίου.
RQ3: Ποια είναι τα κορυφαία επιστημονικά περιοδικά και συνέδρια στα οποία δημοσιεύεται έρευνα σχετική με την επαναχρησιμοποίηση λογισμικού;	Βοηθά τους ερευνητές να επιλέξουν τα κατάλληλα venues για δημοσίευση και να εντοπίσουν τις πιο έγκυρες πηγές βιβλιογραφίας στο πεδίο.
RQ4: Ποιοι είναι οι πιο παραγωγικοί συγγραφείς στον τομέα της επαναχρησιμοποίησης λογισμικού; Ποια είναι η εξέλιξη της συνεργασίας μεταξύ συγγραφέων με την πάροδο του χρόνου;	Επιτρέπει στους ερευνητές να εντοπίσουν τους κορυφαίους συντελεστές του πεδίου και να υποδείχνει την αύξηση ή τη μείωση της συνεργατικότητας μεταξύ των συγγραφέων κατά την πάροδο του χρόνου
RQ5: Ποιες χώρες και ιδρύματα συμβάλλουν περισσότερο στην έρευνα της επαναχρησιμοποίησης λογισμικού;	Αναδεικνύει τη γεωγραφική κατανομή της ερευνητικής δραστηριότητας και τα κέντρα αριστείας που ηγούνται στο πεδίο.
RQ6: Ποιες είναι οι κυρίαρχες θεματικές περιοχές στη βιβλιογραφία της επαναχρησιμοποίησης λογισμικού και πώς εξελίχθηκαν οι ερευνητικές θεματικές περιοχές με την πάροδο του χρόνου;	Επιτρέπει την κατανόηση της θεματικής δομής του πεδίου και τον εντοπισμό των βασικών ερευνητικών αξόνων. Αναδεικνύει τις μεταβολές στο ερευνητικό ενδιαφέρον και βοηθά στον εντοπισμό αναδυόμενων τάσεων και νέων κατευθύνσεων έρευνας.
RQ7: Ποιοι είναι οι πιο χαρακτηριστικοί όροι της βιβλιογραφίας στον τομέα της επαναχρησιμοποίησης λογισμικού;	Επιτρέπει τον εντοπισμό των πιο σημαντικών και διακριτικών εννοιών της βιβλιογραφίας, παρέχοντας μια ποσοτική εικόνα του περιεχομένου των δημοσιεύσεων μέσω τεχνικών ανάλυσης κειμένου όπως το TF-IDF και τα N-Grams.
RQ8: Πώς σχετίζονται οι λέξεις-κλειδιά στη βιβλιογραφία της επαναχρησιμοποίησης λογισμικού με τις θεματικές περιοχές που εντοπίστηκαν;	Αναδεικνύει τις σχέσεις μεταξύ των βασικών εννοιών του πεδίου και τις θεματικές περιοχές που σχηματίζονται, επιτρέποντας την κατανόηση της δομής της γνώσης στον τομέα.

Κεφάλαιο 4

Αποτελέσματα

4.1 Χρονικές Τάσεις Δημοσιεύσεων	28
4.2 Κορυφαία Περιοδικά και Συνέδρια	30
4.3 Κατανομή Τύπων Δημοσιεύσεων	31
4.4 Κορυφαίες Χώρες και Ιδρύματα	32
4.5 Κορυφαίοι Συγγραφείς	35
4.6 Ανάλυση Λέξεων-Κλειδιών	36
4.6.1 Πιο Συχνοί Όροι (N-Grams)	36
4.6.2 Ανάλυση TF-IDF	40
4.6.3 Χάρτης Συν-Εμφάνισης Λέξεων-Κλειδιών	41
4.7 Αποτελέσματα Μοντελοποίησης Θεμάτων	43
4.7.1 Αναγνωρισμένα Θέματα (LDA)	43
4.7.2 Κατανομή Κυρίαρχων Θεμάτων	45
4.7.3 Εξέλιξη Θεμάτων στον Χρόνο	46
4.8 Εξέλιξη Συνεργασίας Συγγραφέων στον Χρόνο	48
4.9 Διαδραστική Διαδικτυακή Εφαρμογή	49

4.1 Χρονικές Τάσεις Δημοσιεύσεων



Γράφημα 4.1: Χρονικές Τάσεις Δημοσιεύσεων

Η ανάλυση της εξέλιξης των δημοσιεύσεων βασίζεται στα δεδομένα που συλλέχθηκαν από τη βάση Scopus και τα οποία έχουν προεπεξεργαστεί και φιλτραριστεί ώστε να περιλαμβάνουν μόνο σχετικές δημοσιεύσεις στον τομέα της επαναχρησιμοποίησης λογισμικού. Η εξέταση της χρονικής κατανομής των δημοσιεύσεων επιτρέπει την αποτύπωση της πορείας και της έντασης της ερευνητικής δραστηριότητας στο συγκεκριμένο πεδίο.

Συγκεκριμένα, το *Γράφημα 4.1* παρουσιάζει την εξέλιξη του αριθμού των δημοσιεύσεων για την περίοδο 2005–2025. Συνολικά παρατηρείται μια γενική αυξητική τάση κατά την πρώτη δεκαετία, με τον αριθμό των δημοσιεύσεων να αυξάνεται σταδιακά από περίπου 100 δημοσιεύσεις το 2005, παρουσιάζοντας ωστόσο ετήσιες διακυμάνσεις. Η κορύφωση παρατηρείται κατά την περίοδο 2015–2016, όπου καταγράφονται περίπου 175–180 δημοσιεύσεις ανά έτος, αποτελώντας το υψηλότερο σημείο ολόκληρης της εξεταζόμενης περιόδου. Στη συνέχεια, ο αριθμός των δημοσιεύσεων παρουσιάζει σταδιακή μείωση κατά τη δεύτερη δεκαετία, με χαρακτηριστικές διακυμάνσεις ανά έτος. Κατά την περίοδο 2018–2022 παρατηρείται μια σχετική σταθεροποίηση γύρω στις 130–170 δημοσιεύσεις ανά έτος, ενώ τα έτη 2023–2024 καταγράφουν αισθητή πτώση στις ~100 δημοσιεύσεις. Αξιοσημείωτη είναι η ανάκαμψη που παρατηρείται το 2025 με ~145 δημοσιεύσεις, αν και η μείωση των τελευταίων ετών ενδέχεται να οφείλεται εν μέρει στην καθυστέρηση ευρετηρίασης των πιο πρόσφατων δημοσιεύσεων στη βάση Scopus.

4.2 Κορυφαία Περιοδικά και Συνέδρια

Top 20 Journals & Conferences (by Number of Documents)

Rank	Venue	Documents
1	Journal of Systems and Software	86
2	Ruan Jian Xue Bao/Journal of Software	37
3	Journal of Software: Evolution and Process	29
4	International Journal of Software Engineering and Knowledge Enginee...	28
5	Jisuanji Xuebao/Chinese Journal of Computers	10
6	Journal of Software	9
7	Journal of Supercomputing	8
8	Software Quality Journal	8
9	Journal of the Brazilian Computer Society	7
10	Journal of Computer Science and Technology	6
11	Journal of Information Science and Engineering	6
12	Journal of Theoretical and Applied Information Technology	6
13	Journal of Web Semantics	6
14	Conference Proceedings of the EUROMICRO	5
15	International Journal of Advanced Computer Science and Applications	5
16	Journal of Logical and Algebraic Methods in Programming	5
17	Journal of Universal Computer Science	5
18	Huazhong Keji Daxue Xuebao (Ziran Kexue Ban)/Journal of Huazhong Un...	4
19	IEEE Systems Journal	4
20	International Journal of Computer Applications in Technology	4

Πίνακας 4.2: Κορυφαία Περιοδικά και Συνέδρια

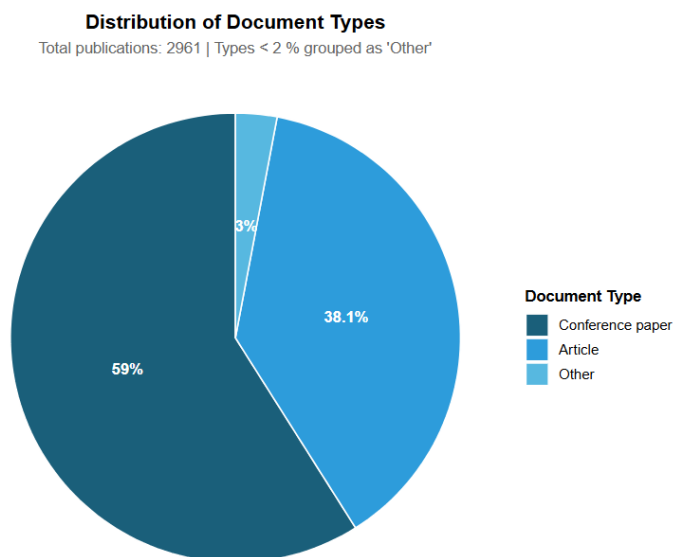
Ο Πίνακας 4.2 παρουσιάζει τα 20 κορυφαία επιστημονικά μέσα δημοσίευσης στον τομέα της επαναχρησιμοποίησης λογισμικού, λαμβάνοντας υπόψη μόνο επιστημονικά περιοδικά και πρακτικά συνεδρίων. Η κυρίαρχη πηγή δημοσίευσης είναι το Journal of Systems and Software, με 86 δημοσιεύσεις, γεγονός που δείχνει τη σημαντική παρουσία του συγκεκριμένου περιοδικού στη σχετική ερευνητική περιοχή. Ακολουθεί το Ruan Jian Xue Bao/Journal of Software με 37 δημοσιεύσεις, ενώ στην τρίτη θέση βρίσκεται το Journal of Software: Evolution and Process με 29 δημοσιεύσεις. Σημαντική παρουσία καταγράφει επίσης το International Journal of Software Engineering and Knowledge Engineering με 28 δημοσιεύσεις.

Στη συνέχεια εμφανίζονται διάφορα επιστημονικά περιοδικά που σχετίζονται με τη μηχανική λογισμικού, την ποιότητα λογισμικού, την επιστήμη υπολογιστών και συναφείς περιοχές.

Ενδεικτικά, το Jisuanji Xuebao/Chinese Journal of Computers έχει 10 δημοσιεύσεις, το Journal of Software έχει 9, ενώ το Journal of Supercomputing και το Software Quality Journal καταγράφουν από 8 δημοσιεύσεις. Επιπλέον, αρκετές πηγές εμφανίζονται με μικρότερο αλλά αξιοσημείωτο αριθμό δημοσιεύσεων, όπως το Journal of the Brazilian Computer Society με 7, το Journal of Computer Science and Technology, το Journal of Information Science and Engineering, το Journal of Theoretical and Applied Information Technology και το Journal of Web Semantics με 6 δημοσιεύσεις το καθένα.

Αξιοσημείωτη είναι η παρουσία του Conference Proceedings of the EUROMICRO με 5 δημοσιεύσεις, που αποτελεί το μοναδικό πρακτικό συνεδρίου στους top 20, υποδηλώνοντας ότι η έρευνα στην επαναχρησιμοποίηση λογισμικού δημοσιεύεται κατά κύριο λόγο σε επιστημονικά περιοδικά όταν εξετάζονται αποκλειστικά έγκριτοι χώροι.

4.3 Κατανομή Τύπων Δημοσιεύσεων

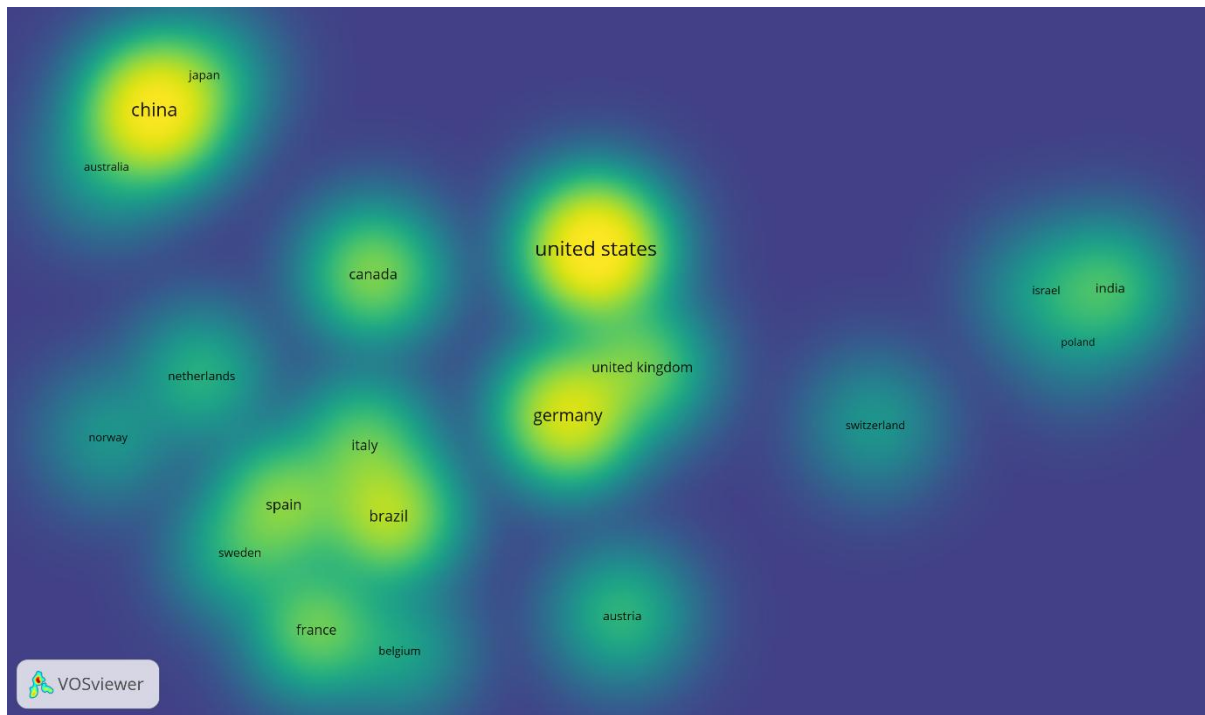


Γράφημα 4.3: Κατανομή Τύπων Δημοσιεύσεων

Το Γράφημα 4.3 παρουσιάζει την κατανομή των τύπων δημοσιεύσεων στο τελικό σύνολο δεδομένων των 2.962 εγγραφών. Για λόγους αναγνωσιμότητας, τύποι δημοσιεύσεων με ποσοστό κάτω από 2% ομαδοποιήθηκαν σε μια ενιαία κατηγορία με την ονομασία "Other". Ο κυρίαρχος τύπος δημοσίευσης είναι τα Conference Papers, τα οποία αντιπροσωπεύουν το 59% του συνόλου, γεγονός που υποδηλώνει ότι η έρευνα στον τομέα της επαναχρησιμοποίησης λογισμικού πραγματοποιείται κυρίως μέσω συνεδρίων. Ο δεύτερος πιο συχνός τύπος είναι τα Articles, τα οποία αντιπροσωπεύουν το 38.1% του συνόλου, αναδεικνύοντας τη σημαντική παρουσία της έρευνας και σε επιστημονικά περιοδικά. Τέλος, η κατηγορία Other, η οποία περιλαμβάνει τύπους όπως Reviews, Book Chapters, Editorials

και Notes, αντιπροσωπεύει μόλις το 3% του συνόλου, επιβεβαιώνοντας ότι αυτές οι μορφές δημοσίευσης δεν αποτελούν κυρίαρχες μορφές διάχυσης γνώσης στο πεδίο.

4.4 Κορυφαίες Χώρες και Ιδρύματα

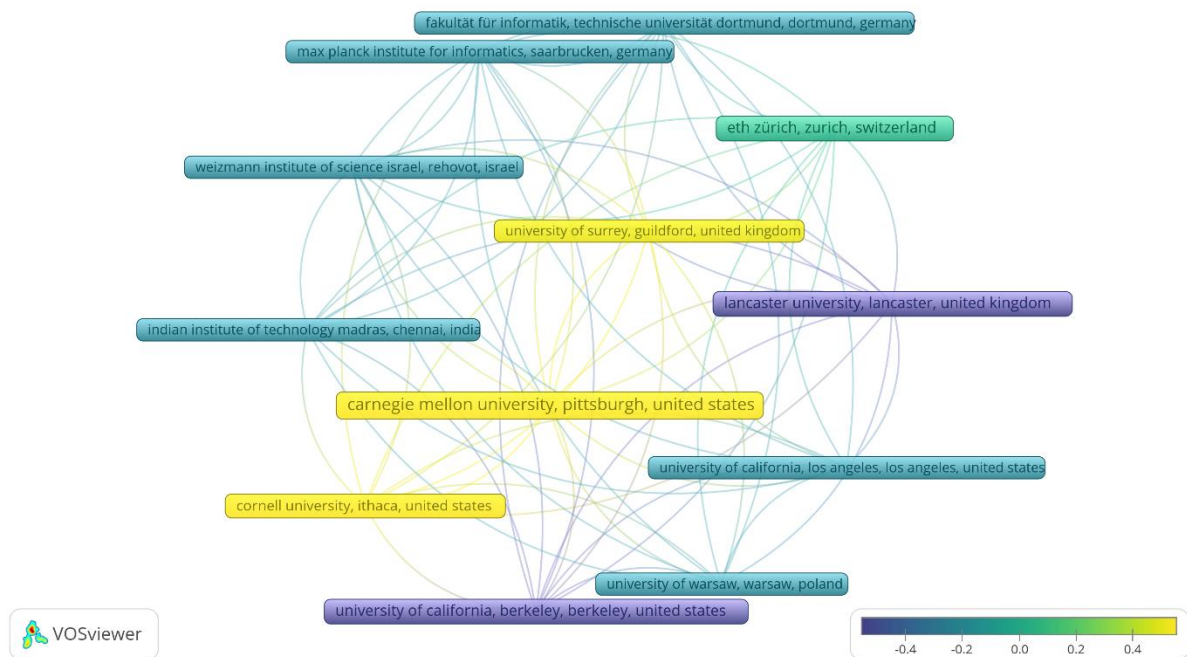


Γράφημα 4.4α: Χάρτης Κορυφαίων Χωρών

Η ανάλυση των κορυφαίων χωρών και ιδρυμάτων πραγματοποιήθηκε με τη χρήση του εργαλείου VOSviewer. Ως είσοδος χρησιμοποιήθηκε το φιλτραρισμένο CSV αρχείο που προέκυψε από την επεξεργασία μέσω της χρήσης R, το οποίο περιείχε τα μεταδεδομένα των 2.962 δημοσιεύσεων του τελικού συνόλου δεδομένων. Στο VOSviewer επιλέχθηκε η ανάλυση τύπου Co-authorship με μονάδα ανάλυσης τις χώρες (Countries), η οποία εξετάζει τη συχνότητα συνεργασίας μεταξύ ερευνητών από διαφορετικές χώρες βάσει κοινών δημοσιεύσεων.

Όπως φαίνεται στο *Γράφημα 4.4α*, ο χάρτης χωρών παρουσιάζεται με τη μορφή density map, όπου το έντονο κίτρινο χρώμα υποδηλώνει υψηλή συγκέντρωση δημοσιεύσεων, ενώ το μπλε αντιστοιχεί σε χαμηλότερη δραστηριότητα. Οι τρεις κυρίαρχοι κόμβοι είναι οι Ηνωμένες Πολιτείες, η Κίνα και η Γερμανία, οι οποίες εμφανίζονται με έντονο κίτρινο χρώμα, υποδηλώνοντας τον πολύ υψηλό αριθμό δημοσιεύσεών τους. Ένας δεύτερος σημαντικός κόμβος αποτελείται από ευρωπαϊκές και λατινοαμερικάνικες χώρες όπως η Ισπανία, η Ιταλία, η Βραζιλία και ο Καναδάς, οι οποίες εμφανίζονται με πράσινο-κίτρινο χρώμα, υποδηλώνοντας μέτρια αλλά αξιοσημείωτη παρουσία. Σε χαμηλότερο επίπεδο

δραστηριότητας, με ανοιχτό πράσινο-μπλε χρώμα, εμφανίζονται χώρες όπως το Ηνωμένο Βασίλειο, η Γαλλία, η Ινδία, η Ιαπωνία, η Αυστραλία και η Ελβετία, αναδεικνύοντας την παγκόσμια διάσταση της έρευνας στην επαναχρησιμοποίηση λογισμικού.



