ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

Μελέτη της Κοινότητας της R στο StackOverflow:Ανάλυση Ερωτήσεων και Θεμάτων χρησιμοποιώντας Τεχνικές Επεξεργασίας Φυσικής Γλώσσας

Study in R community of StackOverflow: An Analysis of Questions and Topics Using Natural Language Processing techniques

Νικόλαος Ζώρος
ΑΕΜ:3533

Εκπόνηση πτυχιακής ως μέρος του Προπτυχιακού Τίτλου Σπουδών
στη
Σχολή Θετικών Επιστημών Τμήμα Πληροφορικής

Επιβλέπων Καθηγητής: Ν.Βασιλειάδης
Συνεπίβλεψη: Δρ. Μαρία Παπουτσόγλου

Θεσσαλονίκη, 30-10-2023

ΠΕΡΙΛΗΨΗ

Το Stack Overflow είναι μία δημοφιλής διαδικτυακή πλατφόρμα, η οποία χρησιμοποιείται ευρέως από προγραμματιστές που μπορούν να υποβάλλουν τεχνικές ερωτήσεις και να απαντήσουν τις ερωτήσεις άλλων χρηστών, που αφορούν διάφορους κλάδους της πληροφορικής. Ανάμεσα στις αμέτρητες κοινότητες που περιβάλουν το SO, η κοινότητα της γλώσσας προγραμματισμού R διακρίνεται μία από τις μεγαλύτερες και ενεργή κοινότητες με παραπάνω από 120.000 χρήστες. Η R χρησιμοποιείται κυρίως για ανάλυση δεδομένων και στατιστικούς υπολογισμούς. Παρόλη την αυξημένη δημοτικότητα της, ύστερα από εκτενής αναζήτηση δεν βρέθηκαν αρκετές έρευνες και αναλύσεις που να αφορούν καθαρά τα θέματα, τις τάσεις και τεχνολογίες που απασχολούν τους χρήστες της.

Σε αυτήν την εργασία χρησιμοποιήθηκε το σύνολο των 412.102 ερωτήσεων που αναρτήθηκαν στο SO οι οποίες είναι συσχετισμένες με την R (tag), από 131.302 χρήστες. Έπειτα, χρησιμοποιήθηκε στατιστική ανάλυση για την εξαγωγή περιγραφικών δεδομένων και εφαρμόστηκαν αλγόριθμοι επεξεργασίας φυσικής γλώσσας (Natural Language Processing) για εντοπισμό θεμάτων και διάφορων τάσεων. Τα αποτελέσματα της έρευνας υποδεικνύουν πως συζητήσεις που αφορούν θέματα όπως **DataFrame Operations** και **Data Visualization** είναι τα πιο δημοφιλή πεδία κατά την κοινότητα της R.Επίσης, συμπεριλαμβάνονται κάποιοι χρήσιμοι ορισμοί και μία συνοπτική ανασκόπηση της σχετικής βιβλιογραφίας. Τα αποτελέσματα και τα συμπεράσματα της συγκεκριμένης έρευνας, καθώς και η μεθοδολογία που χρησιμοποιείται, παρουσιάζονται αναλυτικά. Τέλος, επισημαίνονται ορισμένοι περιορισμοί της εργασίας και προτείνονται ενδιαφέροντα θέματα για μελλοντική έρευνα.

Λέξεις – Κλειδιά: Ανάλυση Δεδομένων, Ανάλυση Κειμένου, Stack Overflow ,NLP,R.

ABSTRACT

Stack Overflow is a popular online platform which is widely used by developers who are able to submit technical questions and answer questions by other users, which are related to various IT disciplines. Among the countless communities surrounding SO, the R community stands out as one of the largest and most active communities with over than 120.000 users. R is mainly used for data analysis and statistical analysis. Despite its increased popularity, after an extensive search there have not been found much research and analysis on R's topics, trends, and technologies which provides insights for the R community.

In this particular thesis, it is used a set of 412.102 questions posted on the SO which are related to R (tag), from 131.302 users. Then, we used statistical analysis to extract descriptive data, and NLP algorithms were applied to identify topics and trends. Our results indicate that discussions about topics as **DataFrame Operations** and **Data Visualization** are the most popular areas in R community. Also, some useful definitions and a brief review of the referenced papers. The results and conclusions of this research, as well as the methodology implementation, are presented in detail. Finally, some limitations of the study are highlighted as long as with suggestions of interesting topics for future research.

Key words: Data analysis, Text analysis, Stack Overflow, NLP, R.

Ευχαριστίες

Θέλω να ευχαριστήσω τον επιβλέποντα καθηγητή μου, Νικόλαο Βασιλειάδη, για την υποστήριξη και εμπιστοσύνη που έδειξε στο πρόσωπο μου καθώς επίσης και την κυρία Μαρία Παπούτσογλου για την πολύτιμη βοήθεια και καθοδήγηση πουμου παρείχε. Τέλος, επιθυμώ να πω ένα μεγάλο ευχαριστώ στους γονείς μου που με βοηθούν καθημερινά να πετύχω τους στόχους μου.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Στόχος της υλοποίησης και Υπόβαθρο

Η ανάπτυξη των κοινοτήτων στις διαδικτυακές πλατφόρμες του διαδικτύου έχει αλλάξει ριζικά τον τρόπο που οι προγραμματιστές συνεργάζονται, μοιράζονται πληροφορίες και ψάχνουν λύσεις στα προβλήματα που αντιμετωπίζουν. Μεταξύ αυτών των πλατφορμών, επικρατεί το Stack Overflow ως η κυρίαρχη σελίδα που θα απευθυνθούν οι προγραμματιστές, παρέχοντας ένα τεράστιο αποθετήριο ερωτήσεων και απαντήσεων σε διάφορες γλώσσες προγραμματισμού και τεχνολογιών. Στην παρούσα εργασία, εμβαθύνουμε στην κοινότητα της γλώσσας προγραμματισμού R της Stack Overflow σελίδας, αναζητώντας διάφορα δημοφιλή θέματα που συζητούν οι χρήστες της στο πέρασμα του χρόνου, καθώς επίσης και διάφορα χαρακτηριστικά που διαμορφώνουν την ίδια την γλώσσα και τους χρήστες της. Αυτό επιτεύχθηκε με την εφαρμογή ενός *Natural language processing (NLP)* αλγόριθμο, τον *Latent Dirichlet allocation(LDA).*

Η R έχει αποκτήσει τεράστια δημοφιλία τα τελευταία χρόνια και πιο συγκεκριμένα στον χώρο της ανάλυσης δεδομένων και της υπολογιστικής στατιστικής. Η ευελιξία της, το εκτεταμένο οικοσύστημα πακέτων που διαθέτει και η ενεργή της κοινότητα την έχουν καθορίσει μία από τις προτιμώμενες γλώσσες για πολλούς ερευνητές, επιστήμονες δεδομένων και στατιστικολόγους. Ως εκ τούτου, η κοινότητα της R στο Stack Overflow θεωρείται μια αξιόπιστη πηγή πληροφοριών, τεχνογνωσίας και συνεργασίας ατόμων που αναζητούν βοήθεια ή εξερευνούν τα όρια αυτής της ισχυρής γλώσσας.

Ο πρωταρχικός στόχος αυτής της πτυχιακής είναι η εξερεύνηση υποκείμενων θεμάτων και τάσεων ανάμεσα στους χρήστες της R κατά την εξέλιξη της στον χρόνο καθώς επίσης και διάφορα χαρακτηριστικά της γλώσσας και των χρηστών της. Η επίλυση αυτού του στόχου μπορεί να γίνει χρήσιμη σε διάφορους τομείς. Μπορεί να αποτελέσει οδηγός για τους προγραμματιστές της R και τους συντηρητές των πακέτων της στο να βελτιώσουν την γλώσσα, συγκεκριμένες βιβλιοθήκες ή ακόμα και κάποιο documentation. Επίσης, μπορεί να βοηθήσει στην καλύτερη κατανόηση των χρηστών της καθώς αναλύοντας το ποιόν των ερωτήσεων και των χρηστών που τις έθεσαν, μπορεί να παραχθούν διάφορα δημογραφικά στοιχεία της κοινότητας. Όλα αυτά θα βοηθήσουν αφενός σε μια ευκολότερη αρχική εκμάθηση της R και αφετέρου σε ένα πιο εύχρηστο και λειτουργικό περιβάλλον της.

Έτσι, με βάση τα παραπάνω η συγκεκριμένη εργασία επιδιώκει να δώσει απαντήσεις στα εξής ερευνητικά ερωτήματα:

- Ποια είναι τα βασικά χαρακτηριστικά της γλώσσας και των χρηστών της.
- Ποια είναι τα κυρίαρχα θέματα και οι τάσεις που αμφιταλαντεύουν τους χρήστες της R στο Stack Overflow και πως αυτά εξελίσσονται με το πέρασμα του χρόνου.

Η οργάνωση των κεφαλαίων της παρούσας εργασίας πραγματοποιείται ως εξής:

- Στο **Κεφάλαιο 2** γίνεται η ανασκόπηση των βιβλιογραφιών που έχουν εκπονηθεί από ερευνητές και μελετητές, οι οποίες χρησιμοποιήθηκαν στην συγκεκριμένη εργασία.
- Στο **Κεφάλαιο 3** παρουσιάζεται αναλυτικά η μεθοδολογία που ακολουθήθηκε για την εκπόνηση της εργασίας.
- Στο **Κεφάλαιο 4** παρουσιάζεται μια αναλυτική αναφορά για τα αποτελέσματα της εργασίας.
- Στο **Κεφάλαιο 5** παρουσιάζονται λεπτομερώς τα συμπεράσματα που εξάχθηκαν καθώς και ιδέες για περεταίρω μελέτη στον συγκεκριμένο τομέα.

Κεφάλαιο 2

Επισκόπηση Βιβλιογραφίας

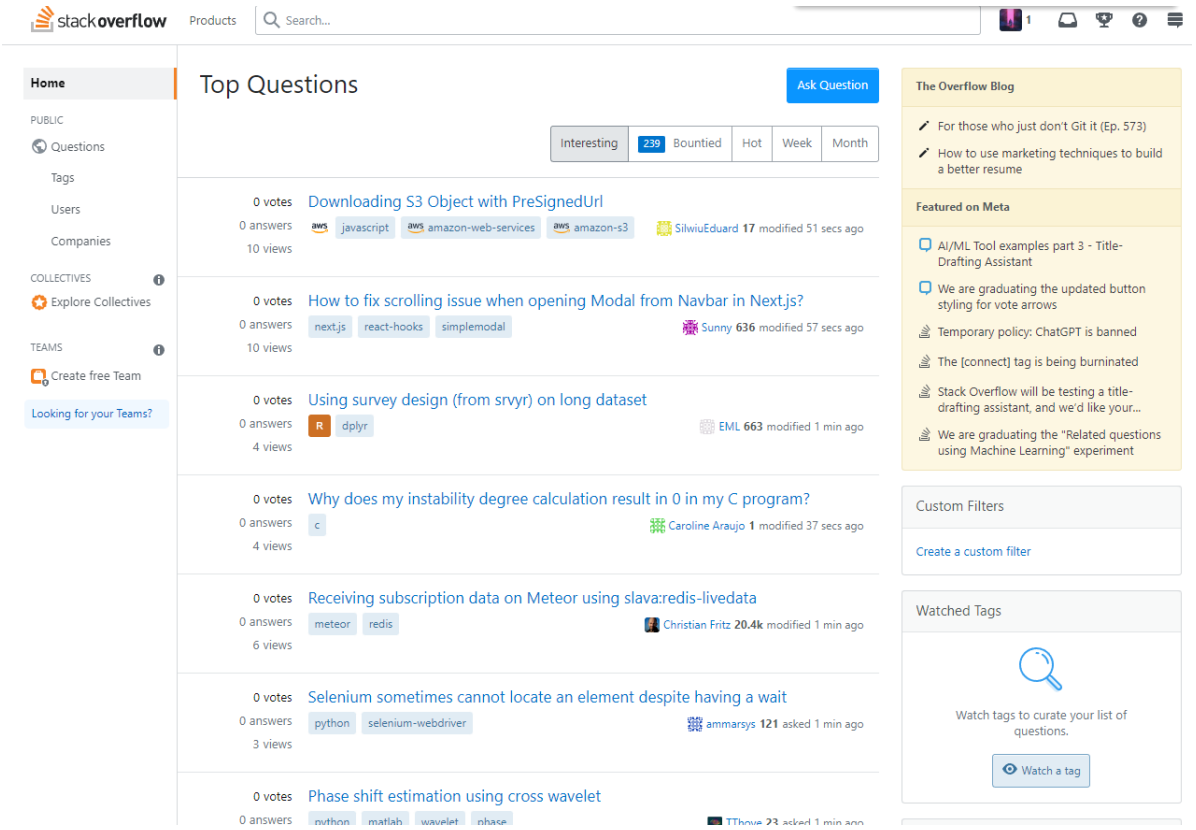
2.1 Stack Overflow

2.1.1 Γενική Περιγραφή

Το Stack Overflow (Εικόνα 1) είναι μία ιστοσελίδα ερωτοαπαντήσεων (Q&As) η οποία ενώνει τους ανθρώπους που ασχολούνται με την Πληροφορική, σε παγκόσμιο βαθμό. Εξυπηρετεί ως πηγή πληροφορίας για τον κόσμο της τεχνολογίας και του προγραμματισμού και διευκολύνει την αμοιβαία συνεργασία, την λύση προβλημάτων και την διαμοίραση πληροφορίας σε μεμονωμένους ανθρώπους, ομάδες και οργανισμούς. Το Stack Overflow αποτελεί ναυαρχίδα στο Stack Exchange Network και φιλοξενεί πάνω από 100 εκατομμύρια χρήστες κάθε μήνα. Κάθε χρήστης έχει την ικανότητα να ρωτήσει κάτι σχετικά με τον πληροφορικό τομέα και αναμένει μέχρι κάποια απάντηση από τους άλλους χρήστες να του είναι αποδεκτή. Οι συμμετέχοντες κατηγοριοποιούνται σε νέους χρήστες (new users), σε συντάκτες (editors) ή συντονιστές (moderators) όπου επίσης ο καθένας έχει επιπρόσθετα χαρακτηριστικά όπως το ψήφισμα (voting), το ιστορικό (history) και την φήμη (reputation). Οι χρήστες της ιστοσελίδας μπορούν να παραλάβουν βαθμούς φήμης και ‘’σήματα’’ (badges) όσο συνεισφέρουν στο σύνολο, κάτι που τους προσδίδει προνόμια όπως το να ψηφίζουν, να σχολιάζουν ακόμα και να επεξεργάζονται τις αναρτήσεις (ερωτήσεις ή σχόλια) άλλων χρηστών.

Το Stack Overflow επίσης εφαρμόζει ένα σύστημα ετικετών για να οργανώνει και να κατηγοριοποιεί αποτελεσματικά το άπλετο περιεχόμενο που διαθέτει. Οι ετικέτες αποτελούν έναν ολιγόλεκτο συσχετισμό που περιγράφει το θέμα ή την τεχνολογία που αναφέρεται η ερώτηση. Όταν οι χρήστες αναρτούν μία ερώτηση στην ιστοσελίδα, είναι αναγκασμένοι να αναθέσουν μία ή παραπάνω ετικέτες σε αυτήν, οι οποίες μπορούν να επιλεχθούν από ένα συγκεκριμένο σύνολο ετικετών ή έστω να δημιουργήσουν μια καινούρια. Σκοπός αυτής της πολιτικής είναι να επιτρέπει στους χρήστες να μπορούν να βρουν γρήγορα και αποτελεσματικά την ερώτηση σχετικά με μια συγκεκριμένη τεχνολογία, γλώσσα προγραμματισμού ή πλαίσια προγραμματισμού (frameworks). Επίσης, οι ετικέτες διευκολύνουν και τους χρήστες που διαθέτουν την τεχνογνωσία, οι οποίοι μπορούν να ακολουθούν συγκεκριμένες ετικέτες των ενδιαφερόντων τους και να λαμβάνουν ειδοποιήσεις και ενημερώσεις σχετικά με νέες ερωτήσεις και απαντήσεις σε αυτούς τους τομείς. Με αυτόν τον τρόπο οι ερωτήσεις των χρηστών έχουν μια γρήγορη και αξιόπιστη

απάντηση από πραγματογνώμονες και μη, πράγμα που καθιστά το Stack Overflow την νούμερα ένα ιστοσελίδα ερωτοαπαντήσεων παγκοσμίως.