Γράφημα 4.4β: Δίκτυο Κορυφαίων Ιδρυμάτων

Όσον αφορά την ανάλυση ιδρυμάτων, για το Γράφημα 4.4β επιλέχθηκε η ανάλυση τύπου Co-authorship με μονάδα ανάλυσης τα ιδρύματα (Organizations), παράγοντας ένα δίκτυο συνεργασιών μεταξύ ιδρυμάτων. Κάθε κόμβος αντιπροσωπεύει ένα ίδρυμα, ενώ οι γραμμές που τους συνδέουν υποδηλώνουν κοινές δημοσιεύσεις και συνεργατικά ερευνητικά έργα. Το υπόμνημα στο κάτω δεξί μέρος του γραφήματος αντιστοιχεί στον αριθμό των συνδέσεων κάθε ιδρύματος, με το κίτρινο-πράσινο χρώμα να υποδηλώνει τα ιδρύματα με τις περισσότερες συνεργασίες και το σκούρο μπλε αυτά με τις λιγότερες. Με βάση αυτή την κλίμακα, τα τρία ιδρύματα με τις περισσότερες διεθνείς συνεργασίες είναι το Carnegie Mellon University (Pittsburgh, ΗΠΑ), το Cornell University (Ithaca, ΗΠΑ) και το University of Surrey (Guildford, Ηνωμένο Βασίλειο), τα οποία εμφανίζονται με το πιο έντονο κίτρινο χρώμα στο δίκτυο. Από την άλλη πλευρά, ιδρύματα όπως το ETH Zürich, το Max Planck Institute for Informatics, το Weizmann Institute of Science, το Indian Institute of Technology Madras και το Lancaster University εμφανίζονται με πιο σκούρο χρώμα, υποδηλώνοντας λιγότερες αλλά εξειδικευμένες συνεργασίες. Συνολικά, το δίκτυο αποκαλύπτει ότι τα κορυφαία ιδρύματα δεν λειτουργούν απομονωμένα, αλλά συμμετέχουν σε πυκνά διασυνδεδεμένα δίκτυα διεθνούς συνεργασίας.

Top 20 Affiliations (by Number of Documents)

Rank	Affiliation	Documents
1	Peking University, Beijing, China	27
2	Johannes Kepler University Linz, Linz, Austria	26
3	Department of Information Systems, University of Haifa, Haifa, Israel	18
4	University of Chinese Academy of Sciences, Beijing, China	18
5	Fraunhofer Institute for Experimental Software Engineering IESE, Kaiserslauter...	17
6	Nanjing University, Nanjing, China	17
7	École de Technologie Supérieure, Montreal, Canada	16
8	Chalmers University of Technology, Gothenburg, Sweden	15
9	Department of Computer Science, Ames, United States	15
10	Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany	15
11	School of Electronics Engineering and Computer Science, Peking University, Be...	15
12	Technische Universität Braunschweig, Braunschweig, Germany	15
13	Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil	15
14	KU Leuven, Leuven, Belgium	14
15	Ministry of Education of the People's Republic of China, Beijing, China	14
16	Mälardalen University, Vasteras, Sweden	14
17	Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil	14
18	Technische Universität München, Munich, Germany	14
19	Carnegie Mellon University, Pittsburgh, United States	13
20	Key Lab of High Confidence Software Technologies, Ministry of Education, Beij...	13

Πίνακας 4.4γ: Κορυφαία Ιδρύματα

Ο Πίνακας 4.4γ παρουσιάζει τα 20 κορυφαία ιδρύματα βάσει του αριθμού δημοσιεύσεων στο τελικό σύνολο δεδομένων. Στην πρώτη θέση εμφανίζεται το Peking University, Beijing, China με 27 δημοσιεύσεις, ενώ ακολουθεί το Johannes Kepler University Linz, Linz, Austria με 26 δημοσιεύσεις. Στην τρίτη και τέταρτη θέση εμφανίζονται το Department of Information Systems, University of Haifa, Haifa, Israel και το University of Chinese Academy of Sciences, Beijing, China, με 18 δημοσιεύσεις το καθένα.

Στις επόμενες θέσεις καταγράφονται ιδρύματα από διαφορετικές γεωγραφικές περιοχές, όπως το Fraunhofer Institute for Experimental Software Engineering IESE στη Γερμανία, το Nanjing University στην Κίνα, το École de Technologie Supérieure στον Καναδά, το Chalmers University of Technology στη Σουηδία, το Carnegie Mellon University στις Ηνωμένες Πολιτείες και πανεπιστήμια από τη Βραζιλία και το Βέλγιο. Η παρουσία ιδρυμάτων από Ασία, Ευρώπη, Βόρεια Αμερική και Λατινική Αμερική δείχνει ότι η ερευνητική παραγωγή στο πεδίο της επαναχρησιμοποίησης λογισμικού κατανέμεται σε διαφορετικές γεωγραφικές περιοχές.

4.5 Κορυφαίοι Συγγραφείς

Top 20 Authors (by number of documents)

Rank	Author	Documents
1	Wang, Y.	28
2	Xie, B.	22
3	Li, Y.	21
4	Reinhartz-Berger, I.	21
5	Zhang, H.	21
6	Zhang, Y.	20
7	De Almeida, E.S.	19
8	Liu, X.	19
9	Zhang, L.	19
10	Assunção, W.K.G.	18
11	Wang, Z.	18
12	Liu, Y.	17
13	Zou, Y.	17
14	Wang, X.	16
15	Egyed, A.	15
16	Martinez, J.	15
17	Werner, C.	15
18	Ampatzoglou, A.	14
19	Berger, T.	14
20	Chen, X.	14

Πίνακας 4.5: Κορυφαίοι Συγγραφείς

Ο Πίνακας 4.5 παρουσιάζει τους 20 πιο παραγωγικούς συγγραφείς στον τομέα της επαναχρησιμοποίησης λογισμικού, βάσει του αριθμού δημοσιεύσεών τους στο τελικό σύνολο δεδομένων. Στην πρώτη θέση βρίσκεται η Yasha Wang με 28 δημοσιεύσεις, ενώ ακολουθεί ο Bing Xie με 22 δημοσιεύσεις. Στην τρίτη θέση εμφανίζονται οι Ying Li, Iris Reinhartz-Berger και Hongyu Zhang με 21 δημοσιεύσεις ο καθένας, ενώ ακολουθεί ο Yu Zhang με 20 δημοσιεύσεις.

Στις επόμενες θέσεις καταγράφονται οι Eduardo Santana de Almeida, Xiao Liu, Lu Zhang, Wesley Klewerton Guez Assunção, Zhi Wang, Yang Liu, Ying Zou, Xiaofang Wang, Alexander Egyed, Jabier Martinez, Cláudia Werner, Apostolos Ampatzoglou, Thorsten Berger και Xiang Chen, με αριθμό δημοσιεύσεων που κυμαίνεται από 14 έως 19. Συνολικά, ο αριθμός δημοσιεύσεων των 20 κορυφαίων συγγραφέων κυμαίνεται από 14 έως 28, γεγονός που δείχνει ότι η ερευνητική παραγωγή δεν συγκεντρώνεται σε έναν μόνο συγγραφέα, αλλά κατανέμεται σε περισσότερους ερευνητές με σταθερή παρουσία στο πεδίο.

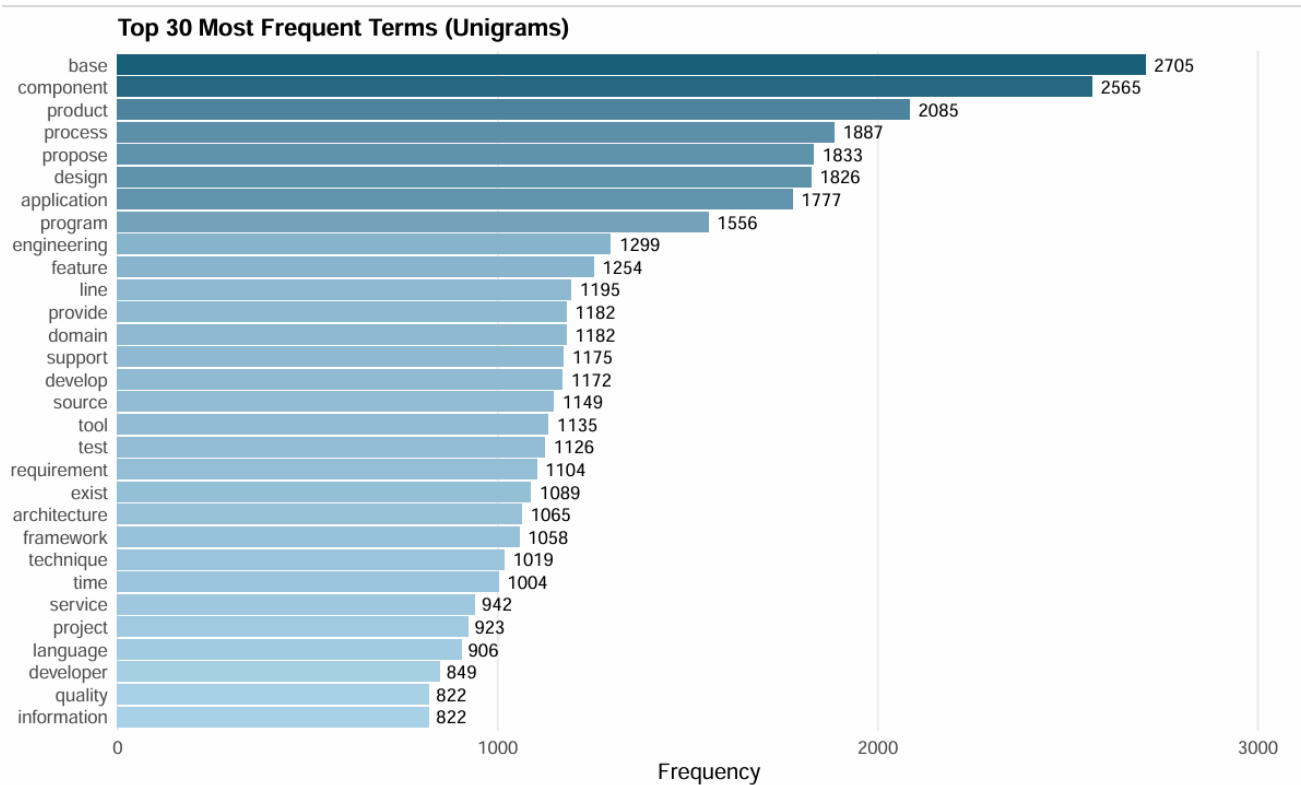
4.6 Ανάλυση Λέξεων-Κλειδιών

Η ανάλυση λέξεων-κλειδιών αποτελεί βασικό στάδιο τόσο της βιβλιομετρικής ανάλυσης όσο και της εξόρυξης κειμένου, καθώς επιτρέπει την ποσοτική αποτύπωση της θεματολογίας μιας βιβλιογραφίας και την ανάδειξη των κυρίαρχων εννοιών ενός ερευνητικού πεδίου. Στο πλαίσιο της παρούσας εργασίας, εφαρμόστηκαν συμπληρωματικές τεχνικές με στόχο την ανάδειξη των συχνότερων και σημαντικότερων όρων που εμφανίζονται στο σύνολο των δημοσιεύσεων. Συγκεκριμένα, η ανάλυση περιλαμβάνει τη μελέτη μονολεκτικών όρων (unigrams), καθώς και συνδυασμών δύο και τριών λέξεων (bigrams και trigrams), οι οποίοι επιτρέπουν την αποτύπωση πιο σύνθετων εννοιών.

Παράλληλα, χρησιμοποιήθηκε η μέθοδος TF-IDF για την ανάδειξη όρων με ιδιαίτερη σημασία σε επιμέρους δημοσιεύσεις, ενώ η ανάλυση συν-εμφάνισης λέξεων-κλειδιών μέσω του εργαλείου VOSviewer παρέχει μία οπτική αναπαράσταση των σχέσεων μεταξύ των βασικών εννοιών του πεδίου. Μέσω των επιμέρους αυτών προσεγγίσεων, επιδιώκεται η συνολική αποτύπωση της δομής και των κύριων θεματικών κατευθύνσεων της βιβλιογραφίας.

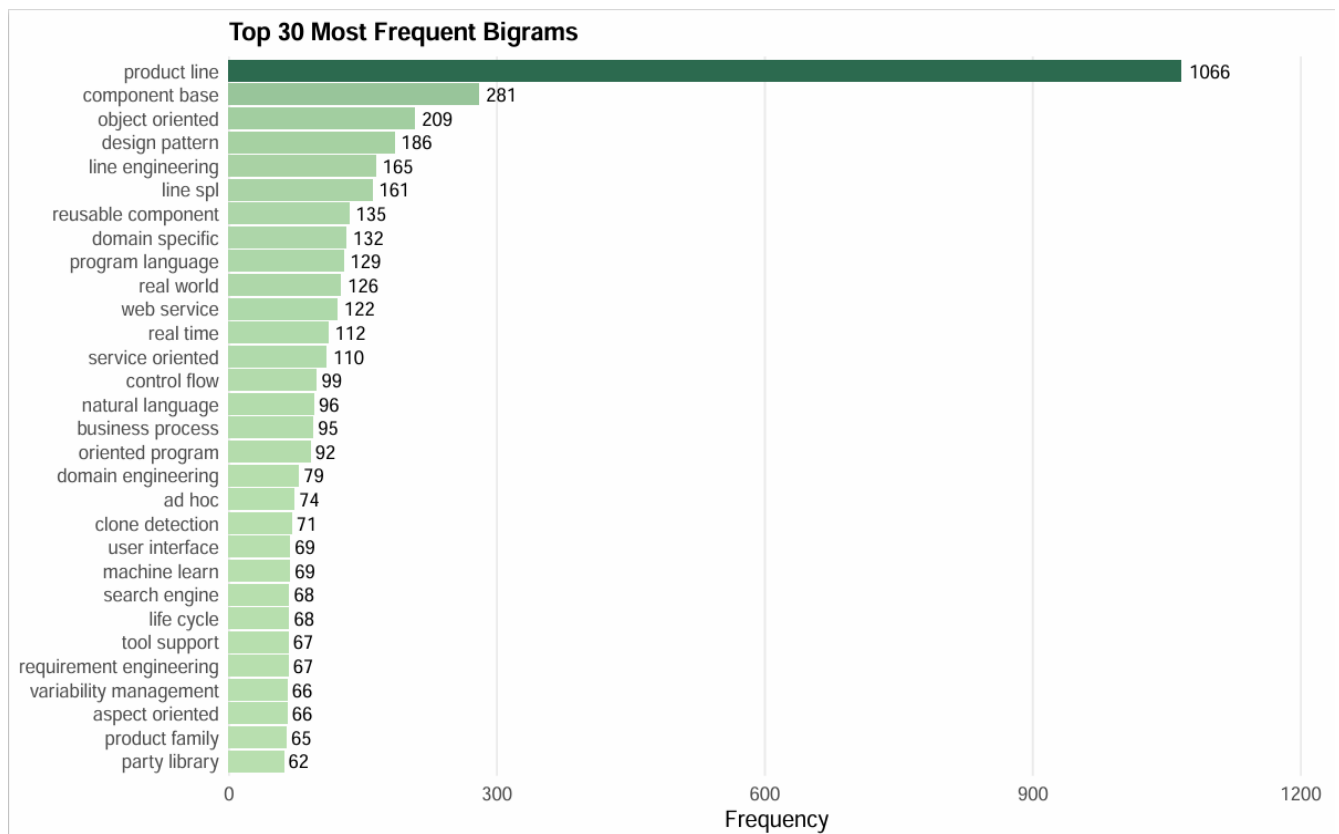
4.6.1 Πιο Συχνοί Όροι (N-Grams)

Η ανάλυση N-Grams πραγματοποιήθηκε σε τρία επίπεδα — unigrams, bigrams και trigrams — στις περιλήψεις των 2,962 δημοσιεύσεων, μετά την εφαρμογή λεμματοποίησης και αφαίρεσης λέξεων διακοπής.



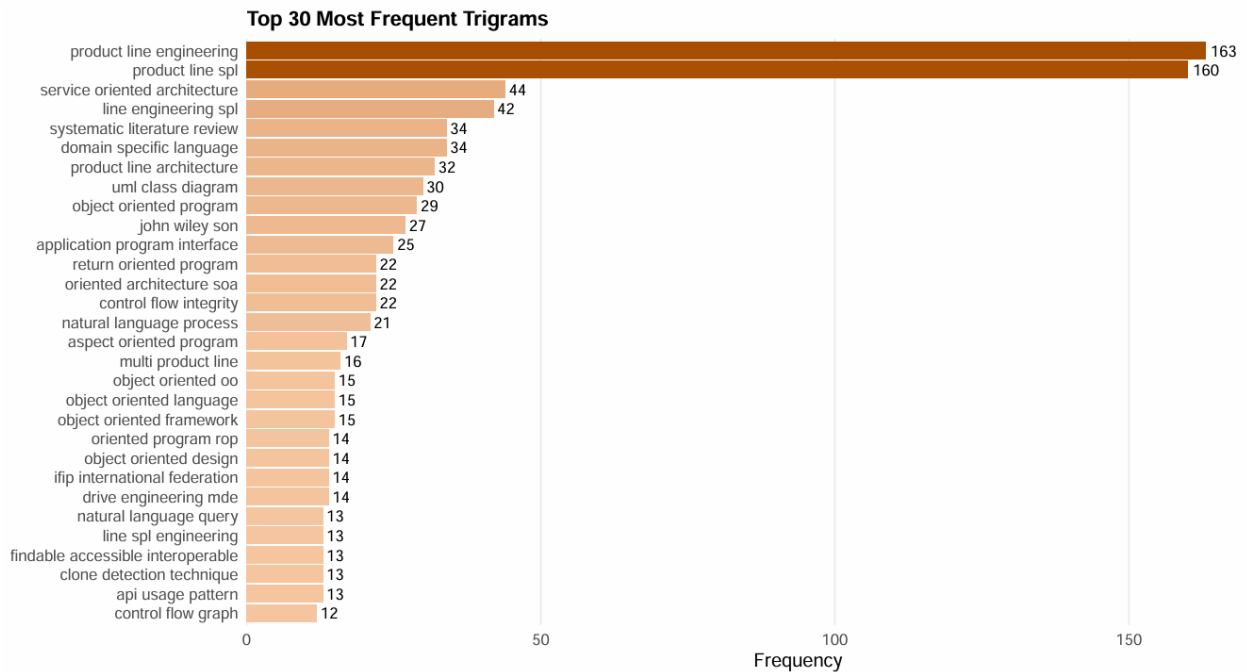
Γράφημα 4.6.1α: Unigrams

Όσον αφορά τα unigrams, ο πιο συχνά εμφανιζόμενος όρος είναι ο “base” με 2.705 εμφανίσεις, ακολουθούμενος από τους “component” (2.565) και “product” (2.085). Στις επόμενες θέσεις εμφανίζονται οι όροι “process” (1.887), “propose” (1.833), “design” (1.826) και “application” (1.777). Επιπλέον, παρατηρείται η παρουσία των όρων “program” (1.556), “engineering” (1.299) και “feature” (1.254), οι οποίοι σχετίζονται με βασικές έννοιες της ανάπτυξης λογισμικού και της επαναχρησιμοποίησης. Τέλος, οι όροι “line” (1.195), “domain” (1.182), “provide” (1.182), “support” (1.175) και “develop” (1.172) συμπληρώνουν τη λίστα των 15 πιο συχνών μονολεκτικών όρων.