Εικόνα 1. Κύρια Σελίδα του Stackoverflow.

2.1.2 Ιστορικά Στοιχεία

Οι δημιουργοί της πλατφόρμας είναι ο Jeff Atwood και ο Joel Spolsky και ιδρύθηκε τον Σεπτέμβριο του 2008, μέχρι και σήμερα είναι η πιο εξεζητημένη σελίδα ερωτοαπαντήσεων. Το όνομά της αποφασίστηκε μέσω ψηφοφορίας από αναγνώστες του blog, που άνηκε στον Atwood. Στις 3 Μαΐου 2010, είχαν συγκεντρωθεί μόλις 6 εκατομμύρια δολάρια σε επιχειρηματικό κεφάλαιο από μία ομάδα επενδυτών με επικεφαλή την Union Square Ventures καθώς και το 2019 ορίστηκε νέος διευθύνοντας σύμβουλος και νέα επικεφαλής της ιστοσελίδας, ο Prashanth Chandrasekar και η Teresa Dietrich αντίστοιχα [25].

2.1.3 Στατιστικά Στοιχεία

Το Stack Overflow αδιαμφισβήτητα καθίσταται η μεγαλύτερη σελίδα ερωτοαπαντήσεων καθώς προσμετράει 21 εκατομμύρια ερωτήσεις και 50.6

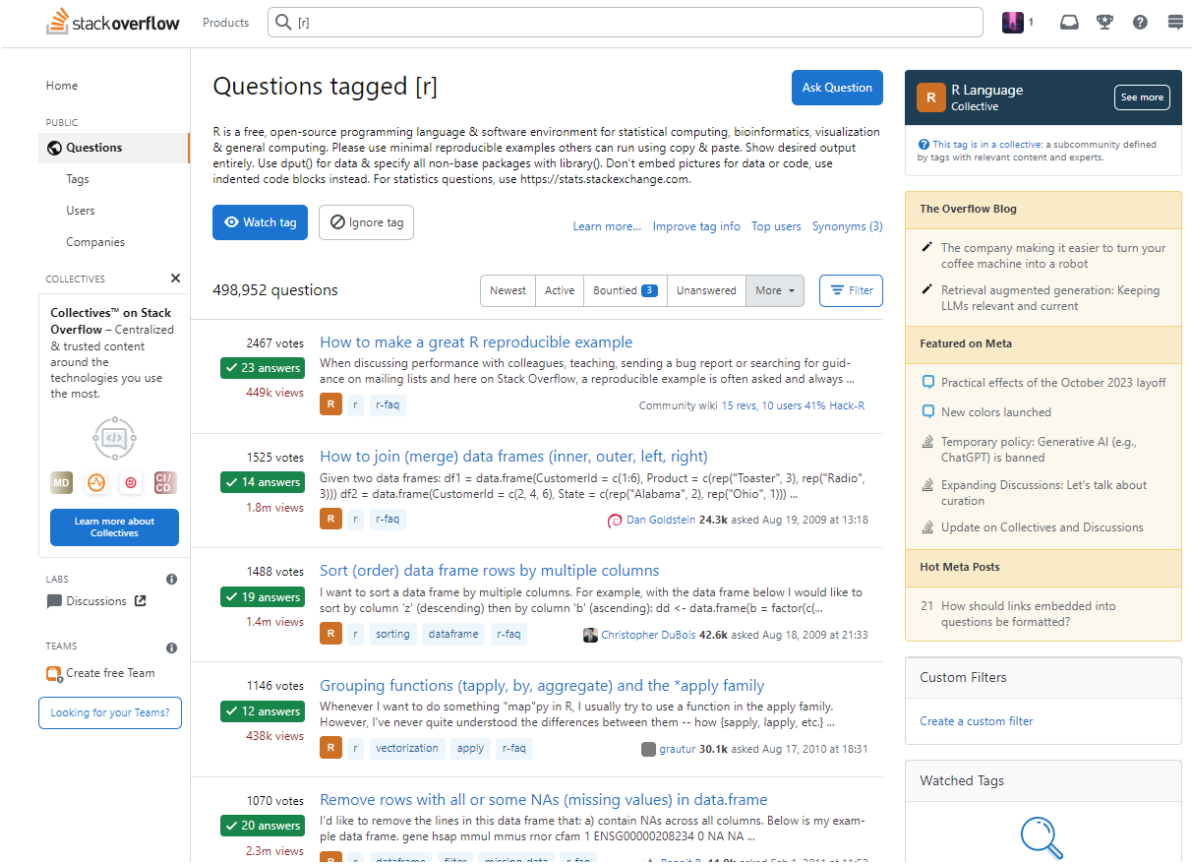
εκατομμύρια απαντήσεις από 100 εκατομμύρια χρήστες μηνιαίως, όπου ο καθένας επισκέπτεται κατά μέσω όρο το site 7 φορές τον μήνα. Το SO από το 2011 και κάθε χρόνο διεξάγει έρευνες και παρέχει στατιστικές πληροφορίες για την σελίδα, οι οποίες μπορούν να βρεθούν στην αντίστοιχη ιστοσελίδα του [15]. Το 2022 την πρώτη θέση στις πιο δημοφιλής τεχνολογίες κατέχει για δέκατη φορά η Javascript, και ακολουθούν στις αμέσως επόμενες θέσεις : η HTML/CSS, SQL, Python, Typescript και η Java. Τα αποτελέσματα των ερευνών και τα παραγόμενα σύνολα δεδομένων είναι διαθέσιμα σε όλους.

2.1.4 Τεχνολογία

Το StackOverflow έχει παραχθεί με έναν συνδυασμό από τεχνολογίες για να υποστηρίξουν την λειτουργικότητα και την επεκτασιμότητά του. Για γλώσσες προγραμματισμού χρησιμοποιήθηκαν οι C# στο back-end και η Javascript στο front-end. Για πλαίσια και βιβλιοθήκες (Frameworks and Libraries) επιλέχθηκαν το ASP.NET MVC (Model-View-Controller) και jQuery αντίστοιχα, καθώς και για βάση δεδομένων το Microsoft SQL Server. Για την υποδομή και την ανάπτυξη (DevOps) βοήθησαν το Docker, Kubernetes και Microsoft Azure.

2.1.5 StackOverflow και R community

Η R είναι μια πολύ δημοφιλής γλώσσα ανοιχτού κώδικα και ειδικότερα στους επιστήμονες δεδομένων, τους στατιστικολόγους, και τους ερευνητές. Το StackOverflow κατέχει έναν σημαντικό ρόλο στην υποστήριξή της και την ακμάζουσα κοινότητά της, και αποτελεί κύρια πηγή για τους χρήστες της. Στην Εικόνα 2, βλέπουμε την ιστοσελίδα Stackoverflow με αναζήτηση το tag της R και στα αποτελέσματά της παρατηρούμε πως αρχικά δίνει κάποιες γενικές πληροφορίες της R καθώς και συγκεκριμένες οδηγίες για το πως οι χρήστες θα πρέπει να υποβάλουν τις ερωτήσεις τους για να είναι πιο κατανοητές και ευανάγνωστες για τους υπόλοιπους χρήστες. Η ενεργητική της κοινότητα συμβάλλει συνεχώς στην ετικέτα της R, παρέχοντας ένα τεράστιο αποθετήριο από ερωτήσεις και απαντήσεις όσων αφορά τις τεχνικές της, τα πακέτα που διαθέτει, την χειραγώγηση και οπτικοποίηση των δεδομένων και την στατιστική ανάλυση. Σύμφωνα με τις έρευνες και τις συμμετοχές της παραπάνω ιστοσελίδας[16], από το 2008 μέχρι το 2021, έχουν ερωτηθεί πάνω από 400.000 ερωτήσεις και πάνω από 500.000 απαντήσεις, με το σύνολο των χρηστών να ξεπερνάει τους 120.000. Επίσης, η ιστοσελίδα παρέχει κι αρκετές ακόμα πληροφορίες οι οποίες είναι προσβάσιμες από όλους, ωστόσο η συγκεκριμένη εργασία χρησιμοποιεί δικά της δεδομένα.



Εικόνα 2. Σελίδα με αναζήτηση το tag R.

2.1.6 Σημασιολογική ανάλυση των δεδομένων.

Η ανάλυση επικεντρώνεται στην ανάπτυξη μια οντολογίας που θα αναπαριστάνει σε έναν ολοκληρωμένο βαθμό τα δεδομένα που συλλέχθηκαν από το Stackoverflow, περικλείοντας όλες τις ερωτήσεις που συσχετίζονται με την R και τις σχετικές πληροφορίες των χρηστών. Η οντολογία εξυπηρετεί σαν μια εμπλουτισμένη αναπαράσταση γνώσης, προσδιορίζοντας σχέσεις και χαρακτηριστικά για τις διάφορες οντότητες.

Η οντολογία περιλαμβάνει μια κεντρική οντότητα, τον χρήστη “R_User”, που αποθηκεύει σημαντικές πληροφορίες για τον χρήστη όπως “user_id” και το “display_name”. Το user id είναι ένας ξεχωριστός αριθμός «κλειδί» που ανατίθεται σε κάθε χρήστη καθώς το display name είναι το όνομα του χρήστη που εμφανίζεται στην ιστοσελίδα. Η οντότητα αυτή συνδέεται με άλλες δυο σημαντικές οντότητες, τις ερωτήσεις και τις απαντήσεις. Στην συγκεκριμένη εργασία επικεντρωνόμαστε μόνο στις ερωτήσεις και η οντότητα ονομάζεται “Questions” με τα εξής χαρακτηριστικά. Κάθε ερώτηση έχει έναν αριθμό κλειδί “question_id” το οποίο διαφέρει σε κάθε ερώτηση, “tags” είναι οι ετικέτες που ανατίθενται στην ερώτηση καθώς και το χαρακτηριστικό ‘’is_answered’’ που δηλώνει αν μια ερώτηση έχει

απαντηθεί. Το “view_count” δηλώνει τον αριθμό των προβολών και το “accepted_answer_id” αποτυπώνει τον αριθμό κλειδί της επιβεβαιωμένης απάντησης της ερώτησης, ο οποίος αριθμός αν είναι 0 σημαίνει πως δεν υπάρχει. Τα χαρακτηριστικά “up_vote_count” και “down_vote_count” τυπώνουν τον αριθμό των υπέρ και κατά ψήφων των χρηστών προς την ερώτηση αντίστοιχα, και έτσι προσμετράτε το “score” της ερώτησης, καθώς υπάρχει και ένας μετρητής απαντήσεων ‘’answer_count’’. Τα χαρακτηριστικά που αποτυπώνουν πληροφορία στον χρόνο είναι δυο, ‘’last_activity_date’’ και “creation_date”, το πρώτο δηλώνει την τελευταία ημερομηνία οποιασδήποτε ενέργειας στην ανάρτηση της ερώτησης καθώς το δεύτερο δηλώνει την ημερομηνία δημιουργίας της. Στην ιστοσελίδα κάθε ερώτηση έχει έναν τίτλο και το περιεχόμενο της ερώτησης τα οποία αποθηκεύονται ως “title” και “body”, καθώς και ένας σύνδεσμος για την εύκολη πρόσβαση προς αυτήν, “link”.Οι ερωτήσεις σχετίζονται με τους χρήστες με την ενέργεια “posted_by” οπότε αποθηκεύουν τον “user_id” και το “display_name” του χρήστη.

2.2 Ανασκόπηση Αρθρογραφίας και Συμβολή

2.2.1 Ανασκόπηση Αρθρογραφίας

Έχουν πραγματοποιηθεί πολλές εργασίες στο παρελθόν όπου γίνεται εξερεύνηση και ανάλυση μιας γλώσσας προγραμματισμού και την κοινότητά της στην ιστοσελίδα Stack Overflow.Όλες οι εργασίες που αναφέρονται είναι δημοσιευμένες στο διαδίκτυο.

Ο *Konstantinos Georgiou, Nikolaos Mittas, Lefteris Angelis και Alexander Chatzigeorgiou* [1] προσπαθούν να εξερευνήσουν το αντίκτυπο που είχε η πανδημία του Κορονοϊού του 2019 στις δραστηριότητες και τις αναρτήσεις των προγραμματιστών και χρηστών στο Stack Overflow. Συγκεκριμένα, προσπάθησαν να επικεντρωθούν στις αναρτήσεις που συσχετιζόντουσαν με τις προγραμματιστικές μεθοδολογίες και projects, τα οποία ασχολούνταν με την κατανόηση και παροχή λύσεων στην παγκόσμια κρίση υγείας Covid-19. Οι συγγραφείς ανέλυσαν τους τίτλους των αναρτήσεων με σκοπό να εξάγουν κάποια κύρια θέματα συζητήσεων με βάση αυτών. Ανακάλυψαν ότι υπήρχε αξιοσημείωτη αύξηση στο νούμερο των αναρτήσεων σχετικά με τον Κορονοϊό κατά την διάρκεια την πανδημίας. Επιπλέον, παρατήρησαν ότι οι προγραμματιστές είχαν επικεντρώσει το ενδιαφέρον τους σε συζητήσεις σχετικά με την ανάλυση των δεδομένων και την ανάπτυξη λογισμικού παρά σε άλλες τομείς της πληροφορικής. Σύμφωνα με τα ευρήματα της ερευνάς τους, οι συγγραφείς κατέληξαν ότι οι κοινότητες που μοιράζονται πληροφορίες και γνώσεις, όπως το StackOverflow, μπορούν να χρησιμοποιηθούν ως πλατφόρμες για ερευνητικές προσπάθειες σε κρίσιμα ζητήματα όπως οι πανδημία, και επίσης υποδεικνύει τον τρόπο που

συνδυάζουν τις ικανότητες και την τεχνογνωσία τους για την επίλυση του προβλήματος.

Οι *Hamed Tahmooresi, Abbas Heydarnoori, Alireza Aghamohammadi* [2] στην εργασία τους προσπαθούν να εξερευνήσουν την κοινότητα της Python και τις τάσεις-κυρίαρχα θέματα που την απασχολεί με το πέρασμα του χρόνου. Για να πετύχουν το σκοπό τους εξέτασαν τα βασικά θέματα που συζητούν οι χρήστες της Python στην ιστοσελίδα StackOverflow, μαζεύοντας δεδομένα από αναρτήσεις και χρησιμοποιώντας ένα μοντέλο ενσωματωμένων λέξεων (word embedding model) για να ανακαλύψουν τεχνολογίες που προσφέρονται από την γλώσσα και πως χρησιμοποιούνται από τους προγραμματιστές. Η έρευνα έδειξε ότι το μεγαλύτερο ποσοστό των συζητήσεων σχετίζονταν με τα βασικά χαρακτηριστικά της Python, προγραμματισμό ιστοσελίδων (web programming), και επιστημονικό προγραμματισμό. Επιπρόσθετα, οι συγγραφείς ξεχώρισαν διάφορες εποχικές τάσεις και διαφορές στην δημοφιλία διάφορων θεμάτων και τεχνολογιών χρησιμοποιώντας γνωστούς μεθόδους ανάλυσης τάσεων (trend analysis) και ομαδοποίησης (clustering). Τέλος, η έρευνα συνιστά τις αντίστοιχες επιλογές που παρέχει η Python σχετικά με τεχνολογίες που παρέχουν άλλες γλώσσες χρησιμοποιώντας την προσέγγιση ενός μοντέλου που ονομάζεται word2vec.

Οι *Mohamad Yazdaninia, David Lo, Ashkan Sami* [3] προσπάθησαν να αναλύσουν τον αυξανόμενο αριθμό των “άλυτων” ερωτήσεων στο StackOverflow και να καθορίσουν την επιρροή που θα υπάρξει στην κοινότητα αν λυθούν αυτά. Η εργασία εξερευνά τα πιο σημαντικά χαρακτηριστικά που έχουν ορισμένες ερωτήσεις τα οποία προβλέπουν αν αυτές θα λάβουν επιβεβαιωμένη απάντηση. Επίσης, συμβάλλει στην εύρεση κάποιων κανόνων και συμπληρωμάτων που πρέπει να ακολουθεί μια ερώτηση για να είναι πιο πιθανό να λάβει μια επιβεβαιωμένη απάντηση, καθώς και ποιες ετικέτες και θεματικές του προγραμματισμού επηρεάζουν αυτές τις άλυτες ερωτήσεις. Οι συγγραφείς συλλέξανε αρκετές ερωτήσεις οι οποίες δεν είχαν καθόλου απαντήσεις, μόνο μη επιβεβαιωμένες απαντήσεις και με επιβεβαιωμένες απαντήσεις και έπειτα από ανάλυση, εξήγαν χαρακτηριστικά. Επιπρόσθετα, ανέλυσαν το υπόβαθρο των χρηστών, και χαρακτηριστικά των θεμάτων, όπως ετικέτες και badges, για να χαρακτηρίσουν τον ρόλο τους στις αναπάντητες ερωτήσεις. Η έρευνα συμπέρανε πως τα πιο σημαντικά χαρακτηριστικά που προβλέπουν ότι μια απάντηση θα παραμείνει αναπάντητη είναι ο αριθμός των προβολών της, ο αριθμός των γενικών απαντήσεων και ο αριθμός των σχολίων της. Παράλληλα, τα αποτελέσματα της εργασίας έδειξαν πως οι ετικέτες έχουν έναν πολύ σημαντικό ρόλο στις άλυτες ερωτήσεις, καθώς και πως οι χρήστες με παραπάνω badges τείνουν να λαμβάνουν περισσότερες επιβεβαιωμένες απαντήσεις. Συμπερασματικά, η εργασία παρέχει χρήσιμες πληροφορίες για τους παράγοντες που επηρεάζουν τις αναπάντητες ερωτήσεις στο StackOverflow και στο πως μπορούν να επιλυθούν.

Οι *Partha Chakraborty, Rifat Shahriyar, Anindya Iqbal, Gias Uddin* [4], εξερευνούν τους τρόπους με τους οποίους οι προγραμματιστές προσπαθούν να βρούνε βοήθεια και υποστήριξη για νέες γλώσσες προγραμματισμού στο StackOverflow. Η συγκεκριμένη εργασία εστιάζει στις γλώσσες Go, Swift, και Rust, και χρησιμοποιούν έναν συνδυασμό από δεδομένα που συλλέχτηκαν από το StackOverflow και το Github για να αναλύσουν κάποια μοτίβα συμπεριφοράς των προγραμματιστών και την ανάπτυξη αυτών των γλωσσών στην πάροδο του χρόνου. Οι συγγραφείς ,από το σύνολο δεδομένων του SO, ξεχώρισαν τα δεδομένα που συσχετίζονταν με αυτές τις γλώσσες χρησιμοποιώντας την ανάλογη ετικέτα, και κράτησαν όλες τις ερωτήσεις και επιβεβαιωμένες απαντήσεις αυτών. Επίσης εξήγαγαν όλα τα θέματα (issues) που αναφέρθηκαν στα αποθετήρια του Github σχετικά με τις τρεις αυτές γλώσσες. Τα συμπεράσματα τις εργασίας ανέδειξαν πως μια ενεργή κοινότητα μπορεί να επηρεάσει σε μεγάλο βαθμό την ανάπτυξη μιας γλώσσας προγραμματισμού καθώς επίσης σημείωσαν και την περίοδο όπου κάθε γλώσσα κατάφερε να μαζέψει αρκετούς πόρους για τους προγραμματιστές στις ιστοσελίδες ερωτοαπαντήσεων.

Οι *Pedro Almir M. Oliveira, Pedro A. Santos Neto, Gleison Silva, Irvayne Ibiapina, Werney L. Lira, Rossana M. C. Andrade* [5] στην εργασία τους προσπαθούν να εξάγουν μια κατατοπιστική προσέγγιση της αντιμετώπισης της προγραμματιστικής κοινότητας στην πανδημία του κορονοϊού Covid-2019. Για να το καταφέρουν αυτό, οι συγγραφείς διεξήγαγαν μια συστηματική εργασία συσχέτισης του Stackoverflow και του Github, δυο δημοφιλής online πλατφόρμες για συζητήσεις και συνεργασίες που αφορούν την ανάπτυξη λογισμικού. Οι συγγραφείς ακολούθησαν ένα σχέδιο εργασίας τεσσάρων επιπέδων, συμπεριλαμβάνοντας την συλλογή, την μοντελοποίηση, την σύνθεση και την ανάλυση των δεδομένων σχετικά με τον Covid-19. Τα αποτελέσματα της εργασίας ανέδειξαν πως η πανδημία είχε μεγάλο αντίκτυπο στην κοινότητα του προγραμματισμού, με μεγάλο ποσοστό των θεμάτων που απασχόλησαν το κοινό καθώς επίσης και των projects να αφορά την πανδημία στο Stackoverflow και στο Github αντίστοιχα. Οι συγγραφείς επίσης εντόπισαν κάποια θέματα και τάσεις σε εκείνη την περίοδο, όπως είναι η αυξημένη διαδικτυακή συνεργασία και συνομιλία ανάμεσα στους προγραμματιστές.

Στο ερευνητικό τους άρθρο, οι *Vishal Johri και Srividya Bansal* [6], παρουσιάζουν μια μεθοδολογία ανάλυσης δεδομένων γραπτού λόγου, που βρίσκεται σε ιστοσελίδες ερωτοαπαντήσεων, με επίκεντρο το StackOverflow. Οι συγγραφείς προσπαθούν να επεκτείνουν τις γνώσεις τους όσων αφορά τις ανάγκες των προγραμματιστών και να εντοπίσουν επίκαιρα θέματα που τους κινούν το ενδιαφέρον εφαρμόζοντας τον αλγόριθμο Latent Dirichlet Allocation (LDA). Αποκαλύπτουν κυρίαρχα θέματα που αμφιταλαντεύουν την κοινότητα, και διεξάγουν περαιτέρω ανάλυση πάνω σε αυτά, όπως την αξιολόγηση της

δημοφιλίας τους, τον υπολογισμό του αντίκτυπου που έχουν, τις τάσεις της εκάστοτε εποχής και τις συσχετίσεις μεταξύ τους. Επιπρόσθετα, συγκρίνουν διάφορες τεχνολογίες μεταξύ συγκεκριμένων τομών, όπως για παράδειγμα HTML/CSS, Javascript, jQuery και PHP στον τομέα του Web Development.Η ανάλυση είναι βασισμένη σε δεδομένα του Stackoverflow και η έρευνα παρέχει χρήσιμες πληροφορίες για τα πρότυπα χρήσης των τεχνολογιών και τον χρόνο που χρειάζονται αυτές ώστε να υιοθετηθούν.

Οι Nischal Shrestha, Colton Botta, Titus Barik, Chris Parnin [7] διεξήγαν μια εμπειρική έρευνα που αφορά ερωτήσεις του Stack Overflow μεταξύ 18 διαφορετικές γλώσσες προγραμματισμού για να διαπιστώσουν αν οι προγραμματιστές δυσκολεύονται να μάθουν επιπρόσθετες γλώσσες. Επίσης, υποθέτουν πως οι ήδη υπάρχουσες γνώσεις ενός ατόμου είναι πολύ πιθανόν να επεμβαίνουν στην εκμάθηση νέων γλωσσών και τεχνολογιών. Οι συγγραφείς βρήκαν 276 παραδείγματα παρεμβάσεων στην ιστοσελίδα που οδήγησαν σε λάθος υποθέσεις που προέρχονταν από γνώσεις άλλης γλώσσας. Για να καταλάβουν καλύτερα γιατί συνέβησαν οι δυσκολίες αυτές, οργάνωσαν ημιδομημένες συνεντεύξεις με 16 επαγγελματίες προγραμματιστές. Τα αποτελέσματα των συνεντεύξεων ανέδειξαν πως οι προγραμματιστές προσπάθησαν, λανθασμένα φυσικά, να συσχετίσουν την “νέα” γλώσσα προς αυτούς με αυτήν που ήδη ήξεραν. Επιπλέον, στην συγκεκριμένη εργασία προτείνονται ιδέες σχεδιασμού για τεχνικούς συγγραφείς, για ανθρώπους που φτιάχνουν τεχνικά εργαλεία και σχεδιαστές γλωσσών.

Οι *Maria Papoutsoglou, Nikolaos Mittas και Lefteris Angelis* [8], στην εργασία τους, προτείνουν ένα λογισμικό (framework) το οποίο συλλέγει, διαδικτυακά, αγγελίες δουλειών που είναι αναρτημένες στο StackOverflow και εξάγει τις προαπαιτούμενες δεξιότητες και γνώσεις καθώς και τις αρμοδιότητες της συγκεκριμένης δουλειάς. Οι συγγραφείς χρησιμοποιήσαν πολυμεταβλητή στατιστική ανάλυση για να ανακαλύψουν τυχόν συσχετίσεις στις δεξιότητες που ζητούνται και τις αρμοδιότητες μιας συγκεκριμένης δουλείας. Η μεθοδολογία της συγκεκριμένης εργασίας είναι να εντοπίσει και να χωρίσει τις ικανότητες-δεξιότητες σε Ήπιες (Soft Skills), Explicit Hard (R, Java) και Implicit Hard (προγραμματισμός, εξόρυξη δεδομένων), αναλύοντας τα κείμενα των αγγελιών και να τις συνδέσει με τα άτομα που ψάχνουν δουλειά.

Οι *Muhammad Asaduzzaman, Ahmed Shah Mashiyat, Chanchal K. Roy, Kevin A. Schneider* [9] προσπαθούν να καταλάβουν γιατί κάποιες ερωτήσεις στην ιστοσελίδα StackOverflow παραμένουν αναπάντητες καθώς επίσης και να κατηγοριοποιήσουν αυτές τις ερωτήσεις. Οι συγγραφείς επίσης διεξάγουν και ένα πείραμα για να διαπιστώσουν αν μπορούν να προβλέψουν πόσο καιρό θα παραμείνει μια συγκεκριμένη ερώτηση αναπάντητη. Τα αποτελέσματα τις

εργασίας εμφανίζουν πως υπάρχουν διάφοροι λόγοι όπου μια ερώτηση παραμένει αναπάντητη, όπως για παράδειγμα διπλότυπες ερωτήσεις, ερωτήσεις οι οποίες δεν είναι πια σχετικές ή υπάρχει ανάγκη για απάντηση καθώς και ερωτήσεις που δεν έχουν απάντηση.

Οι *Shaowei Wang, Tse-Hsun Chen, Ahmed E. Hassan* [10], στην εργασία τους προσπαθούν να αναλύσουν τα μοτίβα των χρηστών με τα οποία αυτοί αναθεωρούν και διορθώνουν απαντήσεις στο StackOverflow, και διερευνούν το αντίκτυπο του σήματος διόρθωσης (revision-related badge) που έχει αυτό στην ποιότητα και ποσότητα των διορθώσεων. Για να πετύχουν αυτόν τον στόχο, οι συγγραφείς συλλέξανε 3,871,966 διορθώσεις από 2,377,692 απαντήσεις στο SO και διεξήγαν ποσοτική και ποιοτική ανάλυση. Χρησιμοποίησαν στατιστικές μεθόδους για να εξετάσουν την σχέση ανάμεσα στον αριθμό και τον τύπο των διορθώσεων και την πιθανότητα μια απάντηση να αποσυρθεί καθώς και για ποιον λόγο. Οι συγγραφείς βρήκαν πως οι χρήστες αναρτούσαν παραπάνω διορθώσεις από το κανονικό όσον επρόκειτο να λάβουν κάποιο σήμα (badge) και σταματούσαν όταν το κατάφερναν.

Οι *Haoxiang Zhang, Shaowei Wang, Tse-Hsun Chen, Ahmed E. Hassan* [11], στην εργασία τους, εξερευνούν τους σχολιασμούς στις αναρτήσεις του Stack Overflow, αναλύοντας το περιεχόμενο τους και τους χρήστες που τα δημιούργησαν, ώστε να μπορέσουν να καταλάβουν καλύτερα την συμβολή τους στις ιστοσελίδες ερωτοαπαντήσεων. Οι συγγραφείς κατηγοριοποίησαν τους διάφορους τύπους θεμάτων που περιείχαν τα σχόλια και διερευνούν τα πλεονεκτήματα και μειονεκτήματα από τα πολλά είδη σχολίων ώστε να τα παραλληλίσουν με τις επίσημες οδηγίες του SO για την κατασκευή σχολίων. Τα αποτελέσματα της έρευνας έδειξε πως η πλειοψηφία των εποικοδομητικών σχολίων, πράγματι ακολουθούν τους επίσημους “κανόνες” και πως οι χρήστες είναι αρκετά ενεργοί στον συγκεκριμένο τομέα. Προτείνονται, επίσης, χρήσιμες ιδέες και βελτιώσεις στους σχεδιαστές και προγραμματιστές της σελίδας ώστε να μπορούν οι χρήστες της να μοιράζονται πληροφορίες ευκολότερα, όπως για παράδειγμα μηχανισμούς για αποτελεσματικότερη διοργάνωση των σχολίων για ευκολότερη ανάκτηση πληροφοριών και ευκολότερη διαχείριση της γνώσης.

Οι *John Sell, Feyzi Bagirov, Mark Newman, Laurel Lord* [12] διεξήγαν μια έρευνα για το πόσο χρόνο χρειάζεται για να αναρτηθεί μια αξιοπρεπή απάντηση στην ιστοσελίδα StackOverflow, για ερωτήσεις που σχετίζονται με τις προγραμματιστικές γλώσσες Python και R. Συλλέξανε αντίστοιχα δεδομένα από την σελίδα για ένα πεπερασμένο χρονικό διάστημα και εφάρμοσαν ανάλυση επιβίωσης (Survival Analysis) για να προβλέψουν δεδομένα ανταπόκρισης συσχετισμένα με τις εξής γλώσσες. Οι ερευνητές χρησιμοποίησαν πολυδιάστατη σχεδίαση για την ανάλυσή τους και εξερεύνησαν διάφορες λεπτομέρειες όπως τον αριθμό των συμβάντων (event counts), την ώρα μέχρι την πρώτη ανταπόκριση μιας

ερώτησης, την ώρα μέχρι την πρώτη επιβεβαιωμένη απάντηση και πολλά άλλα και τελικά βρήκαν πως καμία γλώσσα από τις δύο δεν διαπρέπει σε όλους τους τομείς. Τα συμπεράσματα ανέδειξαν πως ενώ η Python αναδείχθηκε η γλώσσα με την πιο ενεργή κοινότητα καθώς είχε τον μεγαλύτερο αριθμό απαντήσεων, η R αναδείχθηκε αυτή με την καλύτερη ποιότητα των απαντήσεων καθώς είχε καλύτερο βαθμό επιβεβαιωμένων απαντήσεων.

Οι *Jie Zou, Ling Xu, Weikang Guo, Meng Yan, Dan Yang, Xiaohong Zhang* [13] προσπάθησαν να αναλύσουν τις μη λειτουργικές απαιτήσεις (NFRs) που επικεντρώνονται οι προγραμματιστές κατά την διάρκεια της ανάπτυξης λογισμικού. Χρησιμοποίησαν το θεματικό μοντέλο (topic modeling) Latent Dirichlet Allocation (LDA), για να εξάγουν τα βασικά θέματα από τις συζητήσεις που συλλέξανε από την ιστοσελίδα StackOverflow, και της σύνδεσαν με τα NFRs μέσω μιας κατασκευασμένης λίστας λέξεων της επιλογής τους. Τα αποτελέσματα της έρευνας ανέδειξαν πως τα πιο πρόσφατα θέματα που συζητάν οι προγραμματιστές επρόκειτο για την χρηστικότητα και την αξιοπιστία ενώ λιγότεροι ασχολήθηκαν με την συντηρησιμότητα και την αποδοτικότητα. Αυτό αποδόθηκε στο ότι και τα περισσότερα άλυτα προβλήματα υπήρχαν επίσης στον τομέα της χρηστικότητας και της αξιοπιστίας και γι’ αυτό θα πρέπει όλοι να επικεντρωθούμε παραπάνω σε αυτά τα θέματα.