Γράφημα 4.6.1β: Bigrams

Στα bigrams, ο πιο συχνός συνδυασμός είναι το "product line" με 1.066 εμφανίσεις, υπερέχοντας σημαντικά έναντι των υπολοίπων. Ακολουθούν τα "component base" (281) και "object oriented" (209), ενώ σημαντική παρουσία εμφανίζουν και τα "design pattern" (186), "line engineering" (165) και "line spl" (161). Στις επόμενες θέσεις εμφανίζονται οι συνδυασμοί "reusable component" (135), "domain specific" (132), "program language" (129) και "real world" (126). Συμπληρώνουν τη λίστα τα "web service" (122), "real time" (112), "service oriented" (110), "control flow" (99) και "natural language" (96).



Γράφημα 4.6.1γ: Trigrams

Στα trigrams, κυριαρχεί η φράση "product line engineering" με 163 εμφανίσεις, ακολουθούμενη στενά από το "product line spl" (160). Στις επόμενες θέσεις εμφανίζονται τα "service oriented architecture" (44), "line engineering spl" (42) και "systematic literature review" (34). Αξιοσημείωτη είναι η παρουσία των "domain specific language" (34), "product line architecture" (32) και "uml class diagram" (30). Συμπληρώνουν τη λίστα τα trigrams "object oriented program" (29), "john wiley son" (27), "application program interface" (25), "return oriented program" (22), "oriented architecture soa" (22), "control flow integrity" (22) και "natural language process" (21).

Συνολικά, η ανάλυση N-Grams σε όλα τα επίπεδα αναδεικνύει με σαφήνεια την κυριαρχία των Software Product Lines στη βιβλιογραφία, καθώς οι όροι "product line", "product line engineering" και "product line spl" εμφανίζονται με εξαιρετικά υψηλή συχνότητα σε όλα τα επίπεδα ανάλυσης.

4.6.2 Ανάλυση TF-IDF

Top 30 Terms by TF-IDF Score

Rank	Term	TF-IDF Score	Documents
1	component	50.1147	806
2	product	43.8438	672
3	test	30.6584	339
4	process	30.0339	931
5	line	28.9181	536
6	design	28.7936	916
7	feature	28.4535	521
8	service	28.374	294
9	program	28.3177	710
10	application	27.6956	910
11	domain	27.5857	606
12	requirement	27.0729	534
13	pattern	26.7286	294
14	spl	26.6293	217
15	abstract	26.1169	118
16	architecture	24.8158	568
17	framework	24.6152	537
18	engineering	24.6131	780
19	base	24.0755	1534
20	tool	23.7203	657
21	source	23.5254	630
22	variability	23.3529	268
23	project	23.1557	450
24	language	21.9431	485
25	propose	21.5015	1242
26	support	21.0509	811
27	technique	20.845	634
28	develop	20.8365	856
29	quality	19.8571	510
30	knowledge	19.7727	335

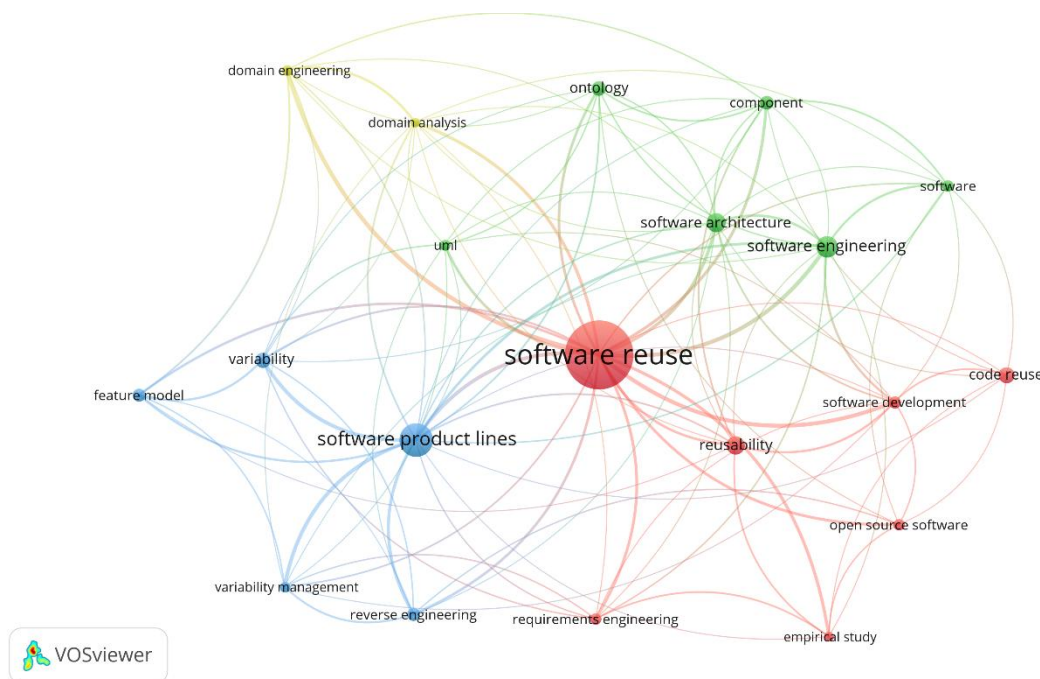
Πίνακας 4.6.2: TF-IDF

Ο Πίνακας 4.6.2 παρουσιάζει τους 30 πιο σημαντικούς όρους βάσει της συνολικής βαθμολογίας TF-IDF στο σύνολο των δημοσιεύσεων. Ο όρος με την υψηλότερη βαθμολογία TF-IDF είναι ο "component" με βαθμολογία 50.11, εμφανιζόμενος σε 806 έγγραφα, γεγονός που υποδηλώνει ότι αποτελεί έναν από τους πιο χαρακτηριστικούς και διακριτικούς όρους

της βιβλιογραφίας. Ακολουθεί ο όρος "product" με βαθμολογία 43.84 και παρουσία σε 672 έγγραφα, ενώ στην τρίτη θέση εμφανίζεται ο όρος "test" με βαθμολογία 30.66 και παρουσία σε 339 έγγραφα.

Στις επόμενες θέσεις εμφανίζονται όροι όπως "process" (30.03, 931 έγγραφα), "line" (28.92, 536 έγγραφα), "design" (28.79, 916 έγγραφα) και "feature" (28.45, 521 έγγραφα), οι οποίοι αντικατοπτρίζουν τη σημασία των διαδικασιών ανάπτυξης και των χαρακτηριστικών στη βιβλιογραφία. Αξιοσημείωτη είναι η παρουσία του όρου "spl" στη 14η θέση με βαθμολογία 26.63, ο οποίος παρά την εμφάνισή του σε μόλις 217 έγγραφα παρουσιάζει υψηλή βαθμολογία TF-IDF, υποδηλώνοντας ότι είναι ιδιαίτερα χαρακτηριστικός για τα έγγραφα στα οποία εμφανίζεται. Τέλος, όροι όπως "service" (28.37, 294 έγγραφα), "domain" (27.59, 606 έγγραφα), "requirement" (27.07, 534 έγγραφα), "pattern" (26.73, 294 έγγραφα) και "abstract" (26.12, 118 έγγραφα) συμπληρώνουν τη λίστα, αναδεικνύοντας τη σημασία των υπηρεσιών, της ανάλυσης πεδίου, των απαιτήσεων και των σχεδιαστικών προτύπων στο πεδίο.

4.6.3 Χάρτης Συν-Εμφάνισης Λέξεων-Κλειδιών



Γράφημα 4.6.3: Χάρτης Συν-Εμφάνισης Λέξεων Κλειδιών

Η ανάλυση συν-εμφάνισης λέξεων-κλειδιών πραγματοποιήθηκε με τη χρήση του εργαλείου VOSviewer, αξιοποιώντας ως είσοδο το φιλτραρισμένο CSV αρχείο που προέκυψε από την επεξεργασία των δεδομένων Scopus μέσω R. Τα αποτελέσματα παρουσιάζονται στο Γράφημα 4.6.3. Στο δίκτυο, κάθε κόμβος αντιπροσωπεύει μια λέξη-κλειδί τόσο των συγγραφέων όσο και των ευρετηριασμένων λέξεων-κλειδιών, το μέγεθος του κόμβου

αντικατοπτρίζει τη συχνότητα εμφάνισής της στη βιβλιογραφία, ενώ οι γραμμές που συνδέουν τους κόμβους υποδηλώνουν τη συχνότητα συν-εμφάνισής τους σε κοινές δημοσιεύσεις. Τα διαφορετικά χρώματα αντιστοιχούν σε διακριτές ομάδες συσχετιζόμενων λέξεων-κλειδιών, οι οποίες εντοπίστηκαν αυτόματα από τον αλγόριθμο του VOSviewer χωρίς καμία χειροκίνητη παρέμβαση, βάσει της συχνότητας συν-εμφάνισης των λέξεων-κλειδιών μεταξύ τους.

Ο κεντρικός και μεγαλύτερος κόμβος του δικτύου είναι ο όρος "software reuse", ο οποίος εμφανίζεται με κόκκινο χρώμα και συνδέεται με σχεδόν όλους τους υπόλοιπους κόμβους, επιβεβαιώνοντας τον κεντρικό του ρόλο στη βιβλιογραφία. Ο δεύτερος μεγαλύτερος κόμβος είναι ο όρος "software product lines", ο οποίος εμφανίζεται με μπλε χρώμα και αποτελεί τον κεντρικό κόμβο μιας ξεχωριστής ομάδας.

Η μπλε ομάδα περιλαμβάνει όρους που σχετίζονται με τα Software Product Lines και τη διαχείριση παραλλαγών, όπως "variability", "feature model", "variability management" και "reverse engineering". Η ομάδα αυτή εμφανίζεται στο αριστερό τμήμα του γραφήματος και παρουσιάζει πυκνές εσωτερικές συνδέσεις μεταξύ των κόμβων της.

Η πράσινη ομάδα εμφανίζεται στο επάνω δεξί τμήμα του γραφήματος και περιλαμβάνει όρους που σχετίζονται με την αρχιτεκτονική λογισμικού και τη μηχανική λογισμικού, συμπεριλαμβανομένων των "software engineering", "software architecture", "component", "software" και "ontology".

Η κόκκινη ομάδα γύρω από τον κεντρικό κόμβο περιλαμβάνει όρους όπως "reusability", "software development", "code reuse", "open source software", "requirements engineering" και "empirical study", αναδεικνύοντας τη στενή σχέση μεταξύ της επαναχρησιμοποίησης λογισμικού και πρακτικών πτυχών της ανάπτυξης λογισμικού.

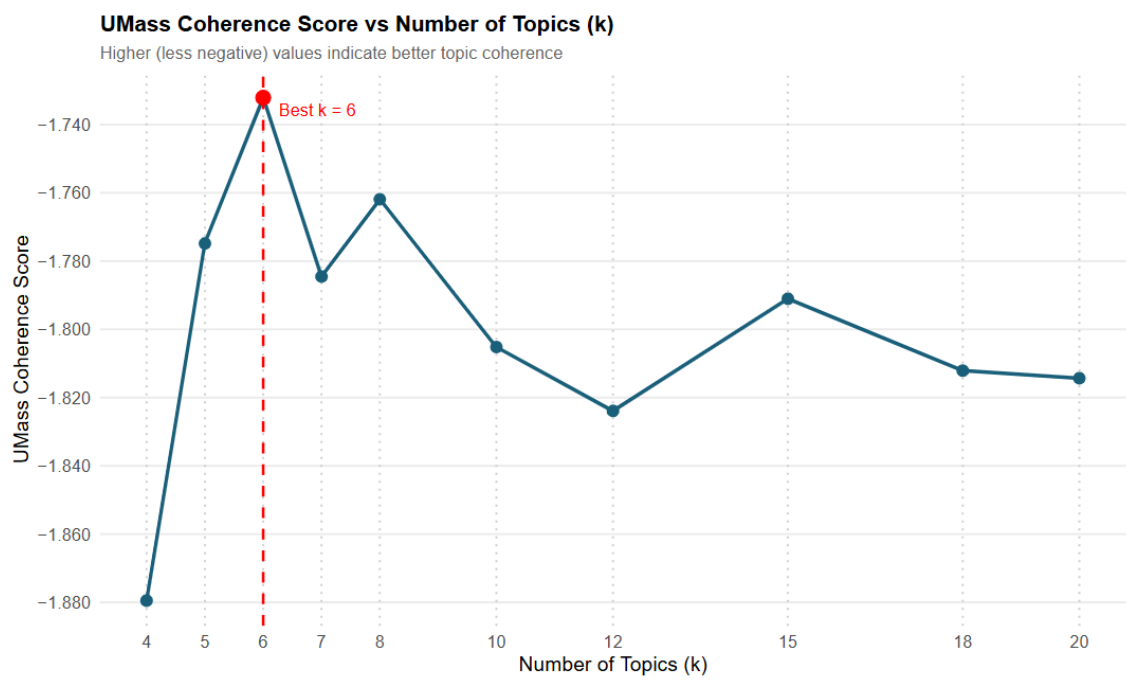
Η κίτρινη ομάδα στο επάνω αριστερό τμήμα περιλαμβάνει όρους όπως "domain engineering" και "domain analysis", ενώ μεταξύ των ομάδων εμφανίζονται επίσης οι όροι "uml" και "requirements engineering", οι οποίοι λειτουργούν ως γέφυρες μεταξύ διαφορετικών θεματικών περιοχών.

Συνολικά, το δίκτυο οργανώνεται σε τέσσερις κύριες ομάδες με πυκνές εσωτερικές συνδέσεις, ενώ παράλληλα υπάρχουν σημαντικές διασυνδέσεις μεταξύ των ομάδων, υποδηλώνοντας ότι οι θεματικές περιοχές δεν είναι απομονωμένες αλλά αλληλοεπικαλύπτονται σε σημαντικό βαθμό.

4.7 Αποτελέσματα Μοντελοποίησης Θεμάτων

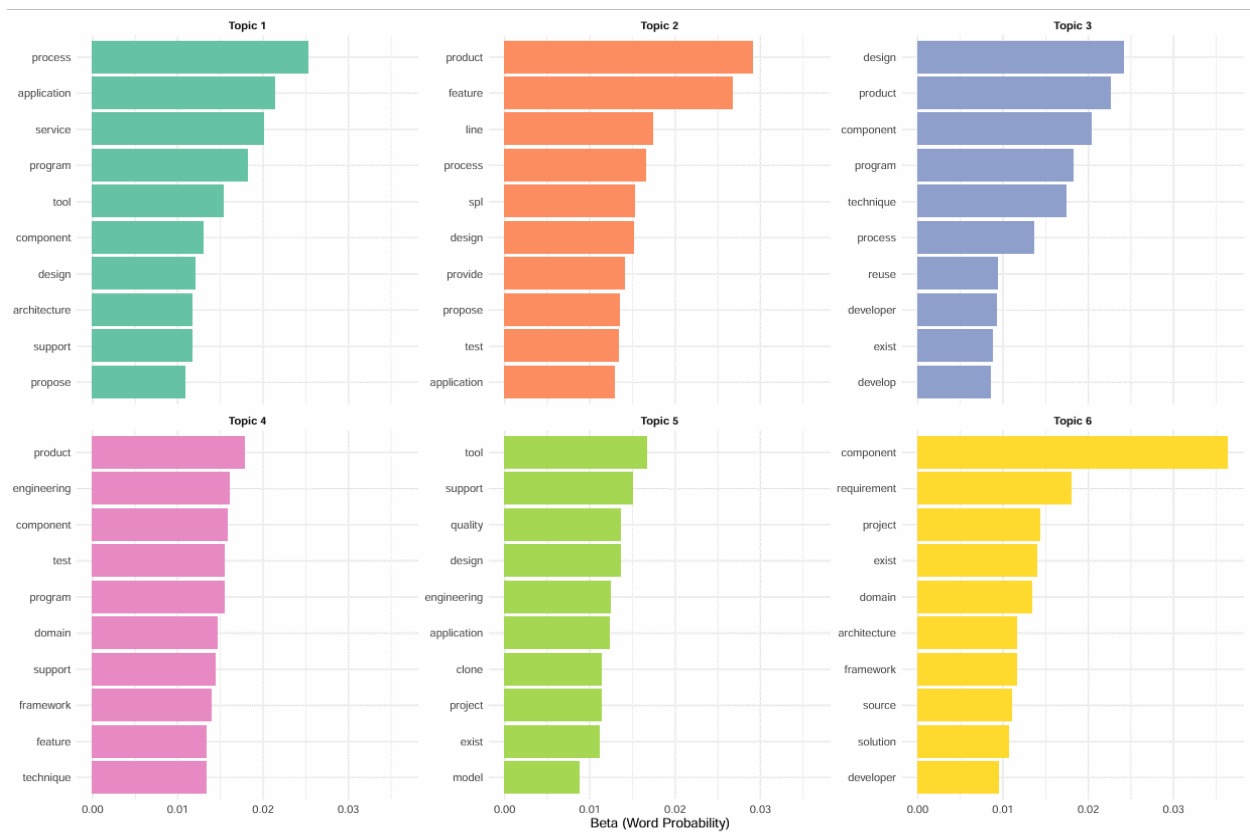
4.7.1 Αναγνωρισμένα Θέματα (LDA)

Η επιλογή του αριθμού των θεμάτων (k) πραγματοποιήθηκε βάσει του UMass Coherence Score, ο οποίος αξιολογεί την ποιότητα των θεμάτων μετρώντας την εσωτερική συνοχή των λέξεων που ανήκουν σε κάθε θέμα. Δοκιμάστηκαν διαφορετικές τιμές του k και επιλέχθηκε η τιμή που απέδωσε την υψηλότερη βαθμολογία coherence, υποδηλώνοντας τη βέλτιστη θεματική διαχωριστικότητα. Το αντίστοιχο γράφημα coherence score παρουσιάζεται στο *Γράφημα 4.7.1α*, όπου φαίνεται καθαρά η κορύφωση στο $k = 6$.



Γράφημα 4.7.1α: Δείκτης Καλύτερου Αριθμού Θεμάτων

Η εφαρμογή του αλγορίθμου LDA με $k = 6$ οδήγησε στην εξαγωγή έξι διακριτών θεματικών περιοχών, καθεμία από τις οποίες περιγράφεται από ένα σύνολο 10 χαρακτηριστικών όρων. Οι όροι αυτοί αντιστοιχούν στις λέξεις με τη μεγαλύτερη πιθανότητα εμφάνισης σε κάθε θέμα και χρησιμοποιούνται για την ερμηνεία του περιεχομένου του. Η διαδικασία ονομασίας των θεμάτων πραγματοποιήθηκε χειροκίνητα, μέσω της ανάλυσης των σημαντικότερων όρων κάθε θέματος και της αντιστοίχισής τους σε ευρύτερες θεματικές έννοιες. Τα έξι θέματα που προέκυψαν συνοψίζονται ως εξής στο *Γράφημα 4.7.1β*:



Γράφημα 4.7.1β: Αναγνωρισμένα Θέματα

Topic 1: Software Processes & Services. Το πρώτο θέμα περιλαμβάνει όρους όπως *process*, *application*, *service*, *program*, *tool*, *component*, *design*, *architecture* και *support*. Οι όροι αυτοί υποδηλώνουν έμφαση στις διαδικασίες ανάπτυξης λογισμικού, στην υποστήριξη εφαρμογών και στην αξιοποίηση υπηρεσιών και αρχιτεκτονικών προσεγγίσεων για την επαναχρησιμοποίηση λογισμικού.

Topic 2: Software Product Lines. Το δεύτερο θέμα περιλαμβάνει όρους όπως *product*, *feature*, *line*, *process*, *spl*, *design*, *test* και *application*. Η ισχυρή παρουσία των όρων *product*, *feature*, *line* και *spl* υποδηλώνει σαφή σύνδεση με τα Software Product Lines και τη διαχείριση χαρακτηριστικών κατά την ανάπτυξη πολλαπλών προϊόντων μέσω κοινών επαναχρησιμοποιήσιμων στοιχείων.

Topic 3: Design Techniques. Το τρίτο θέμα περιλαμβάνει όρους όπως *design*, *product*, *component*, *program*, *technique*, *process*, *reuse*, *developer* και *develop*. Το συγκεκριμένο θέμα σχετίζεται με τεχνικές σχεδίασης, ανάπτυξης λογισμικού και πρακτικές επαναχρησιμοποίησης σε επίπεδο σχεδιασμού και υλοποίησης.