Οι Yunxiang Xiong, Zhangyuan Meng, Beijun Shen, Wei Yin στην εργασία τους προσπάθησαν να αποκτήσουν μια βαθύτερη κατανόηση στις συμπεριφορές των προγραμματιστών αναλύοντας δεδομένα από τις ιστοσελίδες, Stack Overflow και Github. Χρησιμοποίησαν μια μέθοδο δύο φάσεων, σύνδεση ταυτότητας (identity linkage) και την εξόρυξη συμπεριφοράς (behavior mining). Στην πρώτη φάση της σύνδεσης ταυτότητας, εξήγαν χαρακτηριστικά από τα προφίλ των χρηστών και δεδομένα για την συμπεριφορά τους, συμπεριλαμβανομένου και τις ομοιότητες ανάμεσα στα ονόματα χρηστών (usernames) και τους τρόπους γραφής αυτών. Έπειτα, χρησιμοποίησαν αλγορίθμους ομαδοποίησης και οπισθοδρόμησης δέντρων (classification and regression tree/CART) για να αντιστοιχίσουν τους λογαριασμούς των χρηστών ανάμεσα στο StackOverflow και το Github. Στην δεύτερη φάση εξόρυξης συμπεριφοράς, οι συγγραφείς έθεσαν τρία ερευνητικά ερωτήματα για την εξερεύνηση των μοτίβων πάνω στις συμπεριφορές των κοινοτήτων των δύο ιστοσελίδων. Χρησιμοποίησαν στατιστική, επεξεργασία φυσικής γλώσσας (natural language processing/NLP), και Machine Learning τεχνολογίες για την ανάλυση και την συγχώνευση συμπεριφορών. Απέδειξαν ότι η ακρίβεια της πρώτης μεθόδου της έρευνας είναι πιο υψηλή από παλαιότερες μεθόδους. Ανακάλυψαν, επίσης, πως οι πιο ενεργοί χρήστες του Github είναι οι ίδιοι οι οποίοι αναρτούν τις περισσότερες ερωτήσεις στο SO, καθώς επίσης πως και τους περισσότερους προγραμματιστές , τα θέματα από το περιεχόμενο τους στο Github είναι παρόμοια με αυτά από τις ερωτήσεις και απαντήσεις τους. Τέλος, η

έρευνα έδειξε πως τα θέματα που αμφιταλαντεύουν τους προγραμματιστές αλλάζουν με τα τωρινά project που ασχολούνται την προκειμένη περίοδο στο Github, καθώς επίσης και πως τα θέματα του Github αλληλοσυνδέονται περισσότερο με τις απαντήσεις τους ,παρά με τις ερωτήσεις και τα σχόλια στο SO.

Πίνακας 1. Συμβολή εργασιών σχετικών με την ανάλυση των ερωτήσεων.

Τίτλος εργασίας	Χαρακτηριστικά του Stack Overflow	Τεχνολογίες Σημασιολογικού ιστού	Στόχος Εργασίας
[1]	questions (title and body), tags, posting dates, number of received answers, identification number of the post owner	Web-scraping, python, R, Data manipulation, Data visualization.	Εξαγωγή πληροφορίας για το αντίκτυπο της πανδημίας Covid-19 στον τομέα της πληροφορικής.
[2]	Posts(title and body), tags, posting dates.	Python, web-programming, mobile application development, security, blockchain, Django.	Ανάλυση στα θέματα που αφορούν την Python στο Stack Overflow.Προσφορά εναλλακτικών τεχνολογιών άλλων γλωσσών προγραμματισμού, στην python.
[3]	Questions, answers, accepted answer, views, scores, number of answers, badges.	Machine-learning, SQL, XML, metadata, XGBoost, CART, Bayesian Ridge, Lasso, GNB.	Εξερεύνηση άλυτων ερωτήσεων. Παροχή online εργαλείου το οποίο προβλέπει αν μια ερώτηση θα έχει επιβεβαιωμένη απάντηση.
[4]	Questions, accepted answers, tags.	Go, Swift, Rust, C, C++, Java, Github, Objective-C, Javascript, migration.	Εξερεύνηση στο πως οι developers συζητούν και υποστηρίζουν νέες γλώσσες προγραμματισμού στο Stack Overflow.Επικέντρωση στις γλώσσες:Go, Swift, Rust.
[5]	Questions, tags, score of question .	Data science, web-scraping, data visualization, natural language processing, Python(NLTK), Javascript, R, Java.	Εξερεύνηση των ερωτήσεων του Stack Overflow και των repositories του Github για περιεχόμενο που θα βοηθούσε στην καταπολέμηση της πανδημίας και ανάλυση για πιθανή συσχέτιση ανάμεσα των δυο.
[6]	Questions, answers, tags, timestamps, post type.	Topic-modeling, REST-APIs , XML, MySQL database, Regular Expression, Data Structure/Algorithm, Website Design/CSS.	Γενική εξερεύνηση των ερωτήσεων του StackOverflow για υπόδειξη νέων τάσεων στην τεχνολογία και εύρεση κάποιου τομέα για πιθανή βοήθεια.

Τίτλος εργασίας	Χαρακτηριστικά του Stack Overflow	Τεχνολογίες Σημασιολογικού ιστού	Στόχος Εργασίας
[7]	Questions, answers, tags.	Python, Kotlin, Java, C++, Rust, SOTorrent, C#, Visual Basic, PHP.	Εξερεύνηση στις δυσκολίες που αντιμετωπίζουν οι προγραμματιστές στην εκμάθηση νέας γλώσσας προγραμματισμού, όταν κατέχουν την γνώση μιας άλλης.
[8]	Questions, answers, Users, user reputation, Job title, Company, Salary, Type of job, Free text, Joel text, Info tags, Badges, Benefits.	Data mining, e-recruitment, text mining, statical analysis, Data analysis, R, Python, Competence, problem solving.	Παράθεση πλαισίου για συλλογή αγγελιών από το StackOverflow και την εύρεση ικανοτήτων και συσχετίσεων που απαιτούν συγκεκριμένες δουλειές.
[9]	Questions, answers, tags, accepted answers, users, reputation, badges, votes(upvote, downvote), creation dates, comments.	Unanswered questions, classifiers, data-analysis, program complexity, management policy, server configuration, system administration, tagging.	Ανάλυση στων αναπάντητων ερωτήσεων του StackOverflow, και τον λόγο που παραμένουν έτσι. Προτάσεις για το πως οι χρήστες θα διαφοροποιήσουν τις ερωτήσεις τους για να αυξήσουν τις πιθανότητες ώστε να λάβουν απάντηση.
[10]	Questions, answers, comments, user badges, reputation, answer revision, dates, Rollback Body, Edit Body.	Text Correction, Text Description, Code Correction, Code Functionality, Code Formatting, Code Addition/Removal, Reference Modification.	Εξερεύνηση στον τρόπο που οι χρήστες του StackOverflow διορθώνουν και αναθεωρούν τις ερωτήσεις τους με επικέντρωση στο σύστημα revision badge.
[11]	Comments, questions, answers, reputation, votes.	Commenting guidelines, reward reputation, crowdsource knowledge, qualitive analysis, commenting thread, Answer Acceptance, Mann-Whitney U test,	Εξερεύνηση στην συνεισφορά που έχουν τα σχόλια στο StackOverflow.Προτάσεις για βελτιώσεις του συστήματος σχολίων για αποτελεσματική συντήρηση και οργάνωση στην γνώση.
[12]	Posts (Questions and Answers),Accepted Answers, Users, Votes, Comments, PostHistory, PostLinks.	Python , R, RStudio, Survival Analysis, Kaplan-Meier, Fleming-Harrington, Cox Proportional Hazard.	Εξερεύνηση στον χρόνο ανταπόκρισης και την λήψη εγκεκριμένων απαντήσεων στο StackOverflow, στις γλώσσες Python και R.
[13]	Posts(title and body),Comments(text), unanswered questions.	Non-functional requirements (NFRs), Topic model, Latent Dirichlet allocation (LDA).	Έρευνα στα non-functional requirements(NFRs) που επικεντρώνονται οι προγραμματιστές όταν συζητούν προγραμματιστικά θέματα στο Stackoverflow.
[14]	Posts, Users, Answers, comments, age, tags.	Identity Linkage, Developer Behavior Mining, Machine Learning, GitHub.	Ανάλυση των χρηστών και εξερεύνηση για παρόμοιες συμπεριφορές και συσχετίσεις ανάμεσα στο Stackoveflow και το Github.

Κεφάλαιο 3

Μεθοδολογία-Υλοποίηση

3.1 Εισαγωγή

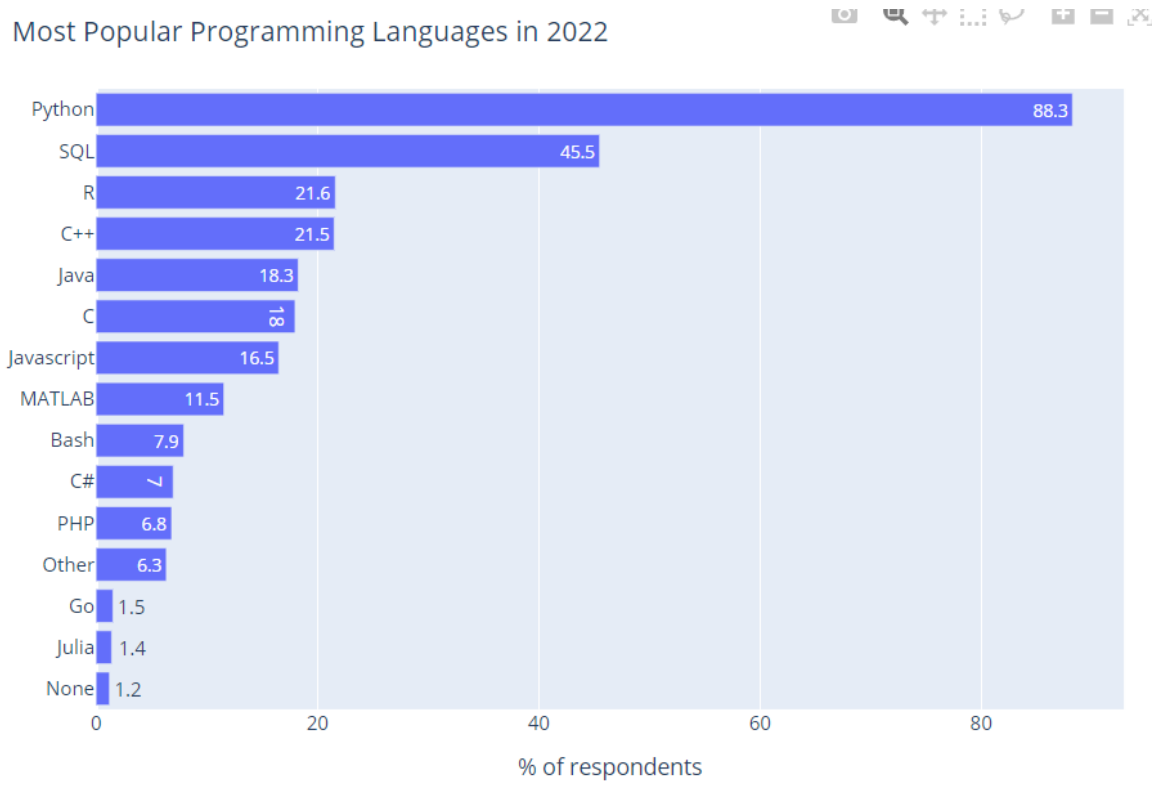
Στο Κεφάλαιο 1 προαναφέρθηκε πως η παρούσα εργασία σκοπεύει να αναλύσει όλες τις ερωτήσεις που καταχωρήθηκαν με ετικέτα την προγραμματιστική γλώσσα της R εξάγοντας κάποια βασικά περιγραφικά χαρακτηριστικά της κοινότητας της, καθώς επίσης και διάφορα θέματα που την αμφιταλαντεύουν. Η ανάλυση θα πραγματοποιηθεί χρησιμοποιώντας κάποιες μεθόδους στατιστικής ανάλυσης καθώς επίσης και αλγορίθμους ανάλυσης φυσικής γλώσσας (Natural Language Processing), όπως είναι ο LDA, NER και μεθόδους word-embedding. Στο παρόν κεφάλαιο, παρουσιάζονται οι χρησιμοποιημένες τεχνολογίες και αναλύονται η μεθοδολογία και η υλοποίηση που ακολουθήθηκε.

3.2 Τεχνολογίες που Χρησιμοποιήθηκαν

3.2.1 Η γλώσσα του προγραμματισμού Python

Η Python [17] είναι διερμηνευόμενη, γενικού σκοπού και υψηλού επιπέδου γλώσσα προγραμματισμού. Είναι μια από τις γλώσσες προστακτικού προγραμματισμού και υποστηρίζει τόσο το διαδικαστικό (procedural programming) όσο και τον αντικειμενοστραφή προγραμματισμό.

Η συγκεκριμένη γλώσσα έχει καταφέρει να αποκτήσει μεγάλη δημοτικότητα και ειδικότερα στην επιστήμη των δεδομένων. Το Kaggle έχει διοργανώσει μια παγκόσμια έρευνα τον Οκτώβριο του 2022 στην οποία ρωτήθηκαν 23.997 επαγγελματίες δεδομένων (data professionals) σχετικά με την επιστήμη των δεδομένων (Data Science), την Μηχανική Μάθηση (Machine Learning) και άλλα. Μια από τις ερωτήσεις της έρευνας ήταν «Ποια γλώσσα προγραμματισμού χρησιμοποιείται σε καθημερινή βάση», το 88.3% των ερωτηθέντων είχαν σαν απάντηση την Python. Στο γράφημα της Εικόνας 3 , που παρέχεται από την προαναφερθείσα έρευνα [18], γίνεται ευδιάκριτη η δημοτικότητα της γλώσσας.



Εικόνα 3. Δημοτικότητα γλωσσών προγραμματισμού το 2022 σύμφωνα με το Kaggle.

3.2.2 Natural Language Processing

Natural Language Processing [19] ή αλλιώς επεξεργασία φυσικής γλώσσας, είναι το ένα πεδίο το οποίο συνδυάζει την επιστήμη των υπολογιστών, γλωσσολογία και μηχανική μάθηση για να εξετάσει τον τρόπο τον οποίο επικοινωνούν και αλληλεπιδρούν ο άνθρωπος με τον υπολογιστή σε φυσική γλώσσα. Εμπεριέχει την χρήση υπολογιστικών τεχνικών για να επεξεργαστεί και να αναλύσει δεδομένα φυσικής γλώσσας, όπως κείμενα και ομιλίες, με στόχο να καταλάβει την σημασία των λέξεων. NLP δίνει την δυνατότητα, δηλαδή, στους υπολογιστές να καταλάβουν πλήρως την φυσική γλώσσα όπως ακριβώς την καταλαβαίνουν οι άνθρωποι.

3.2.3 Latent Dirichlet Allocation(LDA)

Latent Dirichlet Allocation [20] είναι μία διακεκριμένη «topic modeling» τεχνική που χρησιμοποιείται ευρέως στην επεξεργασία φυσικής γλώσσας (NLP) και στην μηχανική μάθηση. Εισήχθη από τον David Blei, τον Andrew Ng και τον Michael I. Jordan το 2003, ο LDA προσπαθεί να ανακαλύψει την υποκείμενη υποδομή από τεράστιες συλλογές από μη δομημένα κείμενα προσδιορίζοντας πιθανά θέματα και την κατανομή τους μέσα στα κείμενα. Η θεμελιώδης υπόθεση πίσω από το συγκεκριμένο μοντέλο είναι πως κάθε αρχείο είναι μια ανάμειξη από πολλαπλά

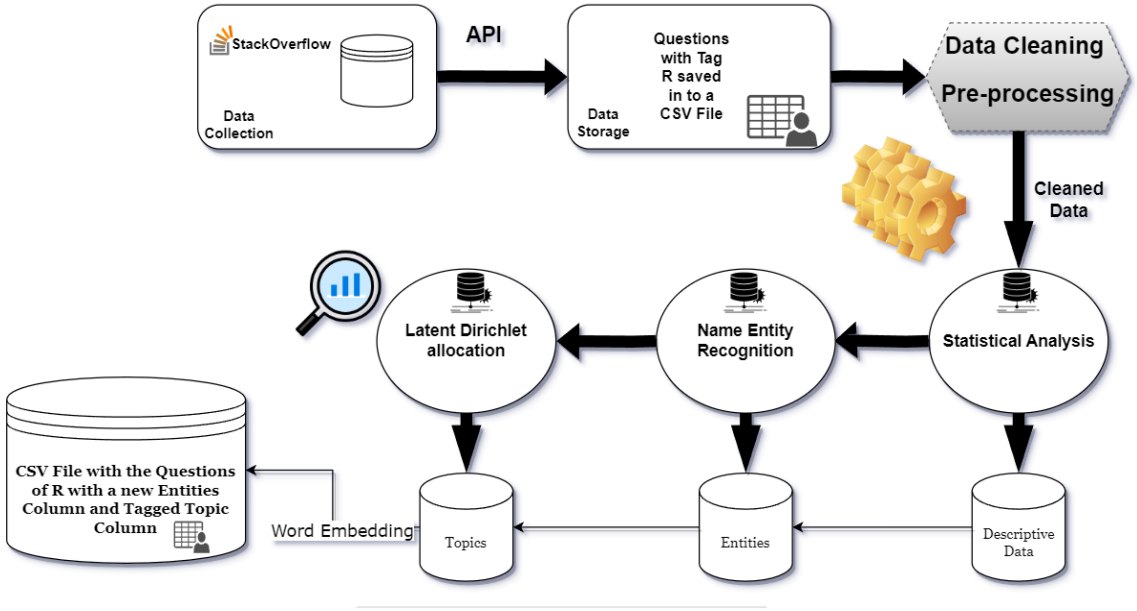
θέματα, και κάθε λέξη μέσα στο αρχείο παράγεται από ένα από αυτά τα θέματα. Ο λειτουργικός μηχανισμός αντιλαμβάνεται ως πιθανολογικό γενετικό μοντέλο. LDA υποθέτει πως υπάρχει συγκεκριμένος σταθερός αριθμός από θέματα (k) σε όλα τα κείμενα των αρχείων και κάθε θέμα εκπροσωπείται από μία κατανομή πιθανοτήτων πάνω στο λεξιλόγιο. Επιπλέον, κάθε αρχείο εκπροσωπείται από μία ανάμειξη αυτών των k θεμάτων που το ποσοστό του εκάστοτε θέματος δηλώνει την σημαντικότητα του στο αρχείο.

3.2.4 Name Entity Recognition (NER)

Name Entity Recognition (NER) [22] ή αλλιώς Αναγνώριση ονοματικών οντοτήτων, είναι μια βασική λειτουργία στο πεδίο Natural Language Processing (NLP) που επικεντρώνεται στην γνωστοποίηση και ταξινόμηση ονοματικών οντοτήτων σε μη δομημένα αρχεία κειμένου. Οι ονοματικές οντότητες είναι συγκεκριμένα στοιχεία στο κείμενο, όπως ονόματα από ανθρώπους, οργανισμούς, τοποθεσίες, ημερομηνίες, ποσότητες/ποσοστά και άλλα κατάλληλα επίθετα. Το NER έχει βασικό ρόλο σε διάφορες εφαρμογές της NLP καθώς χρησιμοποιείται για ανάκτηση πληροφορίας, συστήματα απάντησης ερωτήσεων, ανάλυση συναισθήματος (sentiment analysis) και για την κατασκευή γράφου πληροφορίας (knowledge graph).

3.3 Σχηματική Απεικόνιση Μεθοδολογίας

Η εκπόνηση της εργασίας έγινε με μια αλληλουχία βημάτων από τα οποία το καθένα παρείχαν σημαντικό ρόλο για την ολοκλήρωση του έργου και την έκβαση των τελικών αποτελεσμάτων. Το κάθε βήμα απεικονίζεται στην Εικόνα 4 και εξηγείται λεπτομερώς στην συνέχεια αυτού του κομματιού της εργασίας.

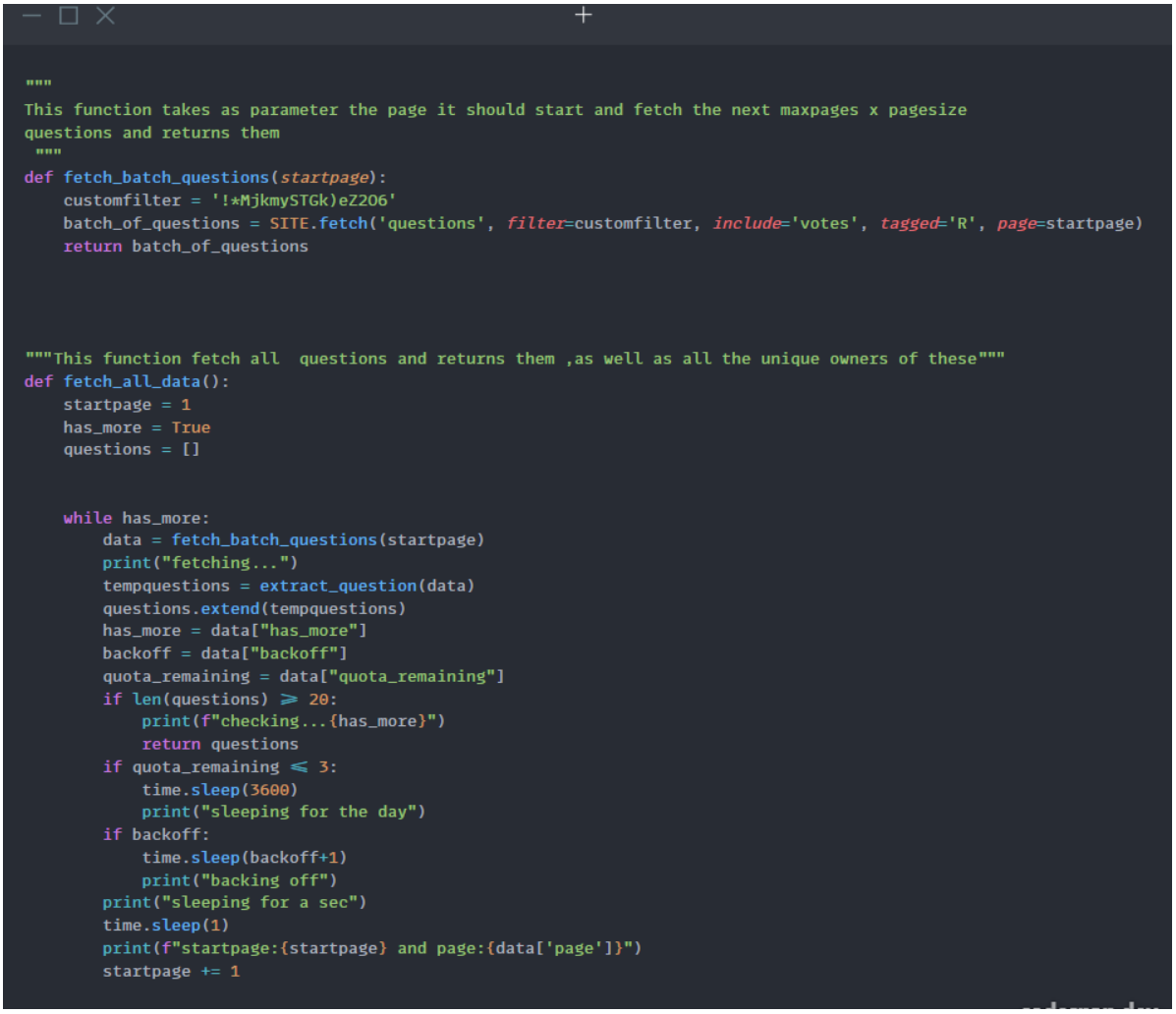


Εικόνα 4. Σχηματική απεικόνιση της εργασίας.

3.4 Συλλογή και Επεξεργασία Δεδομένων

3.4.1 Συλλογή δεδομένων από StackOverflow

Στο αρχικό στάδιο της συλλογής δεδομένων από το StackOverflow, ήταν να μπορέσω να λάβω πρόσβαση στο StackExchange API(Application Programming Interface) [21] κάτι που γίνεται έχοντας API key. Ο αρχικός στόχος ήταν να μαζέψω όλες τις ερωτήσεις με ετικέτα το R μαζί με επιπλέον όλα τα πεδία της κάθε ερώτησης που θα φανούν χρήσιμα, αφού αποθήκευαν χρήσιμές πληροφορίες για την ανάλυσή μας. Σύμφωνα με τους περιορισμούς του API, κάθε κλήση (call) του μπορεί να επιστρέψει μέχρι 100 σελίδες, όπου κάθε σελίδα περιλαμβάνει 100 ερωτήσεις. Με συγκεκριμένη μέθοδο που αναδεικνύεται στην Εικόνα 5 ο στόχος της συλλογής δεδομένων επιτεύχθηκε καθώς συλλέχθηκαν όλες οι απαιτούμενες ερωτήσεις και αποθηκεύτηκαν σχολαστικά στο αντίστοιχο CSV(Comma-separated values) αρχείο για περεταίρω ανάλυση.



```
"""
This function takes as parameter the page it should start and fetch the next maxpages x pagesize
questions and returns them
 """
def fetch_batch_questions(startpage):
    customfilter = '!*MjkmySTGk)eZ2O6'
    batch_of_questions = SITE.fetch('questions', filter=customfilter, include='votes', tagged='R', page=startpage)
    return batch_of_questions



"""This function fetch all  questions and returns them ,as well as all the unique owners of these"""
def fetch_all_data():
    startpage = 1
    has_more = True
    questions = []


    while has_more:
        data = fetch_batch_questions(startpage)
        print("fetching...")
        tempquestions = extract_question(data)
        questions.extend(tempquestions)
        has_more = data["has_more"]
        backoff = data["backoff"]
        quota_remaining = data["quota_remaining"]
        if len(questions) >= 20:
            print(f"checking...{has_more}")
            return questions
        if quota_remaining <= 3:
            time.sleep(3600)
            print("sleeping for the day")
        if backoff:
            time.sleep(backoff+1)
            print("backing off")
        print("sleeping for a sec")
        time.sleep(1)
        print(f"startpage:{startpage} and page:{data['page']}")
        startpage += 1
```

Εικόνα 5. Κώδικας που χρησιμοποιήθηκε για την συλλογή δεδομένων.

Πίνακας 2. Πεδία που συλλέχθηκαν με τα API.

Στήλη	**Εξήγηση**
Tags	Οι ετικέτες που είναι επισημασμένες στην ερώτηση.
Is_answered	Αληθής/Ψευδής το αν μια ερώτηση έχει απαντηθεί.
View_count	Αριθμός προβολών.
Accepted_answer_id	Ξεχωριστός αριθμός επιβεβαιωμένης απάντησης.
Down_vote_count	Αριθμός αρνητικών ψήφων.
Up_vote_count	Αριθμός θετικών ψήφων.
Answer_count	Αριθμός απαντήσεων.
Score	Σκορ ερώτησης(άθροισμα θετικών-αρνητικών ψήφων).
Last_activity_date	Ημερομηνία τελευταίας ενέργειας.
Creation_date	Ημερομηνία δημιουργίας.

Στήλη	Εξήγηση
Question_id	Ξεχωριστός αριθμός ερώτησης.
Link	Διαδικτυακός σύνδεσμος ερώτησης.
Title	Τίτλος ερώτησης.
Body	Σώμα ερώτησης.
Owner_user_id	Ξεχωριστός αριθμός χρήστη της ερώτησης.
Owner_display_name	Όνομα χρήστη στην σελίδα SO.
Text	Κείμενο ερώτησης.

3.4.2 Προεπεξεργασία δεδομένων

Στην φάση της προεπεξεργασίας των δεδομένων, λήφθηκαν δραστικά μέτρα ώστε να επιβεβαιωθεί η ποιότητα και η συνάφεια αυτών. Αρχικά, εντοπίστηκαν και εξαλείφθηκαν διπλότυπες εισαγωγές δεδομένων για να διατηρηθεί η μοναδικότητα κάθε ερώτησης. Έπειτα, αφαιρέθηκαν ερωτήσεις με ατελής ή κενά πεδία για να εξασφαλισθεί η συνοχή. Ως κυρίαρχο σημείο της ανάλυσης να είναι ο τίτλος και το σώμα κάθε ερώτησης, επιβλήθηκε σχολαστική εκκαθάριση των συγκεκριμένων πεδίων καθώς αφαιρέθηκαν όλες οι ετικέτες HTML, ενσωματωμένου κώδικα, διαδικτυακοί σύνδεσμοι (links) και άλλα εξωτερικά στοιχεία. Αυτές οι ενέργειες ήταν απαραίτητες για να επιβεβαιωθεί πως η μεταγενέστερη ανάλυση θα είναι καθαρά στον κορμό της εκάστοτε ερώτησης.

3.4.3 Επεξεργασία δεδομένων

Αρχικά η ανάλυση της κάθε ερώτησης έγινε στον συνδυασμό του καθαρού κειμένου των στηλών τίτλου και σώματος (title and body). Στην ανάλυση NER δεν λήφθηκαν υπόψη οι οντολογίες που είχαν να κάνουν με ημερομηνίες, αριθμούς, ποσοστά κλπ. Χρησιμοποιήθηκε η τεχνολογία Spacy, μια δημοφιλής βιβλιοθήκη επεξεργασία φυσικής γλώσσας της python που υπάρχει ενσωματωμένη μέθοδος NER [23]. Στο επόμενο βήμα, χρησιμοποιήθηκε ο αλγόριθμος LDA στα ίδια δεδομένα καθώς χρησιμοποιήθηκαν όλα τα δεδομένα για το training-process του μοντέλου. Η κατάλληλη τιμή των k θεμάτων, ύστερα από πολλές δοκιμές, αποφασίστηκε να είναι k=10. Έπειτα, αφού εξήχθησαν τα θέματα ακολούθησε η διαδικασία της σύνδεσης των ερωτήσεων με ένα από τα θέματα αυτά. Χρησιμοποιήθηκε το ίδιο εκπαιδευμένο μοντέλο και για τα word embedding χρησιμοποιήθηκε το pretrained μοντέλο ,word2vec, Google-news-vectors. Με αυτόν τον τρόπο δημιουργήθηκε ένα νέο CSV αρχείο με καινούρια στήλη με όνομα topic, στο οποίο αναγράφεται το topic που είναι ανατεθειμένο στην ερώτηση της συγκεκριμένης σειράς.