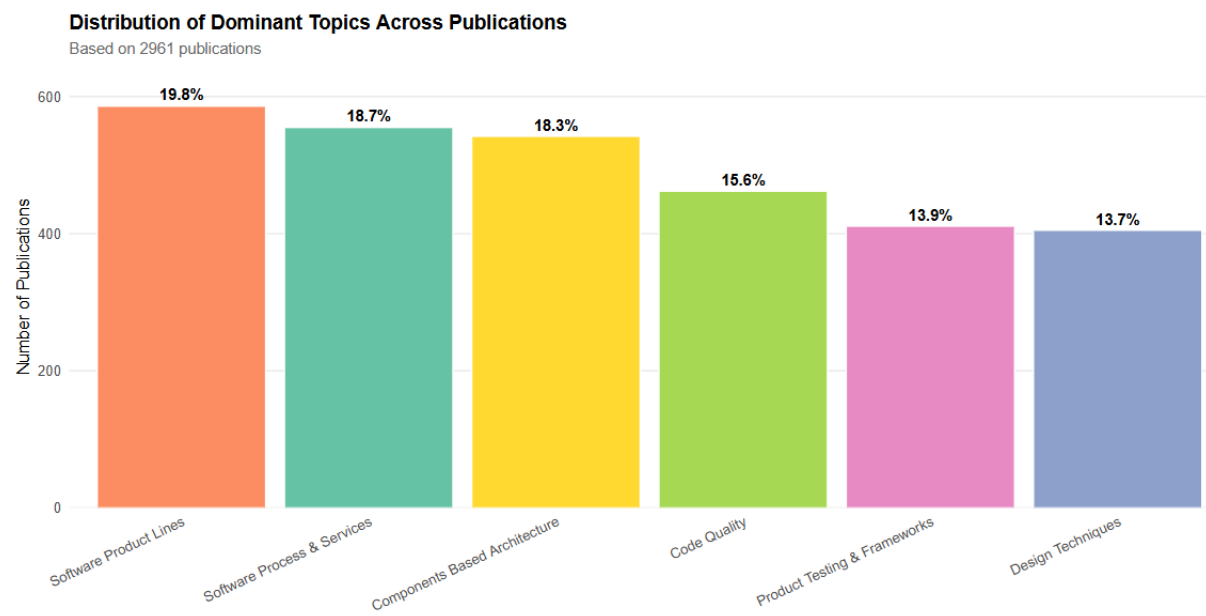
Topic 4: Product Testing & Frameworks. Το τέταρτο θέμα περιλαμβάνει όρους όπως *product*, *engineering*, *component*, *test*, *program*, *domain*, *framework*, *feature* και *technique*. Σε αντίθεση με το προηγούμενο label, αυτό το θέμα φαίνεται να σχετίζεται περισσότερο με

διαδικασίες μηχανικής προϊόντων, δοκιμές λογισμικού και τεχνικά πλαίσια ανάπτυξης που υποστηρίζουν την επαναχρησιμοποίηση.

Topic 5: Code Quality. Το πέμπτο θέμα περιλαμβάνει όρους όπως *tool*, *support*, *quality*, *design*, *engineering*, *application*, *clone*, *project* και *model*. Η παρουσία των όρων *quality* και *clone* δείχνει ότι το θέμα σχετίζεται περισσότερο με την αξιολόγηση ποιότητας λογισμικού, την ανάλυση κώδικα και εργαλεία υποστήριξης επαναχρησιμοποίησης.

Topic 6: Component Based Architecture. Το έκτο θέμα περιλαμβάνει όρους όπως *component*, *requirement*, *project*, *domain*, *architecture*, *framework*, *source*, *solution* και *developer*. Το θέμα αυτό εστιάζει στην ανάπτυξη βασισμένη σε συνιστώσες και στη χρήση αρχιτεκτονικών προσεγγίσεων που υποστηρίζουν την επαναχρησιμοποίηση λογισμικού μέσω δομημένων και επεκτάσιμων λύσεων.

4.7.2 Κατανομή Κυρίαρχων Θεμάτων



Γράφημα 4.7.2: Κατανομή Κυρίαρχων Θεμάτων

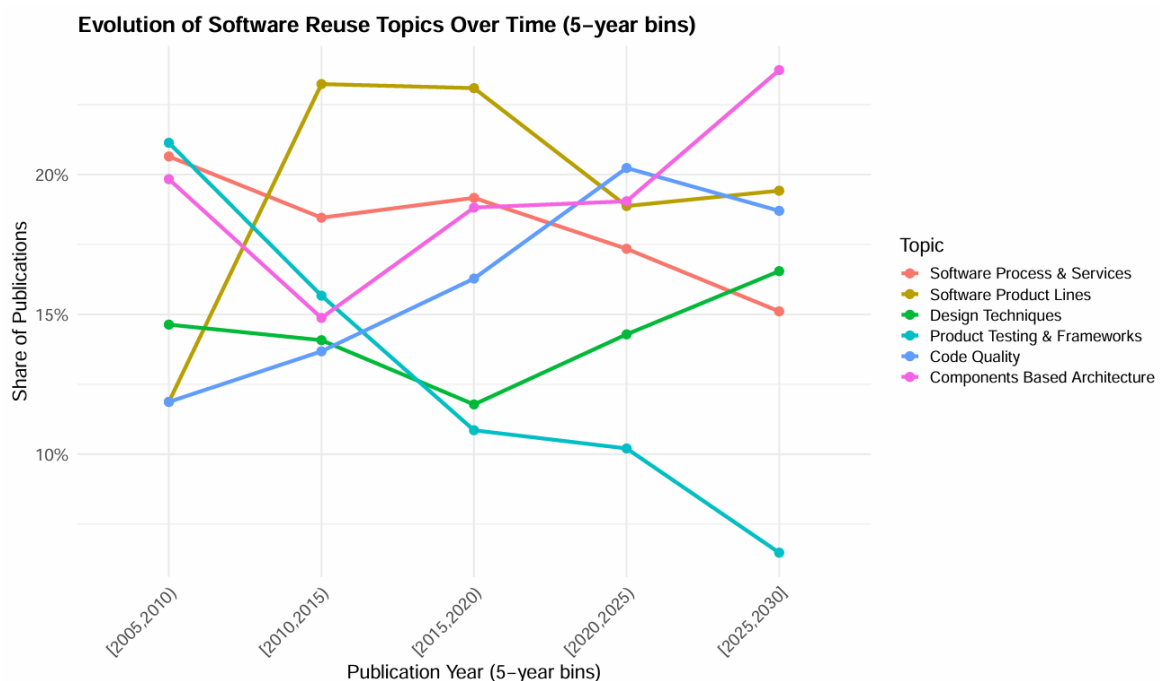
Το Γράφημα 4.7.2 παρουσιάζει την κατανομή των κυρίαρχων θεμάτων στο σύνολο των 2.962 δημοσιεύσεων, αναδεικνύοντας ποιο θέμα (topic) επικρατεί σε κάθε έγγραφο βάσει της υψηλότερης τιμής που υπολογίστηκε από τον αλγόριθμο LDA. Με άλλα λόγια, κάθε δημοσίευση ανατίθεται στο θέμα (topic) στο οποίο ανήκει με τη μεγαλύτερη πιθανότητα, επιτρέποντας έτσι μια συνολική εικόνα της θεματικής κατανομής της βιβλιογραφίας.

Το κυρίαρχο θέμα είναι τα Software Product Lines με 587 δημοσιεύσεις, αντιστοιχώντας στο 19.8% του συνόλου, επιβεβαιώνοντας για ακόμα μια φορά την κυριαρχία αυτής της

θεματικής περιοχής στη βιβλιογραφία της επαναχρησιμοποίησης λογισμικού. Στη δεύτερη θέση εμφανίζεται το Software Processes & Services με 554 δημοσιεύσεις (18.7%), ενώ στην τρίτη θέση ακολουθεί το Component Based Architecture με 542 δημοσιεύσεις (18.3%). Αξιοσημείωτο είναι ότι τα τρία πρώτα topics συγκεντρώνουν συνολικά περίπου το 56.8% του συνόλου των δημοσιεύσεων, υποδηλώνοντας ότι αυτές οι τρεις θεματικές περιοχές αποτελούν τους κυρίαρχους άξονες της έρευνας στο πεδίο.

Στις επόμενες θέσεις ακολουθεί το Code Quality & Tools με 462 δημοσιεύσεις (15.6%), το Software Testing Frameworks με 412 δημοσιεύσεις (13.9%) και τέλος το Design Techniques με 406 δημοσιεύσεις (13.7%). Είναι αξιοσημείωτο ότι παρά τις διαφορές στην κατάταξη, η κατανομή μεταξύ των έξι θεμάτων είναι σχετικά ισορροπημένη, με τη διαφορά μεταξύ του πρώτου και του τελευταίου topic να ανέρχεται μόλις στο 6.1%, γεγονός που υποδηλώνει ότι όλες οι θεματικές περιοχές έχουν σημαντική παρουσία στη βιβλιογραφία.

4.7.3 Εξέλιξη Θεμάτων στον Χρόνο



Γράφημα 4.7.3: Εξέλιξη Θεμάτων στο Χρόνο

Για την ανάλυση της εξέλιξης των θεμάτων στον χρόνο, οι δημοσιεύσεις ομαδοποιήθηκαν σε χρονικές περιόδους πενταετίας και υπολογίστηκε το ποσοστό συμμετοχής κάθε θέματος σε κάθε περίοδο. Το Γράφημα 4.7.3 αποτυπώνει τη δυναμική μεταβολή των έξι θεματικών

περιοχών από το 2005 έως το 2025, με τον άξονα Υ να παρουσιάζει το ποσοστό δημοσιεύσεων που ανατίθεται σε κάθε θέμα ανά πενταετία και τον άξονα Χ να αντιστοιχεί στις πέντε χρονικές περιόδους.

Το Software Product Lines (κίτρινο) εμφανίζει την πιο εντυπωσιακή μεταβολή μεταξύ όλων των θεμάτων. Ξεκινά από χαμηλά επίπεδα την περίοδο [2005,2010), ανερχόμενο απότομα στην περίοδο [2010,2015) όπου κορυφώνεται με ποσοστό περίπου 23%, το υψηλότερο που καταγράφεται για οποιοδήποτε θέμα στη συγκεκριμένη περίοδο. Διατηρεί υψηλά επίπεδα και στην επόμενη περίοδο [2015,2020) με ποσοστό επίσης περίπου 23%, ενώ στις πιο πρόσφατες περιόδους [2020,2025) και [2025,2030] παρουσιάζει σταδιακή μείωση, παραμένοντας ωστόσο σε σχετικά υψηλά επίπεδα γύρω στο 19%.

Το Components Based Architecture (ροζ) ξεκινά με ισχυρή παρουσία γύρω στο 20% κατά την περίοδο [2005,2010), ακολουθεί μια ήπια πτώση στις περιόδους [2010,2015) και [2015,2020) φτάνοντας στο 15%, και στη συνέχεια εμφανίζει σταδιακή ανάκαμψη στην περίοδο [2020,2025). Η πιο αξιοσημείωτη εξέλιξη του θέματος αυτού είναι η έντονη ανοδική τάση στην τελευταία περίοδο [2025,2030], όπου φτάνει το υψηλότερο ποσοστό όλων των θεμάτων με περίπου 24%, αναδεικνύοντάς το ως το κυρίαρχο θέμα της πιο πρόσφατης περιόδου.

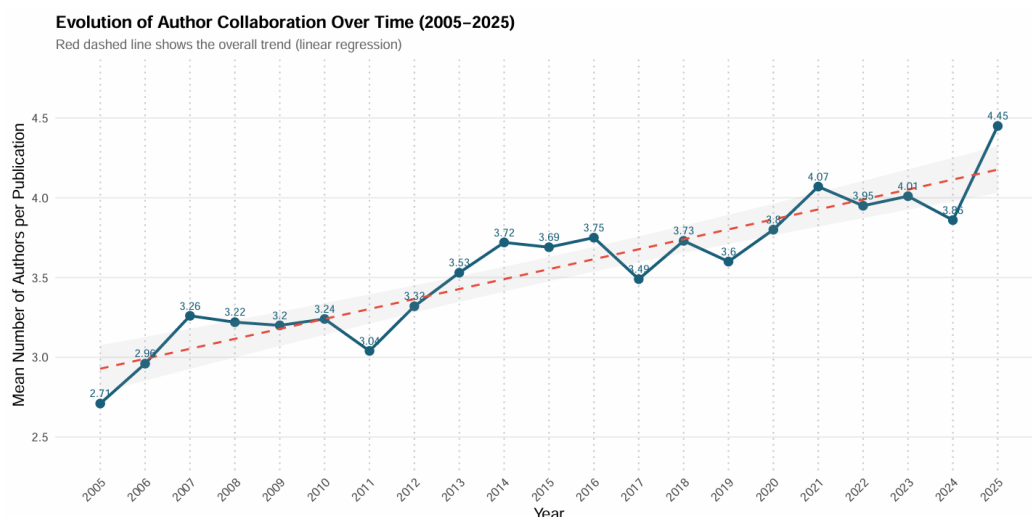
Το Software Process & Services (κόκκινο) παρουσιάζει τη μεγαλύτερη σταθερότητα από όλα τα θέματα, διατηρώντας σχετικά σταθερή παρουσία καθ' όλη τη διάρκεια της εξεταζόμενης περιόδου. Ξεκινά από περίπου 21% την περίοδο [2005,2010), κυμαίνεται μεταξύ 18% και 19% στις μεσαίες περιόδους και παρουσιάζει μια ήπια καθοδική τάση στις τελευταίες περιόδους, κατεβαίνοντας στο 15% στην περίοδο [2025,2030].

Το Code Quality (μπλε) ξεκινά από τα χαμηλότερα επίπεδά του γύρω στο 12% κατά την περίοδο [2005,2010) και παρουσιάζει σταδιακή αλλά σταθερή αύξηση στις επόμενες περιόδους. Φτάνει στο υψηλότερο σημείο του περίπου 20% κατά την περίοδο [2020,2025), ακολουθούμενη από μια ήπια μείωση στην τελευταία περίοδο όπου σταθεροποιείται γύρω στο 19%.

Το Design Techniques (πράσινο) εμφανίζει σταθερά χαμηλά επίπεδα κατά τις πρώτες περιόδους, κυμαινόμενο μεταξύ 14% και 15% κατά τις περιόδους [2005,2010) και [2010,2015). Ακολουθεί μια εμφανής πτώση στην περίοδο [2015,2020) όπου φτάνει στο χαμηλότερο σημείο του γύρω στο 12%, και στη συνέχεια παρουσιάζει σταδιακή ανάκαμψη στις περιόδους [2020,2025) και [2025,2030], φτάνοντας περίπου στο 14% και 17% αντίστοιχα.

Τέλος, το Product Testing & Frameworks (γαλάζιο) παρουσιάζει την πιο έντονη και συνεχή καθοδική τάση από όλα τα θέματα. Ξεκινά από το υψηλότερο σημείο του περίπου 21% κατά την περίοδο [2005,2010) και μειώνεται σταθερά σε κάθε επόμενη πενταετία, φτάνοντας στο 16% κατά [2010,2015), στο 11% κατά [2015,2020), στο 10% κατά [2020,2025) και τέλος στο χαμηλότερο ποσοστό όλων των θεμάτων, περίπου 8%, κατά την περίοδο [2025,2030].

4.8 Εξέλιξη Συνεργασίας Συγγραφέων στον Χρόνο



Γράφημα 4.8: Εξέλιξη Συνεργασίας Συγγραφέων στον Χρόνο

Το Γράφημα 4.8 παρουσιάζει την εξέλιξη του μέσου αριθμού συγγραφέων ανά δημοσίευση για την περίοδο 2005–2025. Στο γράφημα εμφανίζονται δύο στοιχεία: η μπλε γραμμή που αποτυπώνει την πραγματική τιμή του μέσου αριθμού συγγραφέων για κάθε έτος, και η κόκκινη διακεκομμένη γραμμή που αντιπροσωπεύει τη γενική τάση μέσω γραμμικής παλινδρόμησης. Η γραμμική παλινδρόμηση είναι μια στατιστική μέθοδος που υπολογίζει τη βέλτιστη ευθεία γραμμή που προσαρμόζεται στα δεδομένα, αγνοώντας τις ετήσιες διακυμάνσεις και αναδεικνύοντας τη μακροχρόνια κατεύθυνση της μεταβολής. Στην προκειμένη περίπτωση, η ανοδική κλίση της κόκκινης γραμμής επιβεβαιώνει ότι ο μέσος αριθμός συγγραφέων αυξάνεται σταθερά με την πάροδο του χρόνου, ανεξάρτητα από τις ετήσιες διακυμάνσεις που παρατηρούνται στη μπλε γραμμή.

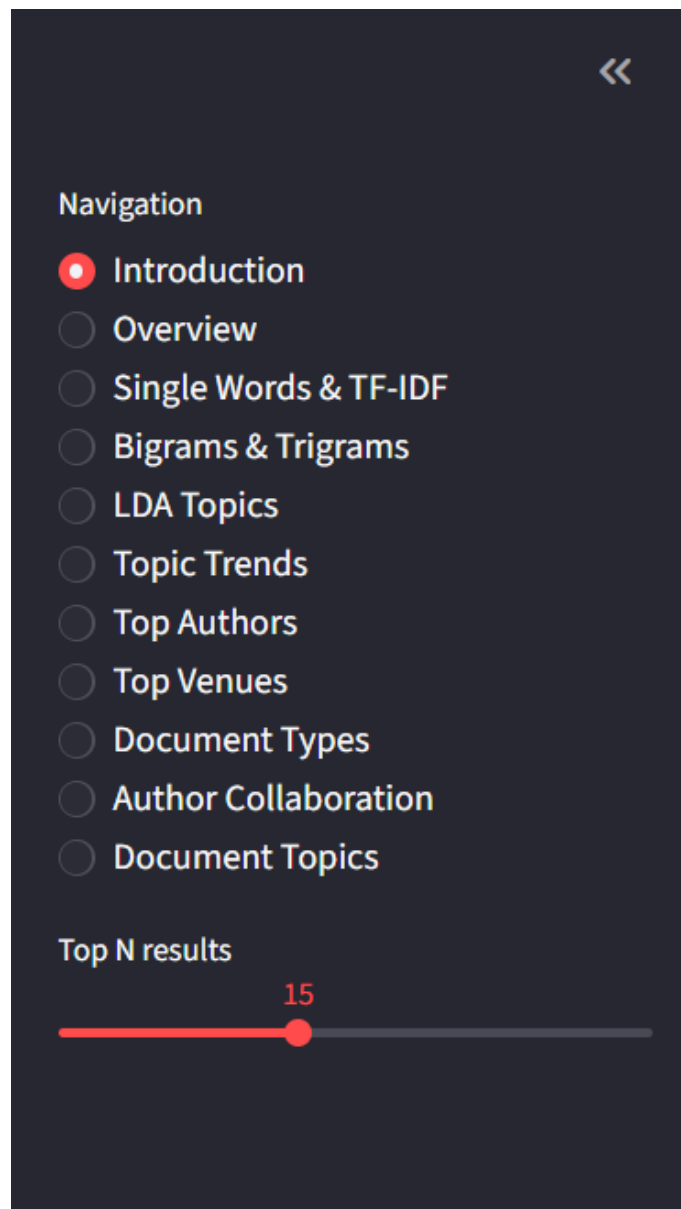
Παρατηρείται σαφής και σταθερή ανοδική τάση στον μέσο αριθμό συγγραφέων καθ' όλη τη διάρκεια της εξεταζόμενης περιόδου. Το 2005, ο μέσος αριθμός συγγραφέων ανά δημοσίευση ήταν 2.71, ενώ το 2025 έφτασε στο 4.45, σημειώνοντας αύξηση της τάξης του 64% σε διάστημα δύο δεκαετιών. Κατά την πρώτη πενταετία 2005–2010, ο μέσος αριθμός συγγραφέων κυμαινόταν μεταξύ 2.71 και 3.24, ενώ κατά την περίοδο 2010–2015 παρατηρείται σταδιακή αύξηση με τιμές που κυμαίνονται μεταξύ 3.04 και 3.75. Στην περίοδο 2015–2020 ο μέσος αριθμός συνεχίζει να αυξάνεται, κυμαινόμενος μεταξύ 3.49 και 3.95, ενώ

στην τελευταία πενταετία 2020–2025 παρατηρούνται οι υψηλότερες τιμές, με αποκορύφωμα το 4.45 το 2025, το υψηλότερο που καταγράφεται σε ολόκληρη την εξεταζόμενη περίοδο. Αξιοσημείωτες είναι οι ήπιες μειώσεις που παρατηρούνται σε ορισμένα έτη, όπως το 2011 (3.04), το 2017 (3.49) και το 2024 (3.86), οι οποίες αποτελούν φυσιολογικές ετήσιες διακυμάνσεις που δεν επηρεάζουν τη γενική ανοδική πορεία όπως αυτή επιβεβαιώνεται από τη γραμμή παλινδρόμησης.

4.9 Διαδραστική Διαδικτυακή Εφαρμογή

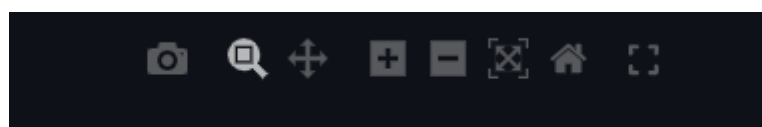
Στο πλαίσιο της παρούσας εργασίας αναπτύχθηκε μια διαδραστική διαδικτυακή εφαρμογή για την παρουσίαση των αποτελεσμάτων της βιβλιομετρικής ανάλυσης. Η εφαρμογή υλοποιήθηκε σε γλώσσα Python με χρήση του framework Streamlit, το οποίο υποστηρίζει την ανάπτυξη και διαμοίραση διαδραστικών εφαρμογών δεδομένων [8]. Επιπλέον, αξιοποιήθηκε η βιβλιοθήκη Plotly για τη δημιουργία διαδραστικών γραφημάτων και οπτικοποιήσεων [9]. Ως πηγή δεδομένων χρησιμοποιούνται τα αρχεία CSV που παράχθηκαν από την ανάλυση σε R, τα οποία φορτώνονται αυτόματα κατά την εκκίνηση της εφαρμογής.

Η εφαρμογή οργανώνεται σε 11 ενότητες πλοήγησης, οι οποίες είναι προσβάσιμες μέσω της πλαϊνής μπάρας. Με αυτόν τον τρόπο, ο χρήστης μπορεί να μεταβαίνει εύκολα από τη μία ανάλυση στην άλλη. Επιπλέον, στην πλαϊνή μπάρα παρέχεται διαδραστικό ρυθμιστικό, μέσω του οποίου ο χρήστης μπορεί να αλλάζει τον αριθμό των δεδομένων που εμφανίζονται στα γραφήματα. Για παράδειγμα, μπορεί να επιλέξει πόσα κορυφαία αποτελέσματα θα παρουσιάζονται, όπως οι κορυφαίοι συγγραφείς, οι κορυφαίες πηγές δημοσίευσης ή οι συχνότεροι όροι.



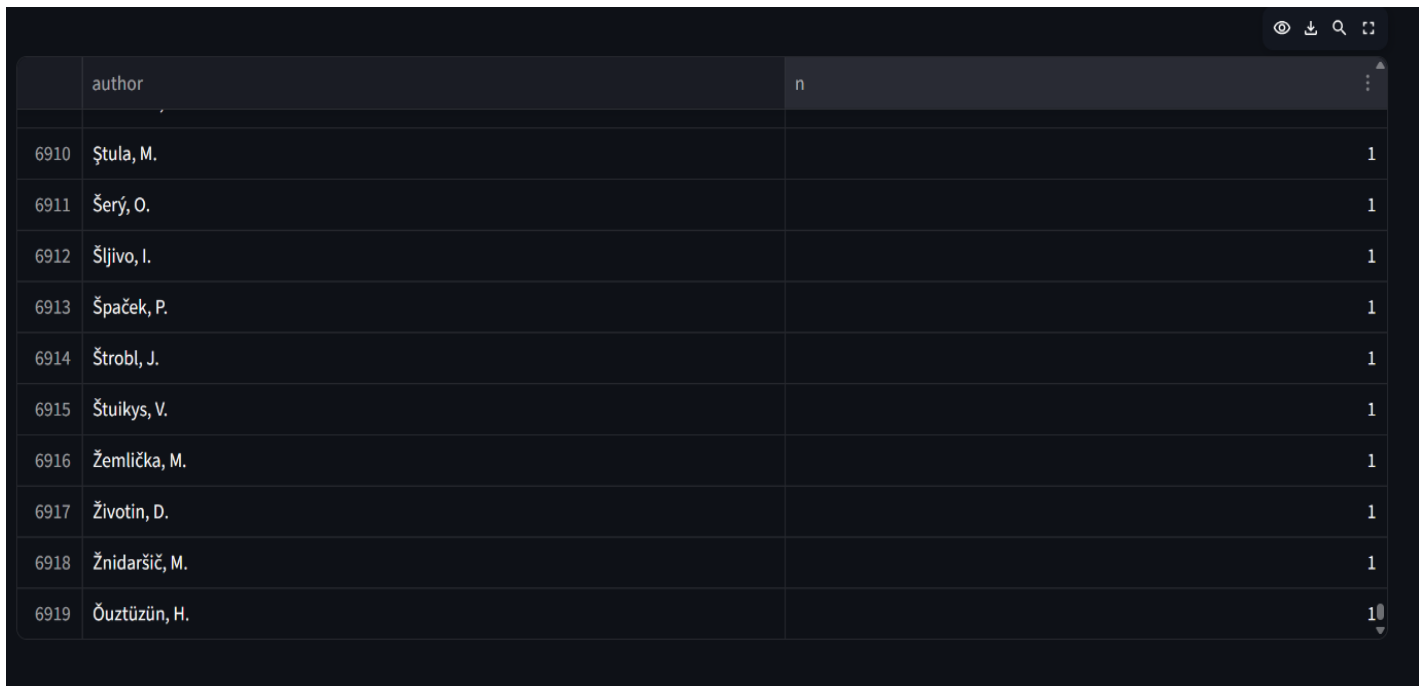
Εικόνα 4.9α: Πλαϊνή Μπάρα

Τα γραφήματα της εφαρμογής είναι διαδραστικά και παρέχουν επιπλέον δυνατότητες εξερεύνησης. Ο χρήστης μπορεί να μεγεθύνει συγκεκριμένα σημεία ενός γραφήματος, να επιλέξει και να εστιάσει σε συγκεκριμένη περιοχή της οπτικοποίησης, καθώς και να αποθηκεύσει τα γραφήματα σε μορφή εικόνας. Οι δυνατότητες αυτές διευκολύνουν την πιο λεπτομερή παρατήρηση των αποτελεσμάτων και καθιστούν την εφαρμογή πιο ευέλικτη για παρουσίαση και ανάλυση.



Εικόνα 4.9β: Διαδραστικές Δυνατότητες Εξερεύνησης Αποτελεσμάτων

Κάτω από κάθε γράφημα εμφανίζεται επίσης ο αντίστοιχος πίνακας δεδομένων, ο οποίος προέρχεται από το αρχείο CSV που χρησιμοποιήθηκε για τη δημιουργία της συγκεκριμένης οπτικοποίησης. Στον πίνακα αυτό ο χρήστης μπορεί να πραγματοποιήσει αναζήτηση, να επιλέξει ποιες στήλες επιθυμεί να εμφανίζονται, καθώς και να αποθηκεύσει το αντίστοιχο αρχείο CSV για περαιτέρω χρήση. Με αυτόν τον τρόπο, η εφαρμογή δεν περιορίζεται μόνο στην οπτική παρουσίαση των αποτελεσμάτων, αλλά παρέχει και άμεση πρόσβαση στα υποκείμενα δεδομένα της ανάλυσης.



	author	n
6910	Štula, M.	1
6911	Šerý, O.	1
6912	Šljivo, I.	1
6913	Špaček, P.	1
6914	Štrobl, J.	1
6915	Štuikys, V.	1
6916	Žemlička, M.	1
6917	Životin, D.	1
6918	Žnidaršič, M.	1
6919	Öztüzün, H.	1

Εικόνα 4.9γ: Αρχείο CSV με όλα τα αποτελέσματα κάθε ανάλυσης

Οι ενότητες της εφαρμογής είναι οι εξής:

Introduction — Παρουσιάζει τους στόχους της εργασίας, την πηγή δεδομένων, τη μεθοδολογία που ακολουθήθηκε και τον σκοπό της εφαρμογής, παρέχοντας στον χρήστη το απαραίτητο πλαίσιο κατανόησης πριν από την εξερεύνηση των αποτελεσμάτων.

Overview — Παρουσιάζει μια συνολική επισκόπηση των δεδομένων μέσω τεσσάρων βασικών μετρικών: συνολικός αριθμός εγγράφων, αριθμός θεμάτων LDA, μοναδικοί συγγραφείς και μοναδικές πηγές δημοσίευσης. Επίσης, περιλαμβάνει γραφήματα για την κατανομή των θεμάτων και την εξέλιξή τους ανά πενταετία.

Single Words & TF-IDF — Παρουσιάζει δύο παράλληλα γραφήματα: τις πιο συχνές μεμονωμένες λέξεις και τους κορυφαίους όρους βάσει βαθμολογίας TF-IDF, μαζί με τους αντίστοιχους πίνακες δεδομένων.

Bigrams & Trigrams — Παρουσιάζει τους πιο συχνούς συνδυασμούς δύο και τριών λέξεων αντίστοιχα, με οριζόντια ραβδόγραμμα και πίνακες δεδομένων.

LDA Topics — Παρουσιάζει τους κορυφαίους όρους κάθε θέματος σε αναγνώσιμη μορφή πίνακα, καθώς και τον αριθμό δημοσιεύσεων που ανήκουν σε κάθε θέμα.

Topic Trends — Παρουσιάζει τη χρονική εξέλιξη των θεμάτων σε περιόδους πενταετίας μέσω διαδραστικού γραφήματος γραμμών, όπου ο χρήστης μπορεί να εστιάσει σε συγκεκριμένα θέματα.

Top Authors — Παρουσιάζει τους πιο παραγωγικούς συγγραφείς με οριζόντιο ραβδόγραμμα και πίνακα δεδομένων.

Top Venues — Παρουσιάζει τα κορυφαία περιοδικά και συνέδρια βάσει αριθμού δημοσιεύσεων, με οριζόντιο ραβδόγραμμα και πίνακα δεδομένων.

Document Types — Παρουσιάζει την κατανομή των τύπων δημοσιεύσεων μέσω διαδραστικού κυκλικού γραφήματος με δακτύλιο.

Author Collaboration — Παρουσιάζει δύο γραφήματα που δείχνουν τον μέσο και τον διάμεσο αριθμό συγγραφέων ανά έτος, επιτρέποντας τη σύγκριση των δύο μετρικών.

Document Topics — Παρέχει τη δυνατότητα φιλτραρίσματος των δημοσιεύσεων ανά θέμα, επιτρέποντας στον χρήστη να δει ποιες δημοσιεύσεις ανήκουν σε κάθε θεματική περιοχή. Επίσης, παρέχεται η δυνατότητα λήψης των φιλτραρισμένων αποτελεσμάτων σε μορφή CSV.

Συνολικά, η εφαρμογή παρέχει μια φιλική προς τον χρήστη διεπαφή που επιτρέπει την εξερεύνηση των αποτελεσμάτων της βιβλιομετρικής ανάλυσης με διαδραστικό τρόπο, χωρίς να απαιτείται τεχνική γνώση από τον χρήστη. Παράλληλα, η δυνατότητα προσαρμογής των γραφημάτων, εξερεύνησης των υποκείμενων δεδομένων και αποθήκευσης των αποτελεσμάτων ενισχύει τη χρησιμότητα της εφαρμογής ως εργαλείο παρουσίασης, επαλήθευσης και περαιτέρω μελέτης των ευρημάτων.



Εικόνα 4.9δ: Ανάλυση Τύπων Δημοσιεύσεων



Εικόνα 4.9ε: Κατανομή των έξι θεμάτων

Σχολιασμός Αποτελεσμάτων

5.1 Επισκόπηση Ευρημάτων	54
5.2 Απαντήσεις στις Ερευνητικές Ερωτήσεις	56
5.3 Ερμηνεία Αποτελεσμάτων	62
5.3.1 Software Product Lines (SPL)	62
5.3.2 Component-Based Architecture	63
5.3.3 Software Process & Services	64
5.4 Πιθανή Επίδραση των LLMs στην Εξέλιξη της Επαναχρησιμοποίησης Λογισμικού	65
5.5 Περιορισμοί	66

5.1 Επισκόπηση Ευρημάτων

Η παρούσα εργασία πραγματοποίησε μια ολοκληρωμένη βιβλιομετρική ανάλυση της επιστημονικής βιβλιογραφίας στον τομέα της επαναχρησιμοποίησης λογισμικού, αξιοποιώντας δεδομένα από τη βάση Scopus για τη χρονική περίοδο 2005–2025. Μετά την εφαρμογή του αυτοματοποιημένου φιλτραρίσματος, το τελικό σύνολο δεδομένων αποτελούνταν από 2.962 δημοσιεύσεις, οι οποίες αποτέλεσαν τη βάση για την περαιτέρω ανάλυση. Τα ευρήματα οργανώνονται γύρω από οκτώ ερευνητικές ερωτήσεις που καλύπτουν χρονικές τάσεις, γεωγραφική κατανομή, θεματική ανάλυση, ανάλυση κειμένου και χαρακτηριστικά της επιστημονικής κοινότητας.

Όσον αφορά τις χρονικές τάσεις, παρατηρήθηκε αυξητική πορεία των δημοσιεύσεων κατά την πρώτη δεκαετία, με κορύφωση γύρω στο 2015–2016, ακολουθούμενη από σταδιακή μείωση στα επόμενα χρόνια. Στην ανάλυση των συγγραφέων, η Yasha Wang αναδείχθηκε ως η πιο παραγωγική συγγραφέας με 28 δημοσιεύσεις, ενώ συνολικά η λίστα των κορυφαίων συγγραφέων περιλαμβάνει ερευνητές από διαφορετικές γεωγραφικές περιοχές. Όσον αφορά τους χώρους δημοσίευσης, μετά τον διαχωρισμό των γενικών σειρών πρακτικών από τα συγκεκριμένα περιοδικά και συνέδρια, σημαντική παρουσία εμφανίζουν καθιερωμένα επιστημονικά περιοδικά της μηχανικής λογισμικού, όπως το Journal of Systems and Software και το IEEE Transactions on Software Engineering, καθώς και εξειδικευμένες πηγές που σχετίζονται με τη μηχανική λογισμικού και την επιστήμη υπολογιστών. Η γεωγραφική ανάλυση μέσω VOSviewer ανέδειξε τις ΗΠΑ, την Κίνα και τη Γερμανία ως τις κυρίαρχες

χώρες, ενώ μεταξύ των σημαντικότερων ιδρυμάτων εμφανίζονται το Carnegie Mellon University, το Cornell University και το University of Surrey με εκτεταμένες διεθνείς συνεργασίες.

Η ανάλυση κειμένου αποκάλυψε ότι οι πιο χαρακτηριστικοί όροι της βιβλιογραφίας βάσει TF-IDF είναι ο «component» με βαθμολογία 50.11, ακολουθούμενος από τους όρους «product» και «test». Η ανάλυση N-Grams ανέδειξε το «product line» ως το πιο συχνό bigram, ενώ στα trigrams κυριαρχούν φράσεις της «product line engineering» και «product line SPL». Το δίκτυο συν-εμφάνισης λέξεων-κλειδιών μέσω VOSviewer αποκάλυψε τέσσερις κύριες ομάδες γύρω από τον κεντρικό κόμβο «software reuse», με κυρίαρχα τα Software Product Lines και την component-based αρχιτεκτονική. Αξιοσημείωτο είναι ότι οι τρεις αυτές τεχνικές ανάλυσης κειμένου συγκλίνουν στο ίδιο αποτέλεσμα, αναδεικνύοντας τα Software Product Lines ως την πιο κυρίαρχη έννοια της βιβλιογραφίας.

Η εφαρμογή του αλγορίθμου LDA με $k=6$ θέματα, το οποίο επιλέχθηκε βάσει του Umass coherence score, αποκάλυψε έξι διακριτές θεματικές περιοχές: Software Processes & Services, Software Product Lines, Design Techniques, Product Testing & Frameworks, Code Quality και Component Based Architecture. Η ανάλυση της κατανομής των κυρίαρχων θεμάτων ανέδειξε τα Software Product Lines ως το κυρίαρχο topic με 587 δημοσιεύσεις (19.8%), ακολουθούμενο από τα Software Processes & Services (18.7%) και το Component Based Architecture (18.3%), ενώ τα τρία αυτά θέματα συγκεντρώνουν συνολικά περίπου το 56.8% του συνόλου των δημοσιεύσεων. Η χρονική ανάλυση των θεμάτων έδειξε ότι το Component Based Architecture εμφανίζει την πιο έντονη ανοδική τάση στην πιο πρόσφατη περίοδο [2025,2030], ενώ το Product Testing Frameworks παρουσιάζει τη μεγαλύτερη καθοδική τάση καθ' όλη τη διάρκεια της εξεταζόμενης περιόδου, κατεβαίνοντας από 21% το 2005 σε μόλις 8% το 2025.

Όσον αφορά την κατανομή των τύπων δημοσιεύσεων, η ανάλυση έδειξε ότι τα Conference Papers αποτελούν τον κυρίαρχο τύπο με ποσοστό 59%, ακολουθούμενα από τα Articles με 38.1%, γεγονός που επιβεβαιώνει ότι η έρευνα στον τομέα διαχέεται κυρίως μέσω συνεδρίων. Τέλος, η ανάλυση της εξέλιξης της συνεργασίας μεταξύ συγγραφέων αποκάλυψε μια σταθερή ανοδική τάση στον μέσο αριθμό συγγραφέων ανά δημοσίευση, από 2.71 το 2005 σε 4.45 το 2025, υποδηλώνοντας ότι η έρευνα στην επαναχρησιμοποίηση λογισμικού γίνεται όλο και πιο συνεργατική με την πάροδο του χρόνου.

5.2 Απαντήσεις στις Ερευνητικές Ερωτήσεις

RQ1: Ποια είναι η χρονική εξέλιξη των δημοσιεύσεων στον τομέα της επαναχρησιμοποίησης λογισμικού κατά την περίοδο 2005–2025;

Η αυξητική τάση που παρατηρείται κατά την πρώτη δεκαετία υποδηλώνει ότι το πεδίο βρισκόταν σε φάση ωρίμανσης, με αυξανόμενο ακαδημαϊκό ενδιαφέρον. Η κορύφωση στην περίοδο 2011–2016 συμπίπτει με την ευρεία διάδοση των Software Product Lines, γεγονός που εξηγεί την έντονη ερευνητική δραστηριότητα της περιόδου. Η ερμηνεία αυτή ενισχύεται και από το *Γράφημα 4.7.3*, όπου το Software Product Lines αναδεικνύονται ως το κυρίαρχο θεματικό πεδίο της βιβλιογραφίας, συγκεντρώνοντας ποσοστό σχεδόν 20% του συνόλου των δημοσιεύσεων. Η μετέπειτα μείωση δεν σημαίνει απαραίτητα ότι το πεδίο χάνει τη σημασία του — αντίθετα, ενδέχεται να αντικατοπτρίζει μια μετατόπιση της έρευνας προς νέες υποπεριοχές που δεν καλύπτονται πλήρως από τους αρχικούς όρους αναζήτησης.