Κεφάλαιο 4

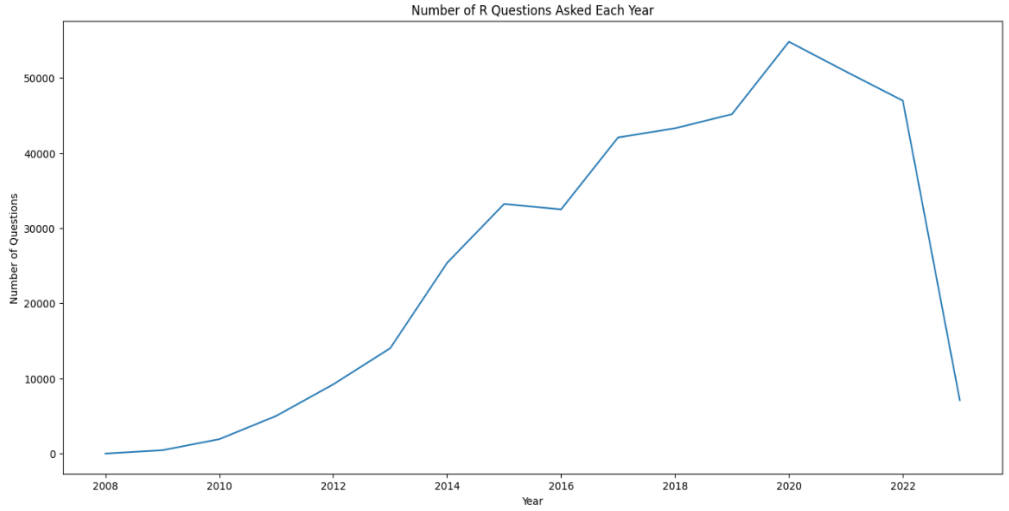
Αποτελέσματα

4.1 Εισαγωγή

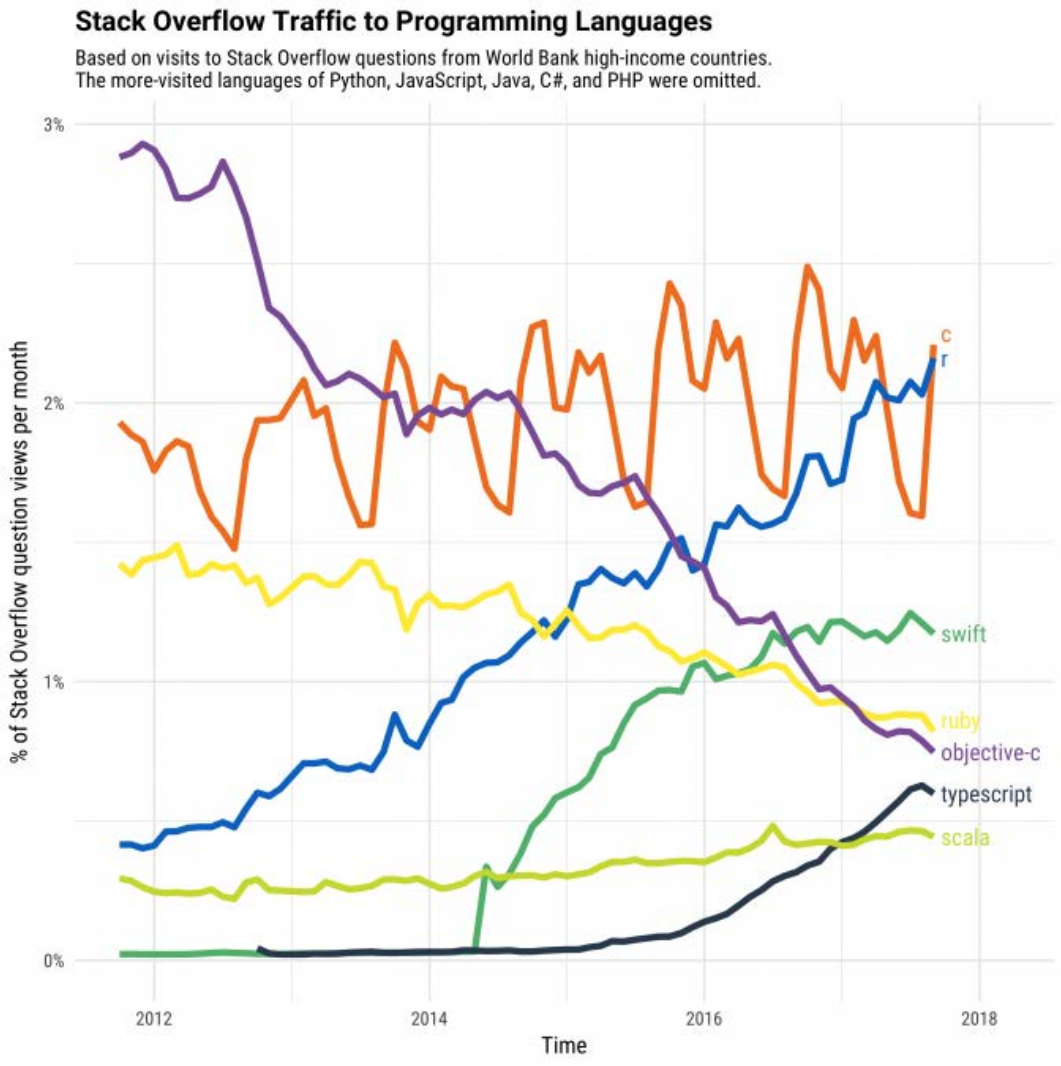
Στο συγκεκριμένο κεφάλαιο παρουσιάζονται και αναλύονται τα αποτελέσματα της έρευνας καθώς και κάποια περιγραφικά στοιχεία ώστε να αποδοθεί η μέθοδος και ο τρόπος της εξερεύνησης που ακολουθήθηκε για να καταλήξουμε στα σχετικά συμπεράσματα.

4.2 Περιγραφική Στατιστική-Εξερεύνηση

Όπως έχει προαναφερθεί νωρίτερα, η κοινότητα της R είναι μια από τις πιο ενεργή κοινότητες στο Stackoverflow με πάνω από 131.000 χρήστες. Η πρώτη ερώτηση που αναρτήθηκε ήταν 16 Σεπτεμβρίου το 2008 και η ερώτηση ήταν «How to access the last value in a vector» που σημαίνει πως να βρω πρόσβαση στην τελευταία τιμή ενός διανύσματος. Από τότε καταγράφηκαν εκατομμύρια ερωτήσεις που είναι συσχετισμένες με την R μέσα στον χρόνο. Στην Εικόνα 7 παρατηρούμε την εξέλιξη της R, στην πάροδο του χρόνου, σε σχέση με άλλες δημοφιλή γλώσσες και αντιλαμβανόμαστε πως είναι η μόνη γλώσσα με μια αύξουσα σταθερή πορεία [26]. Στην Εικόνα 6 το διάγραμμα προβάλλει τον αριθμό των ερωτήσεων που είναι συσχετισμένες με την R κάθε χρόνο. Όπως παρατηρούμε, ο αριθμός των ερωτήσεων του κάθε χρόνου αυξάνεται σταθερά από το 2008, που έγινε η πρώτη ερώτηση, με τον μεγαλύτερο αριθμό των ερωτήσεων να είναι το 2020. Αυτό, εξυπονοεί ότι η κοινότητα της R στην σελίδα Stack Overflow μεγαλώνει και γίνεται πιο ενεργή με τον καιρό.

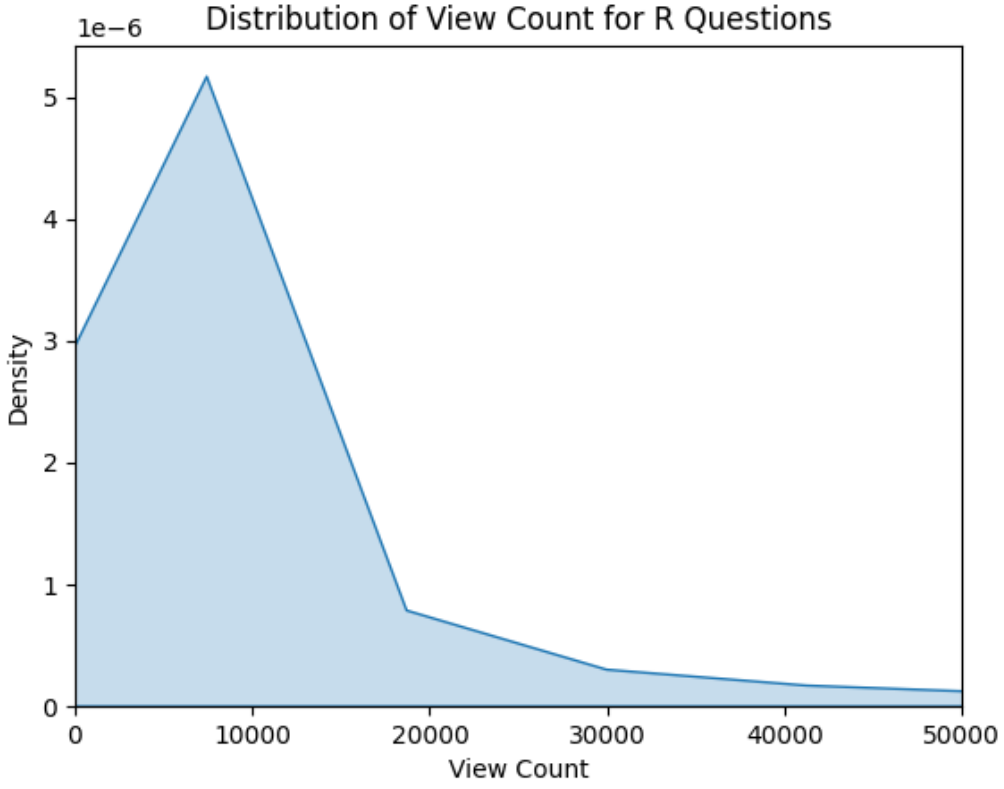


Εικόνα 6. Line Chart με τον αριθμό ερωτήσεων κάθε χρόνο.



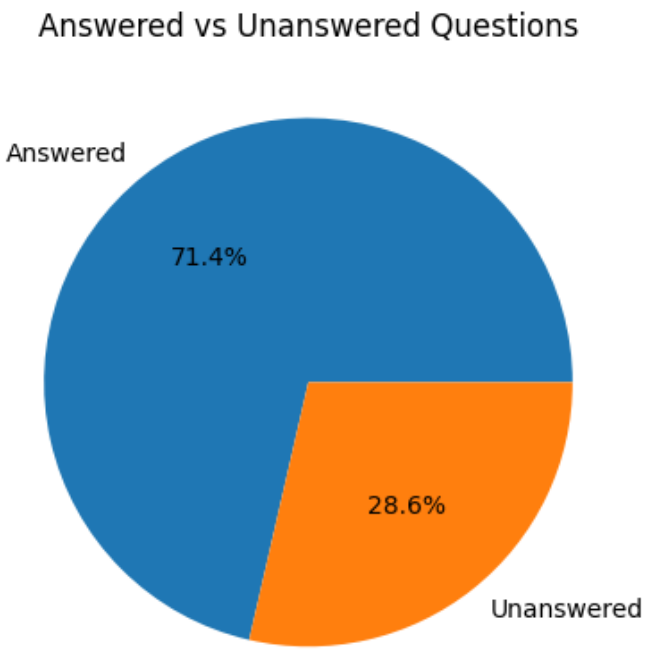
Εικόνα 7. Χρήση της R σχετικά με τις άλλες γλώσσες στην σελίδα SO [26].

Αντιλαμβανόμαστε λοιπόν ότι η κοινότητα από το 2008 έχει κάνει ραγδαίες αυξήσεις και εξελίσσεται ακόμα και σήμερα. Οι προβολές των ερωτήσεων που είναι συσχετισμένες με την R φτάνουν τεράστια νούμερα με την πρώτη να έχει τον αριθμό 2.236.210 προβολές. Στην Εικόνα 8 στο παρακάτω διάγραμμα, το οποίο στον οριζόντιο άξονα εμφανίζει τις τιμές των views και στον κάθετο άξονα την πυκνότητα ,παρέχει πληροφορίες για την διανομή των τιμών των views στις ερωτήσεις που είναι συσχετισμένες με την γλώσσα R στο site Stack Overflow.Από αυτήν την απεικόνιση μπορούμε να καταλάβουμε ότι η πλειοψηφία των ερωτήσεων έχουν σχετικά χαμηλές προβολές, καθώς κορυφή της πυκνότητας βρίσκεται στις 10.000 προβολές. Ωστόσο, υπάρχει μια μεγάλη «ουρά» στην πυκνότητα όπως βλέπουμε ,που ξεκινάει από τα 20.000 και συνεχίζει, το οποίο σημαίνει ότι ένα μικρό ποσοστό των ερωτήσεων έχουν σημαντικά μεγαλύτερο αριθμό προβολών από τις υπόλοιπες. Αυτό υποδεικνύει ότι μερικά θέματα που συσχετίζονται με την R μπορεί να προκαλεί μεγαλύτερο ενδιαφέρον στην κοινότητα και να ελκύει μεγαλύτερο κοινό. Επιπρόσθετα, το σχήμα πυκνότητας φαίνεται χρήσιμο στο να αναγνωρισθούν ερωτήσεις που έχουν λάβει περισσότερη ή λιγότερη συμμετοχή ,το οποίο με την σειρά του θα μας αναδείξει ποιες θεματολογίες χρειάζονται περισσότερη προσοχή και βοήθεια.



Εικόνα 8. Density Plot της κατανομής των προβολών.

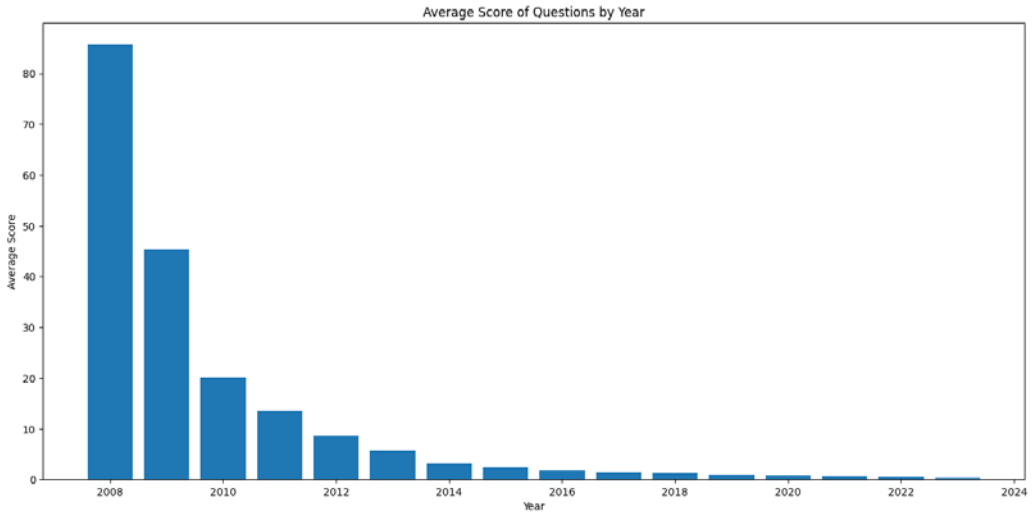
Η κοινότητα της R στο SO ωστόσο δεν είναι ενεργοί μόνο ως «θεατές» των ερωτήσεων καθώς όπως στην Εικόνα 9 του διαγράμματος ποσοστών (pie chart) που ανατίθεται, παρατηρούμε την αναλογία όλων των απαντημένων και των αναπάντητων ερωτήσεων και απευθείας αντιλαμβανόμαστε ότι η πλειοψηφία των ερωτήσεων ,περίπου το 70%, έχουν απαντηθεί ενώ το υπόλοιπο 30% παραμένει αναπάντητο. Ο αριθμός των απαντημένων ερωτήσεων καταφθάνει το 294.261 και των αναπάντητων 117.841. Αυτό υποδεικνύει ότι η Stack Overflow κοινότητα είναι γενικά ανταποκρίνεται στις ανάγκες των χρηστών της και είναι πρόθυμη να βοηθήσει στις ερωτήσεις τους, κάτι που είναι μια θετική ένδειξη ότι αυτή η κοινότητα είναι υποστηρικτική και συνεργατική. Ωστόσο, το ποσοστό που παραμένει αναπάντητο μπορεί υποδηλώνει τους τομείς όπου οι χρήστες δυσκολεύονται να βρουν λύσεις ή ακόμα και που υπάρχει έλλειψη τεχνογνωσίας ώστε να παρέχουν βοήθεια.



Εικόνα 9. Pie Chart Απαντημένων και Αναπάντητων ερωτήσεων.