RQ2: Ποια είναι η κατανομή των τύπων δημοσιεύσεων στον τομέα της επαναχρησιμοποίησης λογισμικού;

Η ανάλυση της κατανομής των τύπων δημοσιεύσεων αποκαλύπτει ότι η έρευνα στον τομέα της επαναχρησιμοποίησης λογισμικού διαχέεται κυρίως μέσω συνεδρίων, με τις δημοσιεύσεις από συνέδρια (Conference Papers) να αντιπροσωπεύουν το 59% του συνόλου. Αυτό είναι χαρακτηριστικό γνώρισμα της μηχανικής λογισμικού ως επιστημονικού πεδίου, όπου τα συνέδρια αποτελούν τον κύριο μηχανισμό διάχυσης νέας γνώσης λόγω της ταχύτητας δημοσίευσης και της άμεσης επιστημονικής ανταλλαγής που προσφέρουν. Η σημαντική παρουσία των δημοσιεύσεων σε περιοδικά (Articles) με 38.1% υποδηλώνει ωστόσο ότι το πεδίο διαθέτει και ισχυρή παράδοση δημοσίευσης πιο ώριμων και εκτενών ερευνητικών αποτελεσμάτων σε επιστημονικά περιοδικά. Η σχεδόν αμελητέα παρουσία άλλων τύπων δημοσιεύσεων, όπως Reviews, Book Chapters και Editorials όπου συνολικά και τα τέσσερα είδη αντιστοιχούσαν στο 3%, επιβεβαιώνει ότι η κύρια ερευνητική παραγωγή στο πεδίο συγκεντρώνεται στις δύο κυρίαρχες μορφές.

RQ3: Ποια είναι τα κορυφαία περιοδικά και συνέδρια στα οποία δημοσιεύεται έρευνα σχετική με την επαναχρησιμοποίηση λογισμικού.

Τα αποτελέσματα δείχνουν ότι η έρευνα στην επαναχρησιμοποίηση λογισμικού δημοσιεύεται κυρίως σε εξειδικευμένα επιστημονικά περιοδικά της μηχανικής λογισμικού και της

επιστήμης υπολογιστών, γεγονός που επιβεβαιώνεται τόσο από την ανάλυση των τύπων δημοσιεύσεων όσο και από την κατάταξη των κορυφαίων πηγών δημοσίευσης. Συγκεκριμένα, η κυρίαρχη πηγή είναι το Journal of Systems and Software με 86 δημοσιεύσεις, ακολουθούμενο από το Ruan Jian Xue Bao/Journal of Software με 37 και το Journal of Software: Evolution and Process με 29 δημοσιεύσεις. Ισχυρή παρουσία παρουσιάζει επίσης το International Journal of Software Engineering and Knowledge Engineering με 28 δημοσιεύσεις.

Όσον αφορά τα συνέδρια, το μοναδικό πρακτικό συνεδρίου που εμφανίζεται στους top 20 είναι το Conference Proceedings of the EUROMICRO με 5 δημοσιεύσεις. Η απουσία εξειδικευμένων συνεδρίων όπως το ICSR από τους κορυφαίους χώρους δεν υποδηλώνει απαραίτητα μικρή σημασία τους για το πεδίο, αλλά ενδέχεται να οφείλεται στον τρόπο ευρετηριάσής τους στο Scopus.

Η κυριαρχία των περιοδικών στους κορυφαίους χώρους δεν έρχεται σε αντίφαση με το εύρημα της RQ2, όπου τα Conference Papers αποτελούν το 59% των δημοσιεύσεων. Αυτό εξηγείται από το γεγονός ότι ο όγκος των συνεδριακών δημοσιεύσεων κατανέμεται σε πολύ μεγαλύτερο αριθμό διαφορετικών συνεδρίων, με αποτέλεσμα κανένα μεμονωμένο συνέδριο να μην συγκεντρώνει αρκετές δημοσιεύσεις ώστε να εμφανιστεί στους top 20, σε αντίθεση με τα περιοδικά όπου η παραγωγή είναι πιο συγκεντρωμένη.

Συνολικά, τα αποτελέσματα υποδηλώνουν ότι η επαναχρησιμοποίηση λογισμικού αποτελεί έναν ώριμο ερευνητικό τομέα με ισχυρή παρουσία σε καθιερωμένα επιστημονικά περιοδικά, ενώ παράλληλα η ερευνητική κοινότητα διαχέει νέα αποτελέσματα μέσω ενός ευρέος φάσματος συνεδρίων.

RQ4: Ποιοι είναι οι πιο παραγωγικοί συγγραφείς στον τομέα της επαναχρησιμοποίησης λογισμικού. Ποια είναι η εξέλιξη της συνεργασίας μεταξύ συγγραφέων με την πάροδο του χρόνου;

Η ισχυρή παρουσία Ασιατών ερευνητών στη λίστα των κορυφαίων συγγραφέων φαίνεται να συνδέεται με τη γενικότερη αυξημένη ερευνητική δραστηριότητα της περιοχής στον τομέα. Η ερμηνεία αυτή ενισχύεται και από τα αποτελέσματα της ανάλυσης χωρών και ιδρυμάτων, όπου η Κίνα αναδείχθηκε ως μία από τις πλέον παραγωγικές χώρες, βλ. *Γράφημα 4.4α*. Το γεγονός αυτό υποδηλώνει ότι η υψηλή παρουσία Ασιατών συγγραφέων δεν αποτελεί μεμονωμένο εύρημα, αλλά μέρος μιας ευρύτερης ερευνητικής δυναμικής της περιοχής.

Παράλληλα, η παρουσία ερευνητών από την Ευρώπη και τη Λατινική Αμερική, όπως οι Ampatzoglou, Berger, Egyed, Werner, Assunção, De Almeida και Martinez, δείχνει ότι το πεδίο διατηρεί διεθνή χαρακτήρα και δεν περιορίζεται γεωγραφικά σε μία μόνο περιοχή.

Επιπλέον, το σχετικά μικρό εύρος διαφοροποίησης μεταξύ των 20 κορυφαίων συγγραφέων — από 14 έως 28 δημοσιεύσεις — υποδηλώνει ότι η επιστημονική παραγωγή δεν συγκεντρώνεται σε έναν πολύ περιορισμένο αριθμό ερευνητών, αντίθετα, φαίνεται να κατανέμεται σε ένα ευρύτερο δίκτυο ακαδημαϊκών.

Η ανάλυση της εξέλιξης του μέσου αριθμού συγγραφέων ανά δημοσίευση αποκαλύπτει μια σαφή και σταθερή ανοδική τάση, η οποία επιβεβαιώνεται από τη γραμμή παλινδρόμησης. Η αύξηση από 2.71 συγγραφείς ανά δημοσίευση το 2005 σε 4.45 το 2025 αντιπροσωπεύει αύξηση της τάξης του 64%, γεγονός που υποδηλώνει ότι η έρευνα στην επαναχρησιμοποίηση λογισμικού γίνεται όλο και πιο συνεργατική με την πάροδο του χρόνου. Η τάση αυτή είναι συνεπής με τη γενικότερη εξέλιξη της επιστημονικής έρευνας, όπου τα σύνθετα ερευνητικά ερωτήματα απαιτούν ολοένα και μεγαλύτερες ομάδες ερευνητών με διαφορετικές εξειδικεύσεις. Οι ήπιες ετήσιες διακυμάνσεις που παρατηρούνται σε ορισμένα έτη, όπως το 2011 και το 2017, δεν αναιρούν τη γενική ανοδική πορεία, αλλά αντικατοπτρίζουν φυσιολογικές μεταβολές στη σύνθεση των ερευνητικών ομάδων από έτος σε έτος.

RQ5: Ποιες χώρες και ιδρύματα συμβάλλουν περισσότερο στην έρευνα της επαναχρησιμοποίησης λογισμικού.

Η ισχυρή παρουσία των Ηνωμένων Πολιτειών, της Κίνας και ευρωπαϊκών χωρών, όπως η Γερμανία και το Ηνωμένο Βασίλειο, μπορεί να εξηγηθεί από το γεγονός ότι οι χώρες αυτές διαθέτουν ανεπτυγμένα ερευνητικά οικοσυστήματα στον τομέα της Πληροφορικής και της μηχανικής λογισμικού. Συγκεκριμένα, φιλοξενούν μεγάλο αριθμό πανεπιστημίων, ερευνητικών κέντρων και τεχνολογικών οργανισμών που δραστηριοποιούνται σε θέματα ανάπτυξης λογισμικού, αρχιτεκτονικής συστημάτων και επαναχρησιμοποίησης. Επιπλέον, η επαναχρησιμοποίηση λογισμικού συνδέεται άμεσα με πρακτικά ζητήματα της βιομηχανίας λογισμικού, όπως η μείωση κόστους, η διαχείριση μεγάλων συστημάτων και η ανάπτυξη οικογενειών προϊόντων. Για τον λόγο αυτό, χώρες με ισχυρή τεχνολογική και ερευνητική δραστηριότητα είναι αναμενόμενο να εμφανίζονται πιο έντονα στη βιβλιογραφία.

Η παρουσία της Κίνας μπορεί επίσης να συνδεθεί με τη γενικότερη αύξηση της ερευνητικής παραγωγής της χώρας σε πεδία που σχετίζονται με την Πληροφορική και την τεχνολογία. Αυτό εξηγεί γιατί η Κίνα εμφανίζεται ισχυρή σε επίπεδο χώρας, ακόμη και αν η παραγωγή

της δεν συγκεντρώνεται απαραίτητα σε ένα μόνο ίδρυμα, αλλά κατανέμεται σε περισσότερα πανεπιστήμια και ερευνητικούς οργανισμούς. Αντίστοιχα, η παρουσία ευρωπαϊκών χωρών δείχνει ότι το πεδίο της επαναχρησιμοποίησης λογισμικού δεν αναπτύσσεται μόνο σε μεγάλα τεχνολογικά κέντρα, αλλά και μέσα από ευρωπαϊκά πανεπιστήμια και ερευνητικά δίκτυα που δραστηριοποιούνται στη μηχανική λογισμικού.

RQ6: Ποιες είναι οι κυρίαρχες θεματικές περιοχές στη βιβλιογραφία της επαναχρησιμοποίησης λογισμικού και πώς εξελίχθηκαν οι ερευνητικές θεματικές περιοχές με την πάροδο του χρόνου;

Τα αποτελέσματα της θεματικής ανάλυσης έδειξαν ότι η βιβλιογραφία της επαναχρησιμοποίησης λογισμικού δεν επικεντρώνεται σε μία μόνο διάσταση, αλλά αναπτύσσεται γύρω από πολλαπλές συμπληρωματικές θεματικές περιοχές. Το κυρίαρχο θέμα αναδείχθηκε το Software Product Lines, το οποίο συγκεντρώνει περίπου το 19,8% των δημοσιεύσεων, γεγονός που υποδηλώνει ότι η συστηματική ανάπτυξη πολλαπλών προϊόντων μέσω κοινών επαναχρησιμοποιήσιμων στοιχείων αποτελεί τον σημαντικότερο άξονα έρευνας στο πεδίο.

Παράλληλα, ιδιαίτερα υψηλή παρουσία καταγράφουν οι θεματικές Software Process & Services (18,7%) και Component-Based Architecture (18,3%), γεγονός που δείχνει ότι η επαναχρησιμοποίηση δεν περιορίζεται μόνο σε επίπεδο κώδικα, αλλά επεκτείνεται σε διαδικασίες ανάπτυξης, αρχιτεκτονικές αποφάσεις και επαναχρησιμοποίηση συνιστωσών.

Επιπλέον, θεματικές όπως Code Quality (15,6%), Product Testing & Frameworks (13,9%) και Design Techniques (13,7%) δείχνουν ότι η βιβλιογραφία δίνει ιδιαίτερη έμφαση και στην ποιότητα του επαναχρησιμοποιούμενου λογισμικού, στις διαδικασίες ελέγχου και στις τεχνικές σχεδιασμού που υποστηρίζουν την αποτελεσματική εφαρμογή της επαναχρησιμοποίησης.

Συνολικά, τα αποτελέσματα δείχνουν ότι η επαναχρησιμοποίηση λογισμικού έχει εξελιχθεί από μια απλή πρακτική επαναχρησιμοποίησης κώδικα σε ένα πολυδιάστατο ερευνητικό πεδίο που περιλαμβάνει αρχιτεκτονική λογισμικού, διαχείριση προϊόντων, διαδικασίες ανάπτυξης, ποιότητα λογισμικού και στρατηγικές σχεδιασμού.

Η χρονική ανάλυση των θεμάτων αποκαλύπτει ενδιαφέρουσες μεταβολές στις ερευνητικές προτεραιότητες του πεδίου, οι οποίες αντικατοπτρίζουν την εξέλιξη της τεχνολογίας λογισμικού τις τελευταίες δύο δεκαετίες.

Το πιο αξιοσημείωτο εύρημα είναι η δραματική άνοδος των Software Product Lines, από 11.9% την περίοδο [2005,2010) σε 23.2% την περίοδο [2010,2015), διατηρώντας την κυρίαρχη θέση του και στην περίοδο [2015,2020) με 23.1%. Αυτή η εκρηκτική αύξηση δεν είναι τυχαία — συμπίπτει με την ευρεία υιοθέτηση των SPL από τη βιομηχανία ως συστηματική μέθοδο επαναχρησιμοποίησης σε μεγάλη κλίμακα, ιδιαίτερα στους τομείς της αυτοκινητοβιομηχανίας, των τηλεπικοινωνιών και του ενσωματωμένου λογισμικού. Η σταδιακή μείωση στις πιο πρόσφατες περιόδους (18.9% και 19.4%) υποδηλώνει ωρίμανση του πεδίου και όχι απώλεια ενδιαφέροντος.

Αντίθετα, το Product Testing & Frameworks παρουσιάζει τη μεγαλύτερη καθοδική τάση, μειούμενο από 21.1% το [2005,2010) σε μόλις 6.5% το [2025,2030]. Αυτή η μείωση ενδέχεται να αντικατοπτρίζει τη μετατόπιση της έρευνας από γενικές τεχνικές testing προς πιο εξειδικευμένες μεθόδους που απορροφώνται από άλλες θεματικές περιοχές.

Το Components Based Architecture παρουσιάζει σταθερή παρουσία με ανοδική τάση στην τελευταία περίοδο, φτάνοντας στο υψηλότερο ποσοστό του (23.7%) το [2025,2030], γεγονός που υποδηλώνει αυξανόμενο ενδιαφέρον για αρχιτεκτονικές προσεγγίσεις βασισμένες σε συνιστώσες.

Το Code Quality εμφανίζει σταθερά αυξητική τάση από 11.9% το [2005,2010) σε 20.2% το [2020,2025), αντικατοπτρίζοντας την αυξανόμενη σημασία που αποδίδει η κοινότητα στην ποιότητα και τη συντηρησιμότητα του κώδικα, ιδιαίτερα στο πλαίσιο του ανοιχτού κώδικα ανάπτυξης. Τέλος, τα Software Process & Services και Design Techniques διατηρούν σχετικά σταθερή παρουσία καθ' όλη την περίοδο, υποδηλώνοντας ότι αποτελούν διαχρονικά θεμέλια του πεδίου που δεν επηρεάζονται σημαντικά από τις τεχνολογικές εξελίξεις.

RQ7: Ποιοι είναι οι πιο χαρακτηριστικοί όροι της βιβλιογραφίας στον τομέα της επαναχρησιμοποίησης λογισμικού.

Η ανάλυση των πιο χαρακτηριστικών όρων πραγματοποιήθηκε μέσω δύο συμπληρωματικών τεχνικών: της ανάλυσης N-Grams και της μεθόδου TF-IDF. Η ανάλυση N-Grams αποκάλυψε ότι οι πιο συχνά εμφανιζόμενοι μεμονωμένοι όροι (unigrams) στις περιλήψεις των δημοσιεύσεων σχετίζονται με βασικές έννοιες της μηχανικής λογισμικού, ενώ στα bigrams κυριαρχούν φράσεις όπως "product line" και "software product", και στα trigrams φράσεις όπως "software product line" και "product line engineering", επιβεβαιώνοντας τη σημασία των Software Product Lines στη βιβλιογραφία. Η ανάλυση TF-IDF, η οποία αναδεικνύει τους πιο διακριτικούς όρους ανεξάρτητα από τη συχνότητά τους, έδειξε ότι ο όρος "component"

κατέχει την υψηλότερη βαθμολογία με 49.07, ακολουθούμενος από τους όρους "product" (42.93) και "test" (30.03). Συνολικά, τα αποτελέσματα και των δύο τεχνικών συγκλίνουν, αναδεικνύοντας τα Software Product Lines, την component-based ανάπτυξη και τις τεχνικές σχεδίασης ως τις κυρίαρχες έννοιες της βιβλιογραφίας.

RQ8: Πώς σχετίζονται οι λέξεις-κλειδιά στη βιβλιογραφία της επαναχρησιμοποίησης λογισμικού με τις θεματικές περιοχές που εντοπίστηκαν;

Η ανάλυση συν-εμφάνισης λέξεων-κλειδιών μέσω του VOSviewer έδειξε ότι η βιβλιογραφία της επαναχρησιμοποίησης λογισμικού οργανώνεται γύρω από συγκεκριμένες έντονα συνδεδεμένες έννοιες, οι οποίες εμφανίζονται συχνά μαζί στις δημοσιεύσεις. Κεντρικός κόμβος του δικτύου αποτελεί ο όρος software reuse, ο οποίος συνδέεται ισχυρά με έννοιες όπως software product lines, software engineering, software architecture, component, reusability και software development. Το εύρημα αυτό δείχνει ότι η επαναχρησιμοποίηση λογισμικού δεν αντιμετωπίζεται ως μεμονωμένη πρακτική, αλλά συνδέεται άμεσα με ευρύτερες διαδικασίες ανάπτυξης, σχεδιασμού και αρχιτεκτονικής λογισμικού.

Ιδιαίτερα σημαντική είναι η παρουσία της περιοχής των Software Product Lines, το οποίο συνδέεται με όρους όπως variability, feature model, variability management και reverse engineering. Αυτό επιβεβαιώνει ότι μεγάλο μέρος της βιβλιογραφίας επικεντρώνεται στη διαχείριση παραλλαγών μεταξύ διαφορετικών προϊόντων και στην αξιοποίηση κοινών επαναχρησιμοποιήσιμων στοιχείων.

Παράλληλα, μία δεύτερη περιοχή περιλαμβάνει όρους όπως software engineering, software architecture, component, ontology και UML, υποδηλώνοντας τη σημασία των αρχιτεκτονικών προσεγγίσεων και της μοντελοποίησης στη διαδικασία επαναχρησιμοποίησης.

Επιπλέον, εμφανίζεται μία ομάδα που σχετίζεται περισσότερο με την πρακτική εφαρμογή της επαναχρησιμοποίησης μέσω όρων όπως code reuse, open source software, requirements engineering και empirical study, γεγονός που δείχνει αυξανόμενο ενδιαφέρον για την αξιολόγηση της αποτελεσματικότητας των τεχνικών επαναχρησιμοποίησης σε πραγματικά περιβάλλοντα ανάπτυξης.

Το εύρημα αυτό ενισχύει σημαντικά τα αποτελέσματα του αλγορίθμου LDA, καθώς οι ίδιες θεματικές περιοχές — όπως τα Software Product Lines και Components Based Architecture — εμφανίστηκαν και στην ανάλυση θεμάτων. Παράλληλα, παρουσιάζει συνέπεια και με τα

αποτελέσματα των N-Grams και του TF-IDF, όπου όροι όπως product, component, design, feature, line και architecture εμφανίστηκαν με υψηλή συχνότητα.

5.3 Ερμηνεία Αποτελεσμάτων

Πέρα από τις απαντήσεις στις επιμέρους ερευνητικές ερωτήσεις, η συνολική ανάλυση αναδεικνύει τρεις κυρίαρχες πτυχές της επαναχρησιμοποίησης λογισμικού που διατρέχουν το σύνολο της βιβλιογραφίας. Οι πτυχές αυτές εντοπίστηκαν μέσω της σύγκλισης πολλαπλών τεχνικών ανάλυσης — LDA, TF-IDF, N-Grams και VOSviewer — και αντιπροσωπεύουν τις πιο ώριμες και καλά τεκμηριωμένες ερευνητικές κατευθύνσεις στο πεδίο.

5.3.1 Software Product Lines (SPL)

Μία από τις σημαντικότερες θεματικές περιοχές που αναδείχθηκαν στην παρούσα βιβλιομετρική ανάλυση είναι τα Software Product Lines (SPL), τα οποία εμφανίστηκαν ως η κυρίαρχη θεματική περιοχή της βιβλιογραφίας. Συγκεκριμένα, από την ανάλυση LDA το Θέμα 2 καταγράφηκε ως το πιο κυρίαρχο θέμα, συγκεντρώνοντας σχεδόν το 20% των δημοσιεύσεων, ενώ όροι όπως product, feature, line, spl, design και application εμφανίστηκαν με υψηλή πιθανότητα στο συγκεκριμένο θέμα. Παράλληλα, στην ανάλυση TF-IDF και N-Grams εμφανίστηκαν επαναλαμβανόμενα όροι όπως product, feature, line και requirement, ενώ στο δίκτυο συν-εμφάνισης λέξεων του VOSviewer οι όροι software product line, variability, feature model και variability management αποτέλεσαν από τους ισχυρότερους κόμβους του δικτύου. Τα αποτελέσματα αυτά δείχνουν ότι τα SPL αποτελούν διαχρονικά τον πιο έντονο ερευνητικό πυρήνα του πεδίου της επαναχρησιμοποίησης λογισμικού.

Η κυριαρχία των SPL εξηγείται από το γεγονός ότι αποτελούν μία πιο οργανωμένη και στρατηγική μορφή επαναχρησιμοποίησης λογισμικού σε σχέση με τις παλαιότερες προσεγγίσεις. Σύμφωνα με τη Northrop L. (2005) [6], ένα Software Product Line ορίζεται ως ένα σύνολο συστημάτων λογισμικού που μοιράζονται ένα κοινό σύνολο χαρακτηριστικών και αναπτύσσονται από κοινά επαναχρησιμοποιήσιμα στοιχεία όπως αρχιτεκτονικές και συνιστώσες. Σε αντίθεση με τις παραδοσιακές μορφές επαναχρησιμοποίησης που βασίζονταν σε μεμονωμένες βιβλιοθήκες κώδικα, τα SPL εισήγαγαν μια πιο συστηματική προσέγγιση, όπου η επαναχρησιμοποίηση σχεδιάζεται από την αρχή. Χαρακτηριστικά παραδείγματα εφαρμογής SPL στη βιομηχανία περιλαμβάνουν την αυτοκινητοβιομηχανία, όπου κατασκευαστές αναπτύσσουν οικογένειες οχημάτων που μοιράζονται κοινές πλατφόρμες

λογισμικού και στον τομέα των τηλεπικοινωνιών όπου εταιρείες διαχειρίζονται οικογένειες προϊόντων με κοινή βάση κώδικα.

Η παρουσίαση της Northrop L. (2005) [6] δείχνει επίσης ξεκάθαρα την ιστορική εξέλιξη της επαναχρησιμοποίησης: από τις υπορουτίνες τη δεκαετία του 1960, στα modules τη δεκαετία του 1970, στα objects τη δεκαετία του 1980, στα components τη δεκαετία του 1990 και τελικά στα Software Product Lines τη δεκαετία του 2000. Αυτό συνδέεται άμεσα με την απάντηση της πρώτης Ερευνητικής Ερώτησης (RQ1), όπου παρατηρήθηκε σημαντική αύξηση των δημοσιεύσεων ιδιαίτερα κατά την περίοδο 2011–2016, περίοδος κατά την οποία τα SPL φαίνεται να είχαν ήδη καθιερωθεί ως βασική ερευνητική κατεύθυνση.

Ένας βασικός λόγος της έντονης παρουσίας των SPL στη βιβλιογραφία είναι τα πρακτικά οφέλη που προσφέρουν στις επιχειρήσεις. Σύμφωνα με τη Northrop L. (2005) [6], οργανισμοί που υιοθέτησαν επιτυχώς πρακτικές SPL πέτυχαν σημαντικές μειώσεις στο κόστος ανάπτυξης, ταχύτερο χρόνο διάθεσης προϊόντων στην αγορά και βελτίωση της ποιότητας.

Συνολικά, τα ευρήματα της παρούσας εργασίας δείχνουν ότι τα Software Product Lines αποτέλεσαν την πιο ώριμη και συστηματική μορφή επαναχρησιμοποίησης λογισμικού στη βιβλιογραφία, εξηγώντας γιατί κυριάρχησαν τόσο στις αναλύσεις θεμάτων όσο και στις αναλύσεις λέξεων-κλειδιών.

5.3.2 Component-Based Architecture

Η δεύτερη σημαντικότερη θεματική περιοχή που αναδείχθηκε στην παρούσα εργασία αφορά την Component-Based Architecture, η οποία εμφανίστηκε ως μία από τις κυρίαρχες θεματικές τόσο στην ανάλυση LDA όσο και στις αναλύσεις λέξεων-κλειδιών. Συγκεκριμένα, το αντίστοιχο θέμα συγκέντρωσε περίπου 18.3% των δημοσιεύσεων, ενώ όροι όπως *component*, *architecture*, *framework*, *domain*, *requirement* και *solution* εμφανίστηκαν με υψηλή πιθανότητα στα αποτελέσματα του LDA. Παράλληλα, στις αναλύσεις unigram, bigram και TF-IDF, ο όρος *component* εμφανίστηκε σταθερά στις υψηλότερες θέσεις, γεγονός που επιβεβαιώνει τη σημαντική θέση των component-based προσεγγίσεων στη βιβλιογραφία.

Η σημασία των component-based αρχιτεκτονικών συνδέεται με την προσπάθεια της μηχανικής λογισμικού να μεταβεί από την επαναχρησιμοποίηση απλού κώδικα σε πιο δομημένες μορφές επαναχρησιμοποίησης μεγαλύτερης κλίμακας. Αντί να

επαναχρησιμοποιούνται μεμονωμένες συναρτήσεις ή βιβλιοθήκες, η προσέγγιση αυτή επιτρέπει την επαναχρησιμοποίηση αυτόνομων λογισμικών μονάδων με σαφώς ορισμένες διεπαφές, οι οποίες μπορούν να ενσωματωθούν σε διαφορετικά συστήματα.

Το εύρημα αυτό υποστηρίζεται και από το άρθρο των Jung, M., & Biersack, E. W. (2000) [10], όπου αναφέρεται ότι η ανάπτυξη βασισμένη σε συνιστώσες (components) προσφέρει «μια νέα διάσταση επαναχρησιμοποίησης και ταχείας δημιουργίας πρωτοτύπων λογισμικού» μέσω της δημιουργίας ανεξάρτητων components που μπορούν να συνδεθούν μεταξύ τους για την ανάπτυξη νέων εφαρμογών. Οι συγγραφείς εξηγούν ότι οι συνιστώσες είναι ανεξάρτητες, διαθέτουν σαφώς καθορισμένες διεπαφές και μπορούν να προσαρμοστούν χωρίς να απαιτείται τροποποίηση του βασικού τους κώδικα, χαρακτηριστικά που ενισχύουν σημαντικά την επαναχρησιμοποίηση. Ένα χαρακτηριστικό παράδειγμα αποτελεί η ανάπτυξη μιας πλατφόρμας ηλεκτρονικού εμπορίου, όπου ένας οργανισμός μπορεί να επαναχρησιμοποιήσει ανεξάρτητα components όπως σύστημα σύνδεσης χρηστών (login), σύστημα πληρωμών, διαχείριση καλαθιού αγορών και υπηρεσίες ειδοποιήσεων. Τα συγκεκριμένα components μπορούν να ενσωματωθούν σε διαφορετικές εφαρμογές χωρίς να απαιτείται εκ νέου ανάπτυξη από την αρχή, μειώνοντας σημαντικά τον χρόνο ανάπτυξης και το κόστος παραγωγής λογισμικού.

Συνολικά, τα αποτελέσματα δείχνουν ότι η Component-Based Architecture αποτέλεσε ένα κρίσιμο ενδιάμεσο στάδιο στην εξέλιξη της επαναχρησιμοποίησης λογισμικού, γεφυρώνοντας την απλή επαναχρησιμοποίηση κώδικα με πιο σύνθετες στρατηγικές όπως τα Software Product Lines.

5.3.3 Software Process & Services

Η τρίτη σημαντική θεματική περιοχή που αναδείχθηκε στην παρούσα εργασία αφορά το Software Process & Services, το οποίο αποτέλεσε μία από τις κυρίαρχες θεματικές περιοχές της βιβλιογραφίας. Στην ανάλυση LDA, το Topic 1 συγκέντρωσε περίπου 18.7% των δημοσιεύσεων, με κυρίαρχους όρους όπως *process*, *application*, *service*, *program*, *tool*, *component* και *architecture*. Παράλληλα, ο όρος *process* καταγράφηκε ως ο τέταρτος πιο συχνός όρος στην ανάλυση unigram με 1.887 εμφανίσεις, ενώ εμφανίστηκε και σε άλλα θέματα της ανάλυσης LDA, γεγονός που δείχνει ότι οι διαδικασίες ανάπτυξης λογισμικού αποτελούν διαχρονικό κομμάτι της βιβλιογραφίας της επαναχρησιμοποίησης.

Σύμφωνα με τη βιβλιογραφία, ο όρος *software process* αναφέρεται στο σύνολο των δραστηριοτήτων, μεθόδων, πρακτικών και διαδικασιών που χρησιμοποιούνται για την

ανάπτυξη, συντήρηση και διαχείριση συστημάτων λογισμικού. Η σύνδεσή του με την επαναχρησιμοποίηση λογισμικού είναι ιδιαίτερα σημαντική, καθώς η αποτελεσματική επαναχρησιμοποίηση δεν εξαρτάται μόνο από τεχνικά στοιχεία, όπως ο κώδικας ή οι συνιστώσες, αλλά και από οργανωμένες διαδικασίες που επιτρέπουν την αναγνώριση, αξιολόγηση, ενσωμάτωση και συντήρηση επαναχρησιμοποιήσιμων στοιχείων.

Για παράδειγμα, μια εταιρεία ανάπτυξης λογισμικού μπορεί να δημιουργήσει μια τυποποιημένη διαδικασία ανάπτυξης εφαρμογών, όπου κάθε νέο project ακολουθεί συγκεκριμένα στάδια επιλογής reusable libraries, testing reusable components και deployment services. Με αυτόν τον τρόπο, η επαναχρησιμοποίηση δεν γίνεται αποσπασματικά, αλλά ενσωματώνεται στην καθημερινή διαδικασία ανάπτυξης λογισμικού.

Η ισχυρή παρουσία όρων όπως service, application και tool δείχνει επίσης ότι η βιβλιογραφία έχει επεκταθεί πέρα από τις παραδοσιακές μορφές επαναχρησιμοποίησης, εξετάζοντας πλέον την επαναχρησιμοποίηση σε επίπεδο υπηρεσιών, εργαλείων ανάπτυξης και λειτουργικών διαδικασιών.

Συνολικά, τα αποτελέσματα δείχνουν ότι η επαναχρησιμοποίηση λογισμικού έχει εξελιχθεί από ένα καθαρά τεχνικό ζήτημα σε μια ευρύτερη οργανωτική και διαδικαστική πρακτική, όπου οι σωστά σχεδιασμένες διαδικασίες ανάπτυξης παίζουν καθοριστικό ρόλο στην επιτυχημένη εφαρμογή της επαναχρησιμοποίησης.

5.4 Πιθανή Επίδραση των LLMs στην Εξέλιξη της Επαναχρησιμοποίησης Λογισμικού

Year	Total Documents	Documents about LLMs	Percentage (%)
2023	94	25	26.6
2024	109	29	26.61
2025	139	41	29.5

Πίνακας 5.4: LLMs κατά την πάροδο του χρόνου

Η πρόσφατη ανάπτυξη των μεγάλων γλωσσικών μοντέλων (Large Language Models – LLMs) δημιουργεί νέες προοπτικές για την επαναχρησιμοποίηση λογισμικού. Σε αντίθεση με τις παραδοσιακές μορφές επαναχρησιμοποίησης, οι οποίες βασίζονται κυρίως στην αναζήτηση και ενσωμάτωση έτοιμων βιβλιοθηκών, συνιστωσών ή τμημάτων κώδικα, τα LLMs μπορούν να υποστηρίξουν πιο δυναμικές μορφές επαναχρησιμοποίησης. Συγκεκριμένα, μπορούν να βοηθήσουν τους προγραμματιστές στην κατανόηση υπάρχοντος

κώδικα, στην αναζήτηση κατάλληλων βιβλιοθηκών ή APIs, στην προσαρμογή υφιστάμενων λύσεων σε νέα προβλήματα και στην παραγωγή νέου κώδικα βασισμένου σε προηγούμενα παραδείγματα.

Για να εξεταστεί αν η επίδραση αυτή αρχίζει να αποτυπώνεται στη βιβλιογραφία της επαναχρησιμοποίησης λογισμικού, πραγματοποιήθηκε συμπληρωματική ανάλυση στο τελικό σύνολο δεδομένων, εστιάζοντας στην περίοδο μετά το 2023. Η ανάλυση βασίστηκε στην αναζήτηση όρων που σχετίζονται με την τεχνητή νοημοσύνη, τα μεγάλα γλωσσικά μοντέλα, τη γενετική τεχνητή νοημοσύνη και την παραγωγή κώδικα, όπως “large language model”, “LLM”, “GPT”, “generative AI”, “artificial intelligence”, “machine learning”, “deep learning”, “code generation”, “program synthesis”, “code recommendation” και “API recommendation”.

Τα αποτελέσματα δείχνουν ότι την περίοδο 2023–2025 η παρουσία AI/LLM-related όρων είναι εμφανώς αυξημένη σε σχέση με την προηγούμενη βιβλιογραφία. Η αύξηση αυτή υποδηλώνει ότι η επαναχρησιμοποίηση λογισμικού αρχίζει να συνδέεται περισσότερο με τεχνικές τεχνητής νοημοσύνης και αυτοματοποιημένης παραγωγής ή υποστήριξης κώδικα. Ωστόσο, η κατηγορία που αφορά αποκλειστικά τα LLMs και τη γενετική τεχνητή νοημοσύνη παραμένει μικρότερη σε σύγκριση με τη γενικότερη κατηγορία AI/ML. Αυτό δείχνει ότι, παρόλο που τα LLMs αποτελούν μια νέα και σημαντική τεχνολογική εξέλιξη, η παρουσία τους στη βιβλιογραφία της επαναχρησιμοποίησης λογισμικού βρίσκεται ακόμη σε αναδυόμενο στάδιο.

Με βάση τα παραπάνω, μπορεί να εκτιμηθεί ότι τα LLMs θα αποκτήσουν σημαντικότερο ρόλο στη μελλοντική εξέλιξη της επαναχρησιμοποίησης λογισμικού. Η επίδρασή τους είναι πιθανό να ενισχυθεί κυρίως σε περιοχές όπως η αυτόματη παραγωγή κώδικα, η αναζήτηση επαναχρησιμοποιήσιμων τμημάτων λογισμικού, η πρόταση κατάλληλων APIs και βιβλιοθηκών, καθώς και η προσαρμογή υπάρχοντος κώδικα σε νέες απαιτήσεις. Επομένως, η μελλοντική έρευνα ενδέχεται να μετατοπιστεί προς πιο έξυπνες και αυτοματοποιημένες μορφές επαναχρησιμοποίησης, όπου τα LLMs θα λειτουργούν ως υποστηρικτικά εργαλεία για την κατανόηση, επιλογή και προσαρμογή λογισμικού. Παρόλα αυτά, η επίδραση αυτή πρέπει να θεωρηθεί αναδυόμενη και όχι πλήρως εδραιωμένη, καθώς η σχετική βιβλιογραφία μετά το 2023 βρίσκεται ακόμη σε πρώιμο στάδιο ανάπτυξης.

5.5 Περιορισμοί

Παρά την ολοκληρωμένη ανάλυση που πραγματοποιήθηκε στην παρούσα εργασία, υπάρχουν ορισμένοι περιορισμοί που θα πρέπει να ληφθούν υπόψη κατά την ερμηνεία των αποτελεσμάτων.

Αρχικά, η συλλογή των δεδομένων βασίστηκε αποκλειστικά στη βιβλιογραφική βάση Scopus. Παρόλο που η συγκεκριμένη βάση παρέχει υψηλής ποιότητας και ευρέως χρησιμοποιούμενα δεδομένα, είναι πιθανό να υπάρχουν σχετικές δημοσιεύσεις σε άλλες βάσεις δεδομένων, όπως το Web of Science ή το IEEE Xplore, οι οποίες δεν συμπεριλήφθηκαν στην ανάλυση.

Επιπλέον, η διαδικασία συλλογής των δεδομένων βασίστηκε σε συγκεκριμένες λέξεις-κλειδιά, ενώ τα δεδομένα που αξιοποιήθηκαν για την ανάλυση περιορίστηκαν αποκλειστικά στα μεταδεδομένα των δημοσιεύσεων, όπως αυτά εξήχθησαν από τη βάση Scopus. Η προσέγγιση αυτή ενδέχεται να αποκλείει σχετικές δημοσιεύσεις που δεν περιλαμβάνουν τους όρους αυτούς, γεγονός που μπορεί να επηρεάσει την πληρότητα του συνόλου δεδομένων.

Ένας ακόμη περιορισμός αφορά τη χρήση τεχνικών ανάλυσης κειμένου και συγκεκριμένα της θεματικής μοντελοποίησης μέσω LDA. Παρόλο που το LDA αποτελεί μία ευρέως χρησιμοποιούμενη μέθοδο, βασίζεται σε στατιστικά μοντέλα και δεν λαμβάνει πλήρως υπόψη τη σημασιολογική ερμηνεία των όρων. Ως αποτέλεσμα, η ερμηνεία των θεμάτων απαιτεί ανθρώπινη κρίση και ενδέχεται να περιλαμβάνει έναν βαθμό υποκειμενικότητας. Επιπλέον, η χειροκίνητη ανάλυση των δημοσιεύσεων πραγματοποιήθηκε σε περιορισμένο βαθμό, κυρίως για σκοπούς επαλήθευσης της συνάφειας και ερμηνείας των αποτελεσμάτων, και όχι ως πλήρης συστηματική αξιολόγηση κάθε δημοσίευσης. Συνεπώς, ορισμένες ταξινομήσεις ή θεματικές αποδόσεις ενδέχεται να επηρεάζονται από τους περιορισμούς της αυτοματοποιημένης ανάλυσης.

Τέλος, η ανάλυση βασίστηκε κυρίως σε περιλήψεις (abstracts) των δημοσιεύσεων και όχι στο πλήρες κείμενο. Η προσέγγιση αυτή, αν και συνηθισμένη σε βιβλιομετρικές μελέτες, ενδέχεται να περιορίζει το βάθος της ανάλυσης.

Κεφάλαιο 6

Συμπεράσματα

6.1 Σύνοψη της Εργασίας	68
6.2 Κύρια Συμπεράσματα	69
6.3 Συνεισφορές της Εργασίας	69
6.4 Μελλοντική Έρευνα	70

6.1 Σύνοψη της Εργασίας

Η παρούσα διπλωματική εργασία είχε ως στόχο τη διεξαγωγή μιας ολοκληρωμένης βιβλιομετρικής ανάλυσης της επιστημονικής βιβλιογραφίας στον τομέα της επαναχρησιμοποίησης λογισμικού για την περίοδο 2005–2025. Τα δεδομένα συλλέχθηκαν από τη βιβλιογραφική βάση Scopus μέσω στοχευμένης αναζήτησης (advanced query), και μετά την εφαρμογή αυτοματοποιημένου φιλτραρίσματος σε R, το τελικό σύνολο δεδομένων αποτελούνταν από 2.962 δημοσιεύσεις από ένα αρχικό σύνολο 12.673 εγγραφών.