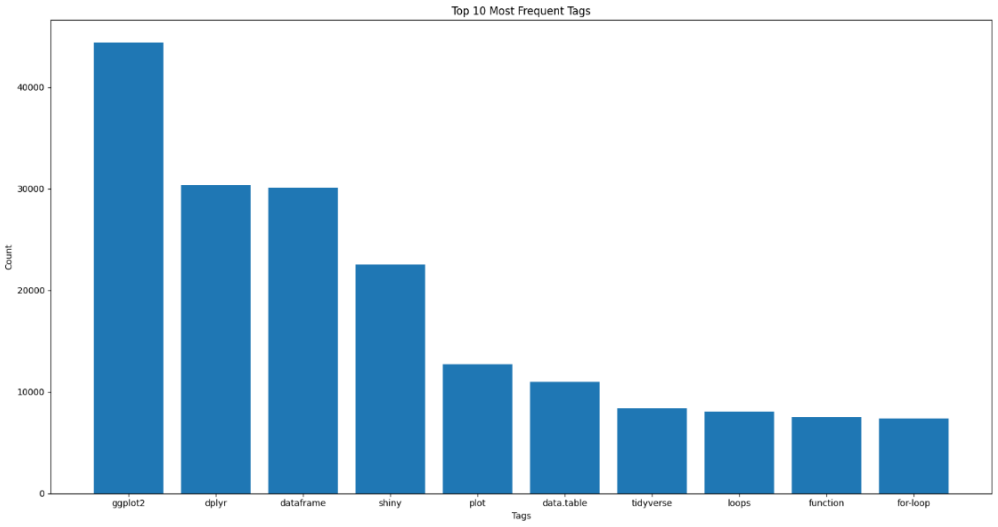
Υπάρχει επίσης και μια ποιότητα στην κοινότητα της R καθώς ένα μεγάλο ποσοστό των ερωτήσεων έχουν επιβεβαιωμένες απαντήσεις (56%). Ο αριθμός των

ερωτήσεων με επιβεβαιωμένη απάντηση είναι 231.215 και των ερωτήσεων με χωρίς είναι 180.887. Η ποιότητα των χρηστών διακρίνεται επίσης από τους ψήφους των ερωτήσεων καθώς ο μεγαλύτερος αριθμός θετικού ψήφου είναι 2476 και αρνητικού ψήφου 35. Από την άλλη μεριά, παρατηρώντας το διάγραμμα στην Εικόνα 10 καταλαβαίνουμε ότι ο μέσος όρος των scores τείνει να μειώνεται στο πέρας του χρόνου. Πιο συγκεκριμένα, υπάρχει μια γενική μείωση στον μέσο όρο των scores των ερωτήσεων, που σημαίνει ότι οι ερωτήσεις είναι πιο πιθανό να λάβουν λιγότερα υψηλά scores τα τελευταία χρόνια. Αυτό μπορεί να αποδοθεί σε διάφορους τομείς, όπως είναι η πτώση της δημοτικότητας της γλώσσας, η εφεύρεση νέων τεχνολογιών που αντικαθιστούν τις τεχνολογίες που είναι χρήσιμη η R ή ακόμα και στον τρόπο που βαθμολογούνται οι ερωτήσεις . Επιπρόσθετα, το line chart μπορεί να φανεί χρήσιμο για να προσδιοριστούν οι περίοδοι όπου οι ερωτήσεις τείνουν να λαμβάνουν περισσότερα ή χαμηλότερα scores,το οποίο θα μα υποδείξει το ποιόν της κοινότητας και τους τύπους των ερωτήσεων που είναι πιο αξιότιμες.



Εικόνα 10. Bar chart με το μέσο όρο των scores.

Όπως προαναφέρθηκε, η R επικεντρώνεται στην Ανάλυση δεδομένων και τα περισσότερα πιο δημοφιλή πακέτα που προσφέρει είναι χρήσιμα στο συγκεκριμένο πεδίο. Αυτό μπορούμε για άλλη μια φορά να το διακρίνουμε από τα πρώτα 10 πιο χρησιμοποιημένα tags στο dataset που αναλύουμε ,όπως φαίνεται στον bar chart στην Εικόνα 11. Παρατηρώντας το, το tag “ggplot2”, ένα αρκετά δημοφιλή πακέτο της R που χρησιμοποιείται για Απεικόνιση δεδομένων, κυριαρχεί στην λίστα.



Εικόνα 11. Bar chart με τα 10 πιο χρησιμοποιημένα tags εκτός το R ,στα δεδομένα μας.

4.3 Αποτελέσματα Ερωτήσεων

4.3.1 Εισαγωγή

Όπως προαναφέρθηκε νωρίτερα η εργασία επικεντρώνεται σε δύο είδη αναλύσεων. Σε αυτήν την ενότητα θα αναφερθούμε στην δεύτερη κατηγορία ανάλυσης η οποία είναι η ανάλυση φυσικής γλώσσας των ερωτήσεων που έχουν αναρτηθεί στην ιστοσελίδα Stackoverflow, με ετικέτα (tag) την προγραμματιστική γλώσσα R. Η ανάλυση που διεξήχθη είναι να εφαρμοστούν οι συγκεκριμένοι αλγόριθμοι NLP ώστε να εξαχθούν κάποιες συγκεκριμένες οντότητες και να διαπιστωθούν συγκεκριμένα *topics* που αναπαράγονται σε αυτές τις ερωτήσεις για να καταλάβουμε τα θέματα και τα σημεία που δυσκολεύουν τους χρήστες. Τέλος, θα ακολουθήσει η ανάθεση της κάθε ερώτησης σε ένα από τα παραπάνω θέματα ώστε να έχουμε μια πιο καθαρή αντίληψη στην κατανομή των θεμάτων μέσα στα δεδομένα.

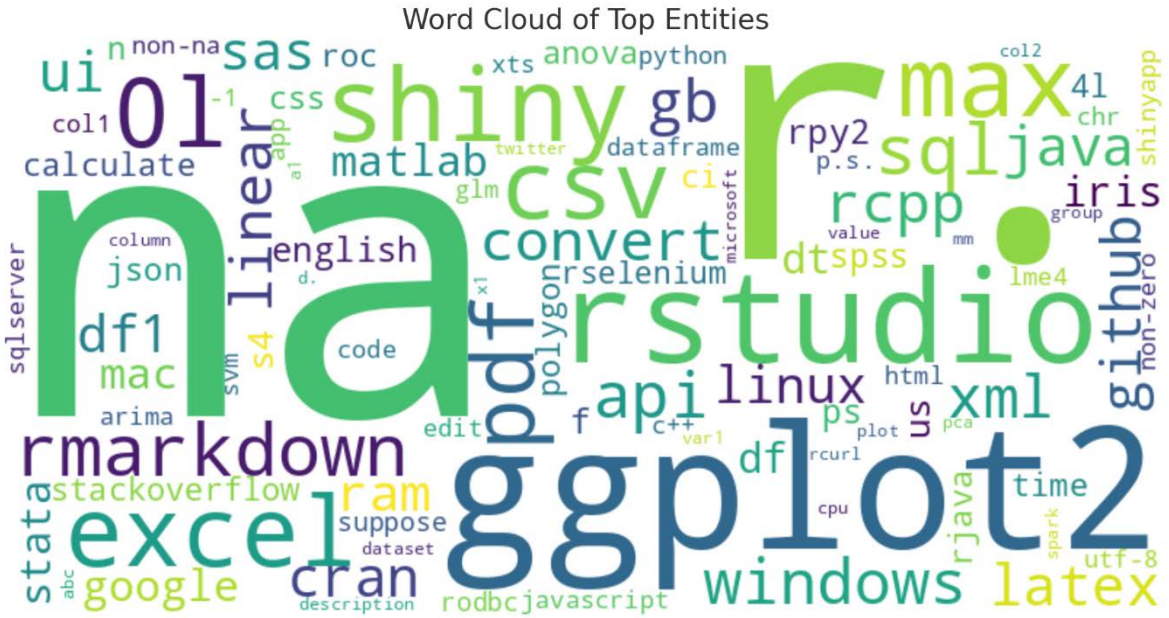
4.3.2 Αποτελέσματα Ανάλυσης NER (Name Entity Recognition).

Αρχικά, από την συλλογή των δεδομένων εξήχθησαν όλες οι οντότητες της κάθε ερώτησης και έπειτα ακολούθησε η μέτρηση εμφάνισης της εκάστοτε οντότητας. Έτσι, αναδείχθηκαν οι 50 πιο εμφανιζόμενες οντότητες και χαρακτηρίστηκαν οι κύριες οντότητες όλου του συνόλου(Εικόνα 12).

	Entity	Count
0	na	37477
1	r.	26950
2	ggplot2	22482
3	rstudio	11029
4	shiny	6407
5	max	5986
6	excel	5757
7	0l	5640
8	csv	5461
9	pdf	4391
10	rmarkdown	3970
11	sql	3864
12	api	3092
13	windows	2919
14	gb	2613
15	xml	2566
16	convert	2483
17	rcpp	2462
18	linear	2454
19	nan	2425
20	latex	2372
21	java	2222
22	github	2166
23	ui	2120
24	cran	2085
25	sas	1969
26	df1	1850
27	ram	1828
28	linux	1736
29	google	1719
30	matlab	1679
31	mac	1676
32	df	1595
33	dt	1584
34	stata	1570
35	iris	1542
36	english	1499
37	s4	1490
38	f	1475
39	rpy2	1418
40	us	1409
41	spss	1321
42	ps	1305
43	stackoverflow	1254
44	n	1250
45	calculate	1236
46	json	1233
47	4l	1197
48	polygon	1167
49	roc	1152

Εικόνα 12. 50 πιο εμφανιζόμενες οντότητες και αριθμός εμφάνισής τους.

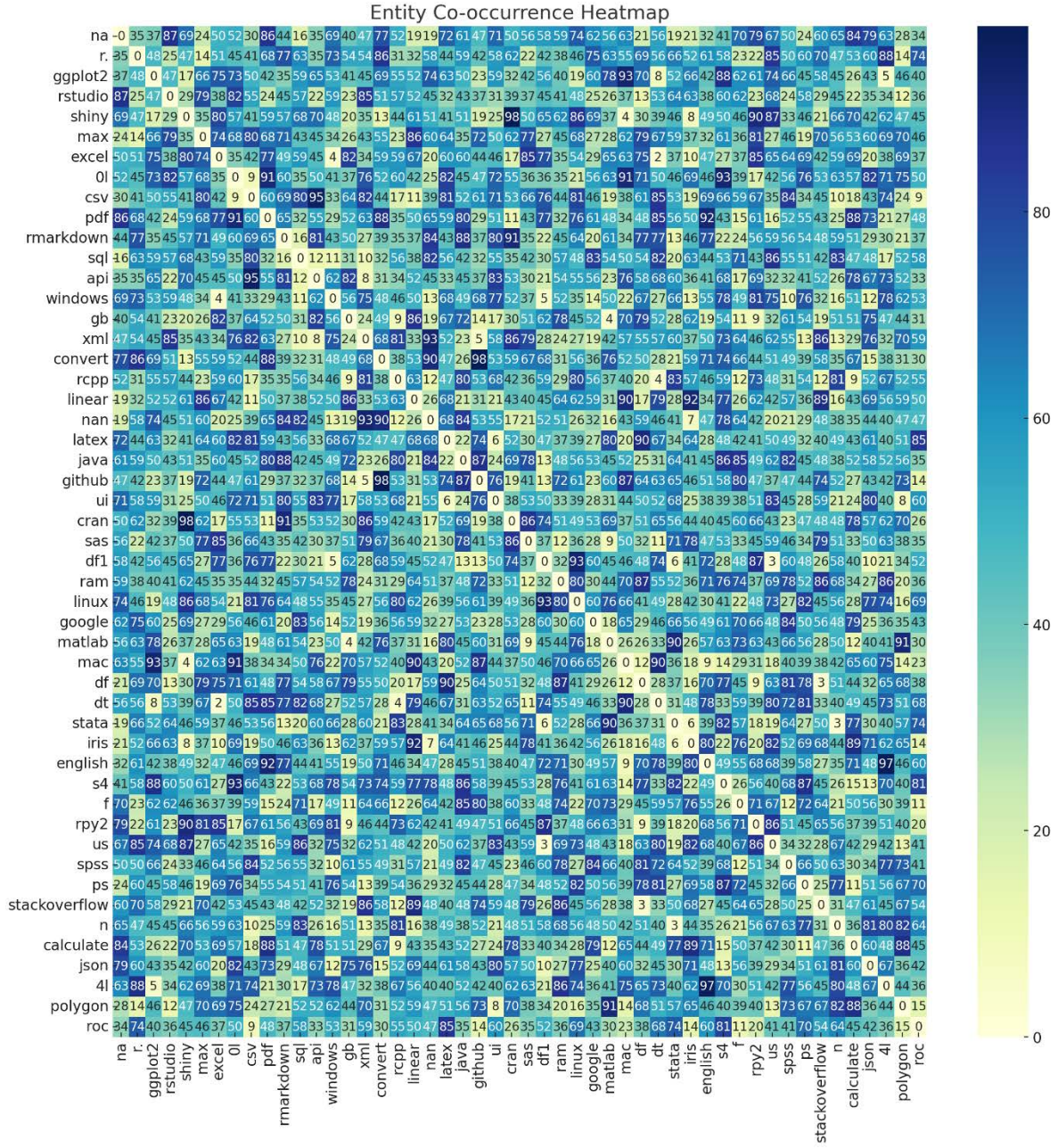
Δημιουργήθηκε ένα Word Cloud για την βέλτιστη οπτικοποίηση της συγκεκριμένης ανάλυσης του οποίου το μέγεθος κάθε λέξης (οντότητας) είναι ακριβώς ανάλογη με την συχνότητα αυτής στα δεδομένα. Με την πρώτη ματιά καταλαβαίνουμε ποια θέματα κυριαρχούν. Για παράδειγμα, λέξεις όπως “dataframe”, “column”, “function” είναι σημαντικά μεγαλύτερες, κάτι που υποδεικνύει πως είναι οι πιο συχνές οντότητες (Εικόνα 13).



Εικόνα 13. Word Cloud με τις πιο συχνές οντότητες.

Έπειτα, σχεδιάστηκε ένας πίνακας “Heatmap” (Εικόνα 14), ο οποίος αναδεικνύει την συνύπαρξη-συνεμφάνιση των οντοτήτων μέσα στην συλλογή των δεδομένων. Με τον όρο συνύπαρξη εννοούμε το πόσο συχνά δυο οντότητες εμφανίζονται μαζί στην ίδια ερώτηση. Κάθε κουτάκι του πίνακα αντιπροσωπεύει το σημείο τομής των δύο οντοτήτων και το χρώμα του, το πόσο συχνή είναι η συνεμφάνιση αυτών. Όσο πιο σκούρο είναι το κλουβί, τόσο πιο συχνή είναι η συνεμφάνιση των οντοτήτων. Κατά τον άξονα x και y, βλέπουμε την λίστα με τις οντότητες και στο σημείο που τέμνονται στο πλέγμα, είναι η τιμή της συνεμφάνισής τους. Παρατηρούμε πως οντότητες που αφορούν συγκεκριμένα πεδία της R σκιαγραφούνται πιο σκούρα τα κουτάκια τομής τους, όπως είναι για παράδειγμα το πεδίο DataFrame Operations, καθώς τα κουτάκια τομής των οντοτήτων “rstudio[1]”, “df[2]”, “na[3]”, “dt[4]” σκιαγραφούνται με σκούρο μπλε.

[1] [2] [3] [4] Είναι λεξιλόγιο της R ,όπου το κάθε ένα αναπαριστά αντικείμενο προγραμματισμού της.



Εικόνα 14. Heatmap με την συνύπαρξη των οντοτήτων.