Για την ανάλυση των δεδομένων αξιοποιήθηκε συνδυασμός ποσοτικών και υπολογιστικών τεχνικών. Συγκεκριμένα, εφαρμόστηκαν στατιστική ανάλυση για την εξαγωγή πληροφοριών σχετικά με τους κορυφαίους συγγραφείς, χώρους δημοσίευσης, χώρες και ιδρύματα, ανάλυση κειμένου μέσω N-Grams και TF-IDF για τον εντοπισμό των πιο χαρακτηριστικών όρων, topic modelling μέσω LDA για την ανακάλυψη θεματικών δομών, και οπτικοποίηση δικτύων μέσω VOSviewer για την ανάδειξη συν-εμφανίσεων λέξεων-κλειδιών και γεωγραφικών συνεργασιών. Επιπλέον, αναλύθηκε η κατανομή των τύπων δημοσιεύσεων και η εξέλιξη της συνεργασίας μεταξύ συγγραφέων, παρέχοντας μια πολυδιάστατη εικόνα της επιστημονικής κοινότητας του πεδίου. Τα αποτελέσματα παρουσιάστηκαν μέσω γραφημάτων, πινάκων και διαδραστικής διαδικτυακής εφαρμογής που αναπτύχθηκε σε Python και Streamlit.

6.2 Κύρια Συμπεράσματα

Η ανάλυση οδήγησε σε μια σειρά σημαντικών συμπερασμάτων για τη δομή και την εξέλιξη της έρευνας στην επαναχρησιμοποίηση λογισμικού.

Το πιο σημαντικό και αξιοσημείωτο εύρημα της εργασίας είναι η σύγκλιση όλων των τεχνικών ανάλυσης στην ανάδειξη των Software Product Lines (SPL) ως της κυρίαρχης θεματικής περιοχής της βιβλιογραφίας. Συγκεκριμένα, η ανάλυση N-Grams ανέδειξε το "product line" ως το πιο συχνό bigram και το "software product line" ως το πιο συχνό trigram. Η ανάλυση TF-IDF κατέταξε τους όρους "product" και "component" στις κορυφαίες θέσεις. Το δίκτυο συν-εμφάνισης λέξεων-κλειδιών μέσω VOSviewer αποκάλυψε τα SPL ως την κυρίαρχη ομάδα γύρω από τον κεντρικό κόμβο "software reuse". Τέλος, ο αλγόριθμος LDA εντόπισε ένα ξεχωριστό θέμα αποκλειστικά αφιερωμένο στα Software Product Lines. Το γεγονός ότι τέσσερις εντελώς διαφορετικές τεχνικές ανάλυσης καταλήγουν στο ίδιο αποτέλεσμα ενισχύει σημαντικά την αξιοπιστία του ευρήματος και επιβεβαιώνει ότι τα SPL αποτελούν αναμφισβήτητα την πιο διαδεδομένη και μελετημένη μορφή επαναχρησιμοποίησης λογισμικού.

Όσον αφορά τις χρονικές τάσεις, η έρευνα στο πεδίο παρουσίασε αυξητική πορεία κατά την πρώτη δεκαετία με κορύφωση γύρω στο 2011-2016, ακολουθούμενη από σταδιακή μείωση, γεγονός που υποδηλώνει ότι το πεδίο έχει φτάσει σε φάση ωρίμανσης. Η γεωγραφική ανάλυση ανέδειξε τις ΗΠΑ, την Κίνα και τη Γερμανία ως τις κυρίαρχες χώρες, ενώ η ανάλυση ιδρυμάτων αποκάλυψε πυκνά δίκτυα διεθνών συνεργασιών με επίκεντρο ιδρύματα όπως το Carnegie Mellon University. Η έρευνα διαχέεται κυρίως μέσω συνεδρίων (59%), γεγονός χαρακτηριστικό της μηχανικής λογισμικού ως επιστημονικού πεδίου, ενώ η σταθερή αύξηση του μέσου αριθμού συγγραφέων από 2.71 το 2005 σε 4.45 το 2025 επιβεβαιώνει ότι η έρευνα γίνεται όλο και πιο συνεργατική. Τέλος, η χρονική ανάλυση των θεμάτων αποκάλυψε ότι το Component Architecture εμφανίζει την πιο έντονη ανοδική τάση στην πιο πρόσφατη περίοδο, αναδεικνύοντάς το ως την αναδύομενη θεματική περιοχή του πεδίου για τα επόμενα χρόνια.

6.3 Συνεισφορές της Εργασίας

Η παρούσα εργασία συνεισφέρει στο πεδίο της επαναχρησιμοποίησης λογισμικού με διάφορους τρόπους.

Πρώτον, παρέχει μια ενημερωμένη και ολοκληρωμένη βιβλιομετρική ανάλυση του πεδίου για την περίοδο 2005–2025, καλύπτοντας δύο ολόκληρες δεκαετίες ερευνητικής

δραστηριότητας. Προηγούμενες βιβλιομετρικές μελέτες στη μηχανική λογισμικού, όπως αυτή των Wong, W. E., Mittas, N., Arvanitou, E. M., & Li, Y. (2021) [16], κάλυπταν μικρότερα χρονικά εύρη, αφήνοντας κενό στην κατανόηση της πιο πρόσφατης εξέλιξης του πεδίου.

Δεύτερον, η εργασία εισάγει έναν αυτοματοποιημένο μηχανισμό φιλτραρίσματος δημοσιεύσεων που αναπτύχθηκε σε R, ο οποίος συνδυάζει φίλτρο αξιόπιστων τύπων δημοσιεύσεων, μηχανισμό βαθμολόγησης συνάφειας και φίλτρο αποκλεισμού άσχετων εννοιών. Ο μηχανισμός αυτός επιτρέπει την αποτελεσματική επεξεργασία μεγάλου όγκου δεδομένων με αυξημένη ακρίβεια και αναπαραγωγικότητα.

Τρίτον, η εργασία αξιοποιεί έναν συνδυασμό πολλαπλών τεχνικών ανάλυσης — στατιστική ανάλυση, N-Grams, TF-IDF, LDA και VOSviewer — παρέχοντας μια πολυδιάστατη εικόνα του πεδίου που δεν θα ήταν δυνατή με μία μόνο τεχνική. Για την εφαρμογή των τεχνικών εξόρυξης κειμένου σε R αξιοποιήθηκε η μεθοδολογική προσέγγιση του Text Mining with R: A Tidy Approach [7], η οποία παρουσιάζει τεχνικές ανάλυσης κειμένου όπως η ανάλυση TF-IDF, η ανάλυση n-grams και η θεματική μοντελοποίηση με LDA. Η σύγκλιση όλων αυτών των τεχνικών στην ανάδειξη των Software Product Lines ως κυρίαρχης θεματικής περιοχής αποτελεί από μόνη της σημαντικό εύρημα που ενισχύει την αξιοπιστία των αποτελεσμάτων.

Τέλος, αναπτύχθηκε μια διαδραστική διαδικτυακή εφαρμογή σε Python και Streamlit για την παρουσίαση των αποτελεσμάτων, η οποία επιτρέπει στους χρήστες να εξερευνούν τα δεδομένα με διαδραστικό τρόπο. Η εφαρμογή αυτή καθιστά τα αποτελέσματα της ανάλυσης προσβάσιμα σε ευρύτερο κοινό, πέρα από την ακαδημαϊκή κοινότητα.

6.4 Μελλοντική Έρευνα

Παρά την ολοκληρωμένη προσέγγιση που ακολουθήθηκε, υπάρχουν αρκετές κατευθύνσεις στις οποίες η παρούσα εργασία θα μπορούσε να επεκταθεί στο μέλλον.

Πρώτον, η συλλογή δεδομένων θα μπορούσε να επεκταθεί σε επιπλέον βιβλιογραφικές βάσεις, όπως το DBLP, η IEEE Digital Library και το Web of Science, ώστε να διασφαλιστεί πληρέστερη κάλυψη της βιβλιογραφίας και να μειωθεί η εξάρτηση από μία μόνο πηγή δεδομένων.

Δεύτερον, μια σημαντική μελλοντική επέκταση θα μπορούσε να είναι η ανάλυση αναφορών, με στόχο τον εντοπισμό των πιο επιδραστικών δημοσιεύσεων, συγγραφέων και

επιστημονικών πηγών στο πεδίο της επαναχρησιμοποίησης λογισμικού. Η παρούσα εργασία επικεντρώθηκε κυρίως στην ποσότητα, τη θεματολογία και τη χρονική εξέλιξη των δημοσιεύσεων. Η ανάλυση αναφορών θα πρόσθετε μια επιπλέον διάσταση, καθώς θα επέτρεπε την αξιολόγηση της επιστημονικής επιρροής των δημοσιεύσεων και της ροής της γνώσης μέσα στο πεδίο.

Τρίτον, μελλοντική εργασία θα μπορούσε να εξετάσει διαφορετικές προσεγγίσεις θεματικής μοντελοποίησης. Στην παρούσα εργασία χρησιμοποιήθηκε ο αλγόριθμος LDA με έξι θέματα, ωστόσο θα μπορούσαν να δοκιμαστούν εναλλακτικές μέθοδοι, όπως NMF ή BERTopic. Μια τέτοια σύγκριση θα βοηθούσε στην αξιολόγηση του κατά πόσο διαφορετικές τεχνικές μπορούν να αποκαλύψουν πιο λεπτομερείς ή πιο ερμηνεύσιμες θεματικές περιοχές.

Τέλος, η χρήση πιο σύγχρονων τεχνικών επεξεργασίας φυσικής γλώσσας, όπως τα γλωσσικά μοντέλα μεγάλης κλίμακας (Large Language Models — LLMs), θα μπορούσε να αξιοποιηθεί για πιο ακριβή κατανόηση και ταξινόμηση των περιλήψεων. Οι τεχνικές αυτές θα μπορούσαν να λειτουργήσουν συμπληρωματικά με το LDA, προσφέροντας πιο σημασιολογικά ακριβή αποτελέσματα και βαθύτερη κατανόηση των θεματικών περιοχών της βιβλιογραφίας.

Βιβλιογραφία

- [1] **Krueger, C. W.** (1992). Software reuse. *ACM Computing Surveys*, 24(2), 131–183.
- [2] **Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M.** (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, Volume 133, 285–296.
- [3] **Spärck Jones, K.** (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21. (Reprinted 2004, Vol. 60, No. 5, pp. 493–502)
- [4] **Blei, D. M., Ng, A. Y., & Jordan, M. I.** (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- [5] **Stănescu, A.; Delcea, C.** (2024). A Bibliometric Analysis of Text Mining: Exploring the Use of Natural Language Processing in Social Media Research. *Applied Sciences*, 14(8), 3144.
- [6] **Northrop, L.** (2005). Software Product Lines. Software Engineering Institute, Carnegie Mellon University.
- [7] **Silge, J., & Robinson, D.** (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media.
- [8] **Streamlit.** (n.d.). Streamlit: A faster way to build and share data apps. Retrieved May 5, 2026, from <https://streamlit.io/>
- [9] **Plotly.** (n.d.). Plotly Python Graphing Library. Retrieved May 5, 2026, from <https://plotly.com/python/>
- [10] **Jung, M., & Biersack, E. W.** (2000). A component-based architecture for software communication systems. In *Proceedings of IEEE ECBS 2000*, Edinburgh, Scotland.
- [11] **Öztürk, O., Kocaman, R., & Kanbach, D. K.** (2024). How to design bibliometric research: An overview and a framework proposal. *Review of Managerial Science*, 18, 3333–3361.

- [12] **Sandu, A., Cotfas, L.-A., Stănescu, A., & Delcea, C. (2024).** A bibliometric analysis of text mining: Exploring the use of Natural Language Processing in social media research. *Applied Sciences*, 14(8), 3144.

- [13] **Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012).** Exploring topic coherence over many models and many topics. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952–961.

- [14] **Van Eck, N. J., & Waltman, L. (2023).** *VOSviewer manual: Manual for VOSviewer version 1.6.20*. Universiteit Leiden; Erasmus Universiteit Rotterdam.

- [15] **Tomaszewski, R. (2023).** Visibility, impact, and applications of bibliometric software tools through citation analysis. *Scientometrics*, 128(7), 4007-4028.

- [16] **Wong, W. E., Mittas, N., Arvanitou, E. M., & Li, Y. (2021).** A bibliometric assessment of software engineering themes, scholars and institutions (2013-2020). *The Journal of Systems & Software*, 180, 111029.

- [17] **Elsevier. (n.d.).** *Scopus database*, from <https://www.scopus.com>

Παράρτημα Α

Α1. Λίστα έγκριστων περιοδικών και συνεδρίων της Πληροφορικής

Journal of Computational Design and Engineering
Journal of Computational Information Systems
Journal of Computational Methods in Sciences and Engineering
Journal of Computational Social Science
Journal of Computer Assisted Learning
Journal of Computer Information Systems
Journal of Computer Networks and Communications
Journal of Computer Science
Journal of Computer Science and Technology
Journal of Computer Security
Journal of Computer Virology and Hacking Techniques
Journal of Computers
Journal of Computers (Taiwan)
Journal of Computing and Information Science in Engineering
Journal of Computing and Information Technology
Journal of Control Theory and Applications
Journal of Convergence Information Technology
Journal of Cultural Analytics
Journal of Cyber Security
Journal of Data and Information Quality
Journal of Digital Imaging
Journal of Digital Information
Journal of Electronic Imaging
Journal of Field Robotics
Journal of Frontiers of Computer Science and Technology
Journal of Functional Programming
Journal of Geomatics
Journal of Global Optimization
Journal of Graphics
Journal of Grid Computing
Journal of Image and Graphics
Journal of Imaging
Journal of Imaging Science and Technology
Journal of Information Processing
Journal of Information Processing Systems

International Arab Journal of Information Technology
"International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives"
"International Conference for High Performance Computing, Networking, Storage and Analysis, SC"
"International Conference on Advanced Computing and Communication Technologies, ACCT"
International Conference on Advanced Technologies for Communications
"International Conference on Cloud Computing and Services Science, CLOSER - Proceedings"
"International Conference on Communication Technology Proceedings, ICCT"
"International Conference on Control, Automation and Systems"
"International Conference on Cyber Conflict, CYCON"
"International Conference on Enterprise Information Systems, ICEIS - Proceedings"
"International Conference on Evaluation of Novel Approaches to Software Engineering, ENASE - Proceedings"
International Conference on Geoinformatics
International Conference on Information Networking
"International Conference on Intelligent User Interfaces, Proceedings IUI"
"International Conference on Internet of Things, Big Data and Security, IoTBDS - Proceedings"
International Conference on Model-Driven Engineering and Software Development
"International Conference on Signal Processing Proceedings, ICSP"
International Conference on Transparent Optical Networks
"International Conference on Ubiquitous and Future Networks, ICUFN"
"International Conference on Wireless and Mobile Computing, Networking and Communications"
International Congress on Ultra Modern Telecommunications and Control Systems and Workshops
International Design and Test Workshop
International Geoscience and Remote Sensing Symposium (IGARSS)
"International IEEE Conference proceedings, IS"
International Joint Conference on Computational Intelligence
International Journal of Advanced Computer Science and Applications
International Journal of Advanced Intelligence Paradigms
International Journal of Advanced Manufacturing Technology
International Journal of Advanced Robotic Systems
International Journal of Advanced Science and Technology
International Journal of Advanced Trends in Computer Science and Engineering
International Journal of Advancements in Computing Technology
International Journal of Advances in Soft Computing and its Applications
International Journal of Agent-Oriented Software Engineering
International Journal of Applied Mathematics and Computer Science
International Journal of Applied Systemic Studies

IEEE Transactions on Cloud Computing
IEEE Transactions on Cognitive and Developmental Systems
IEEE Transactions on Computational Social Systems
IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems
IEEE Transactions on Computers
IEEE Transactions on Cybernetics
IEEE Transactions on Dependable and Secure Computing
IEEE Transactions on Emerging Topics in Computational Intelligence
IEEE Transactions on Emerging Topics in Computing
IEEE Transactions on Evolutionary Computation
IEEE Transactions on Industrial Informatics
IEEE Transactions on Information Forensics and Security
IEEE Transactions on Knowledge and Data Engineering
IEEE Transactions on Learning Technologies
IEEE Transactions on Mobile Computing
IEEE Transactions on Multimedia
IEEE Transactions on Network Science and Engineering
IEEE Transactions on Neural Networks
IEEE Transactions on Neural Networks and Learning Systems
IEEE Transactions on Parallel and Distributed Systems
IEEE Transactions on Pattern Analysis and Machine Intelligence
IEEE Transactions on Quantum Engineering
IEEE Transactions on Robotics
IEEE Transactions on Services Computing
IEEE Transactions on Signal Processing
IEEE Transactions on Software Engineering
IEEE Transactions on Sustainable Computing
"IEEE Transactions on Systems, Man, and Cybernetics: Systems"
IEEE Transactions on Vehicular Technology
IEEE Transactions on Very Large Scale Integration (VLSI) Systems
IEEE Transactions on Visualization and Computer Graphics
IEEE Transactions on Wireless Communications
IEEE Vehicular Technology Conference
IEEE Wireless Communications
"IEEE Workshop on Signal Processing Systems, SiPS: Design and Implementation"

Complex and Intelligent Systems
Complexity
Complexity International
Computacion y Sistemas
Computation
Computational Intelligence and Neuroscience
Computational Mechanics
Computational Statistics and Data Analysis
Computational Toxicology
Computer
Computer Animation and Virtual Worlds
Computer Bulletin (London)
Computer Communication Review
Computer Communications
Computer Engineering and Applications
Computer Fraud and Security
Computer Graphics Forum
Computer Journal
Computer Methods in Applied Mechanics and Engineering
Computer Networks
Computer Science - Research and Development
Computer Science Review
Computer Science and Information Systems
Computer Software
Computer Standards and Interfaces
Computer Supported Cooperative Work
Computer Systems Science and Engineering
Computer Vision and Image Understanding
Computer-Aided Design and Applications
Computers
Computers and Electrical Engineering
Computers and Geosciences
Computers and Industrial Engineering
Computers and Mathematics with Applications
Computers and Operations Research
Computers and Security

Παράρτημα Β

Β1. Πίνακας με πρόσθετες λέξεις διακοπής και χειροκίνητες κανονικοποιήσεις όρων

Κατηγορία	Όροι
Όροι ειδικού περιεχομένου	“software”, “reuse”, “reusability”
Γενικοί όροι υψηλής συχνότητας	“using”, “use”, “used”, “based”, “approach”, “study”, “method”, “methods”, “result”, “results”, “system”, “systems”, “analysis”, “case”, “cases”, “proposed”, “problem”, “problems”, “research”, “development”, “new”, “code”, “data”, “model”, “models”
Βιβλιογραφικοί και εκδοτικοί όροι	“paper”, “article”, “elsevier”, “rights”, “reserved”, “ltd”, “copyright”, “publisher”, “springer”, “ieee”, “acm”, “press”, “journal”, “science”, “verlag”, “berlin”, “heidelberg”, “reserve”

```
extra_stop <- c(  
  "software", "reuse", "reusability", "using", "use", "used", "based", "approach",  
  "paper", "article", "study", "method", "methods", "result", "results", "system",  
  "systems", "analysis", "case", "cases", "proposed", "problem", "problems",  
  "research", "development", "new", "elsevier", "rights", "reserved", "ltd", "copyright", "publisher",  
  "springer", "ieee", "acm", "press", "journal", "science", "code", "data", "based", "model", "models",  
  "verlag", "berlin", "heidelberg", "reserve"  
)
```

Αρχικός όρος	Κανονικοποιημένος όρος
spls	spl
sple	spl
engineer	engineering
orient	oriented
datum	data

```
normalize_domain_terms <- function(x) {
  dplyr::case_when(
    x %in% c("spls") ~ "spl",
    x %in% c("sple") ~ "spl",
    x %in% c("engineer") ~ "engineering",
    x %in% c("orient") ~ "oriented",
    x %in% c("datum") ~ "data",
    TRUE ~ x
  )
}
```