4.3.3 Αποτελέσματα Ανάλυσης LDA

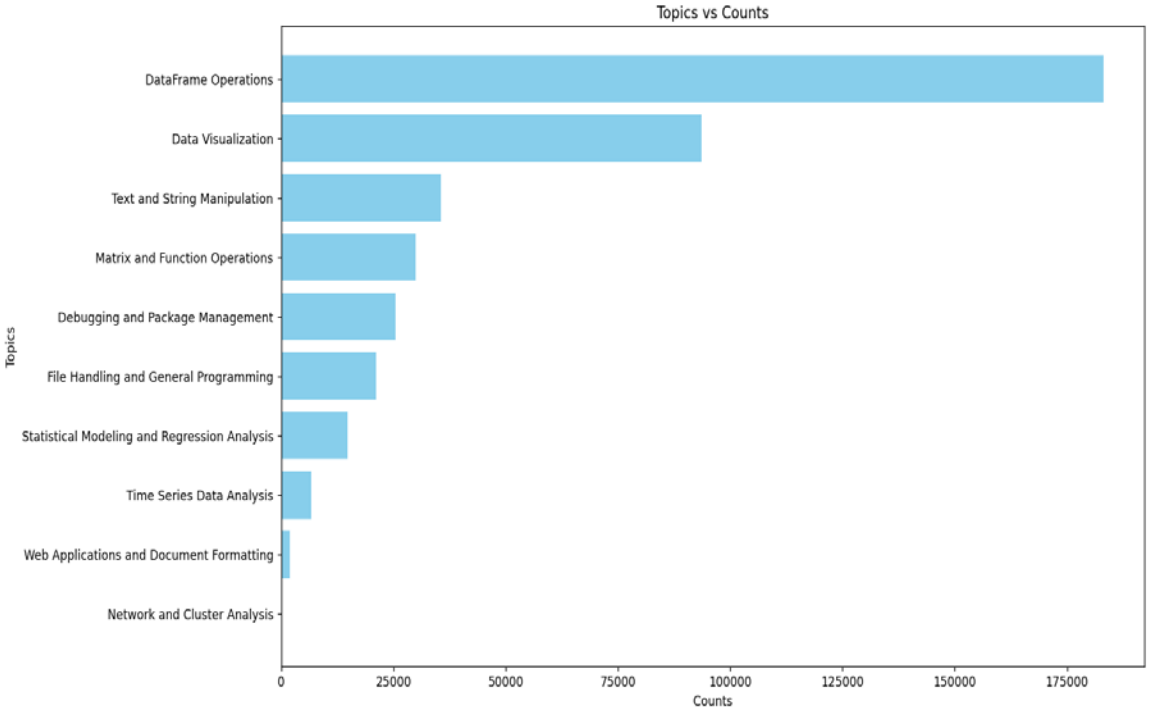
Στο δεύτερο βήμα της ανάλυσης πραγματοποιήθηκε ανάλυση LDA στον τίτλο και σώμα της κάθε ερώτησης. Έπειτα από διάφορες δοκιμές με την μεταβλητή k που προαναφέρθηκε για την εξαγωγή των θεμάτων, αποφασίστηκε ότι η κατάλληλη τιμή του να είναι 10. Τα θέματα παρουσιάζονται αναλυτικά στον Πίνακα 3 καθώς δημιουργήθηκε και ένα σχήμα στην Εικόνα 15 με την κατανομή θεμάτων στις

ερωτήσεις των δεδομένων. Σημαντική διαφορά στην κατοχή των ερωτήσεων παρατηρούμε να έχει το **DataFrame Operation** με αριθμό ερωτήσεων να καταφθάνει το **183.097** το οποίο είναι σχεδόν το 40% όλων των ερωτήσεων, κάτι που είναι απολύτως φυσιολογικό αφού ο προγραμματισμός με τα dataframes είναι μια από τις κύριες λειτουργίες της R καθώς επίσης υπάρχουν και πολλές δυσκολίες στο συγκεκριμένο πεδίο εξού και ο μεγάλος αριθμός των ερωτήσεων.

Πίνακας 3. Θέματα που εξάχθηκαν από τις ερωτήσεις.

Topic	**Keywords**	**Description**	**Questions**
Time Series Data Analysis.	time, series, date, forecast, trend, seasonal, xts, zoo, frequency, period.	Χειρισμός δεδομένων χρονολογικών σειρών,ανάλυση δεδομένων χρονολογικά, χειρισμός ελλειπόντων τιμών .	6.634
File Handling and General Programming.	file, read, write, import, export, csv, txt, script, execute, load.	Λειτουργίες αρχείων, χειρισμός, γενικές ερωτήσεις προγραμματισμού της R.	21.078
Statistical Modeling and Regression Analysis.	model, regression, statistic, linear, anova, glm, forecast, fit, coefficient, predict.	Στατιστική μοντελοποίηση, ειδικότερα ανάλυση παρεμβολής.	14.706
DataFrame Operations.	dataframe, operation, manipulation, rows, columns, join, merge, filter, select, groupby, column, data, row, value, frame.	Λειτουργίες σχετικές με τα DataFrame της R.Χειρισμός γραμμών, στηλών και τιμών μέσα στα data frames.	183.097
Data Visualization.	plot, graph, ggplot, chart, visualize, histogram, scatter, line, bar, color, ggplot2.	Οπτικοποίηση δεδομένων, δημιουργία των plots, graphs και άλλα.	93.619
Matrix and Function Operations.	matrix, function, apply, vector, calculate, operations, sapply, lapply, map, loop.	Λειτουργίες πινάκων ,συνάρτηση επαναλήψεων και άλλες λειτουργίες της R όπως vectors και λίστες.	29.901
Network and Cluster Analysis.	network, cluster, node, edge, graph, community,	Ανάλυση διαδικτύου και ομαδοποίησης δεδομένων.	58

Topic	Keywords	Description	Questions
	hierarchy, dendrogram, k-means, hclust.		
Debugging and Package Management.	error, debug, package, install, library, version, update, fix, issue, resolve.	Αποσφαλμάτωση, αντιμετώπιση προβλημάτων και χειρισμός των πακέτων.	25.496
Text and String Manipulation	text, string, manipulate, gsub, regex, pattern, extract, replace, substring, match.	Επεξεργασία κειμένου, και χειρισμός αλφαριθμητικών τιμών.	35.549
Web Applications and Document Formatting.	shiny, web, app, server, UI, html, pdf, markdown, rmarkdown, render.	Δημιουργία web εφαρμογών με την χρήση του framework “Shiny”.	1.964



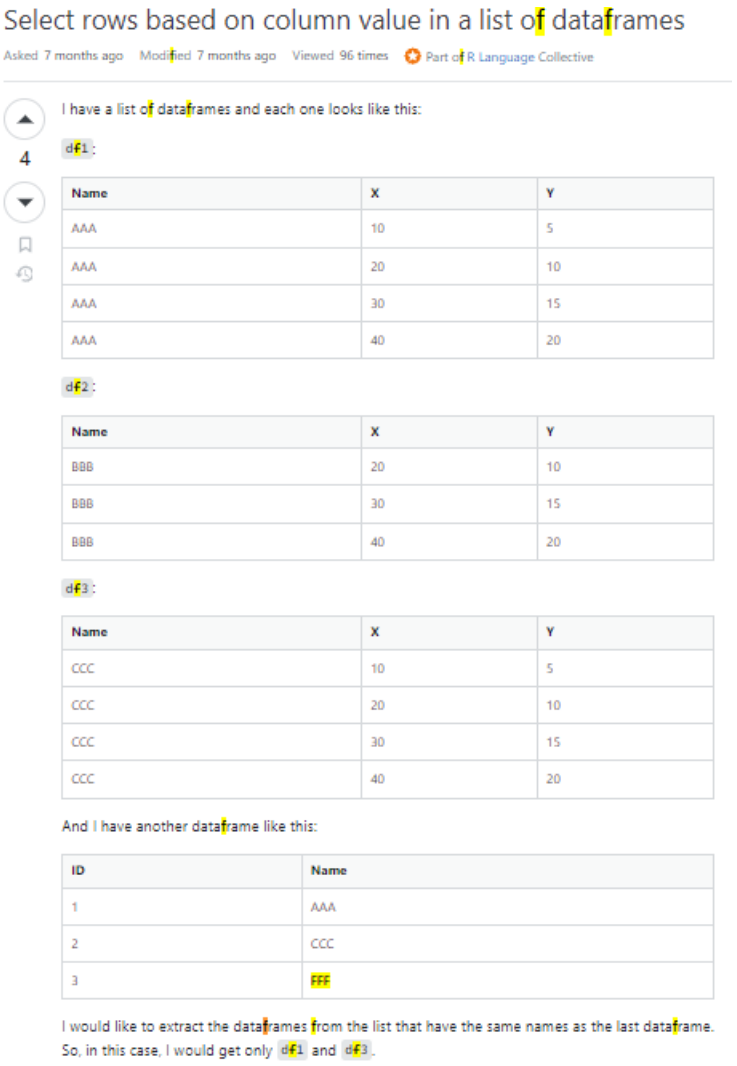
Εικόνα 15. Bar Chart με την κατανομή θεμάτων.

Ένα παράδειγμα χαρακτηριστικής ερώτησης χρηστών που αφορά Dataframe Operations μπορούμε να δούμε στον Πίνακα 4 καθώς και το στιγμιότυπο της

αυτούσιας ερώτησης στην Εικόνα 16.

Πίνακας 4. Παράδειγμα Ερώτησης DataFrame Operations

Τίτλος	*Select rows based on column value in a list of dataframes*
Σώμα	*I have a list of dataframes and each one looks like this:(*table of dataframe) *I would like to extract the dataframes from the list that have the same names as the last dataframe. So, in this case, I would get only df1 and df3*.
Link	https://stackoverflow.com/questions/75511910/select-rows-based-on-column-value-in-a-list-of-dataframes



Select rows based on column value in a list of dataframes

Asked 7 months ago Modified 7 months ago Viewed 96 times Part of R Language Collective

4

I have a list of dataframes and each one looks like this:

df1 :

Name	X	Y
AAA	10	5
AAA	20	10
AAA	30	15
AAA	40	20

df2 :

Name	X	Y
BBB	20	10
BBB	30	15
BBB	40	20

df3 :

Name	X	Y
CCC	10	5
CCC	20	10
CCC	30	15
CCC	40	20

And I have another dataframe like this:

ID	Name
1	AAA
2	CCC
3	FFF

I would like to extract the dataframes from the list that have the same names as the last dataframe. So, in this case, I would get only df1 and df3.

Εικόνα 16.Ερώτηση Dataframe Operation του SO.

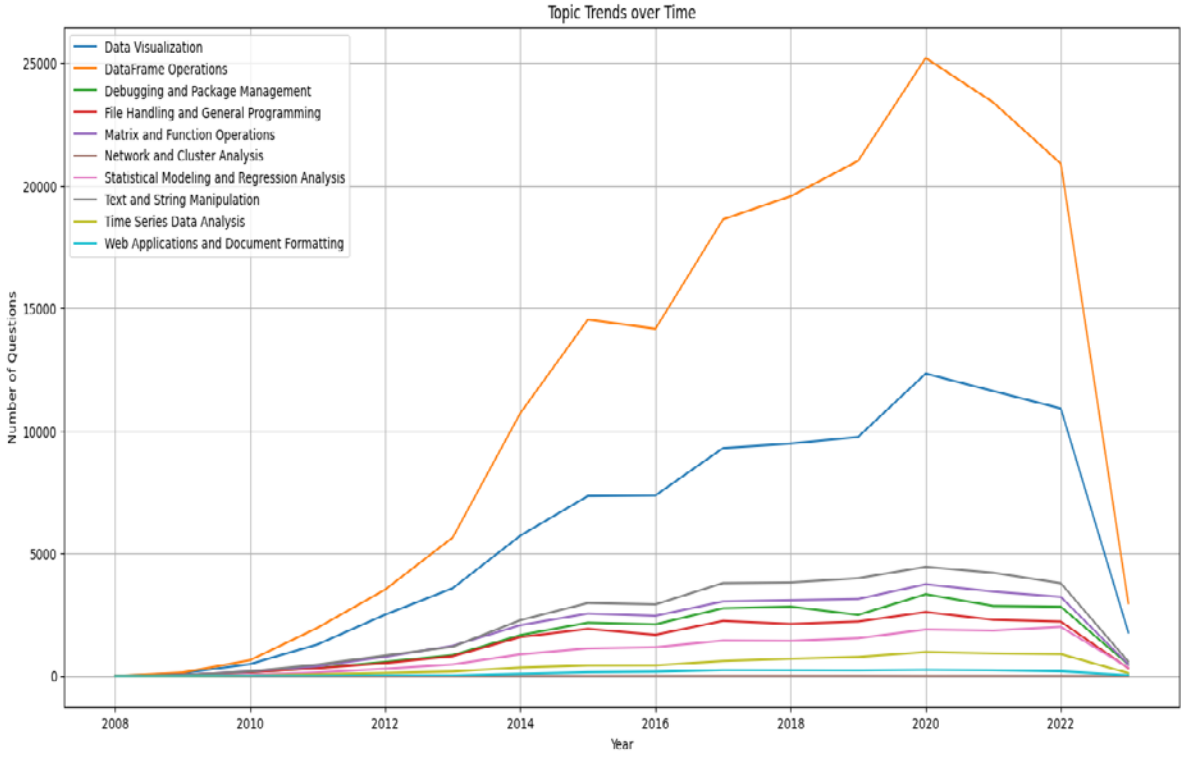
Στην δεύτερη θέση του πιο συχνά εμφανιζόμενου θέματος βρίσκεται, φυσικά, το **Data Visualization**, κάτι που είναι εξίσου αρκετά φυσιολογικό αφού η R έχει πολύ χρήσιμες και ευρέως χρησιμοποιούμενες βιβλιοθήκες για αυτόν τον σκοπό. Οι ερωτήσεις που ανήκουν σε αυτό το θέμα είναι **93.619** καθώς οι χρήστες προσπαθούν να δημιουργήσουν όμορφα και ευανάγνωστα διαγράμματα, προφανώς αντιμετωπίζουν κάποιες δυσκολίες. Μια χαρακτηριστική ερώτηση του συγκεκριμένου θέματος αναγράφεται στον Πίνακα 5 καθώς και το στιγμιότυπό της στην ιστοσελίδα του Stackoverflow στην Εικόνα 17.

Πίνακας 5. Παράδειγμα Ερώτησης Data Visualization.

Τίτλος	*How to add vertical lines and text to time series plot?*
Σώμα	*I have some time series data and used autoplot function in R to plot my time series. I would like add vertical lines to the plot and text. For example, a line between 2003-2010 with text "train data", 2010-2015 "test" data and 2015- 2018 "predictions". I did it in ordinary R plot, but it is not fancy. I would like to do it with ggplot or autoplot. My code is as follows.*
Link	https://stackoverflow.com/questions/75511910/select-rows-based-on-column-value-in-a-list-of-dataframes

Εικόνα 17.Ερώτηση Data Visualization του SO.

Στο τελικό στάδιο της ανάλυσης χρησιμοποιήθηκαν οι ημερομηνίες δημιουργίας της εκάστοτε ερώτησης, σε συνδυασμό με τα θέματα, ώστε να ανακαλυφθούν ποια θέματα ήταν δημοφιλή στην εξέλιξη της R ανά περιόδους. Παρατηρούμε στο διάγραμμα της Εικόνας 18 ότι τα πιο δημοφιλή θέματα που προαναφέρθηκαν παραπάνω, DataFrame Operation και Data Visualization, έχουν από την αρχή ραγδαία αύξηση και κατέχουν τον κυρίαρχο ρόλο μέχρι και σήμερα. Τα δύο αυτά θέματα αρχίζουν να γίνονται δημοφιλή το 2014, την περίοδο που και η R άρχισε να γίνεται γνωστή στο ευρύ κοινό, με την ακμή τους να βρίσκεται το 2020.



Εικόνα 18. Διάγραμμα με τις Ερωτήσεις των θεμάτων ανά περίοδο.

Κεφάλαιο 5

Συμπεράσματα

Στόχος της παρούσας εργασίας ήταν η μελέτη διάφορων χαρακτηριστικών της γλώσσας της R και των χρηστών της, καθώς και την εξερεύνηση υποκείμενων θεμάτων και τάσεων που αμφιταλαντεύουν την κοινότητα της R στο StackOverflow. Για τον πρώτο στόχο εφαρμόστηκαν τεχνικές Exploratory Analysis και Descriptive Statistics και αντίστοιχα για τον δεύτερο στόχο εφαρμόστηκαν διάφοροι αλγόριθμοι NLP, όπως είναι ο NER και LDA.

Αρχικά, όπως προαναφέραμε, η γλώσσα R υπέστη μια ραγδαία αύξηση στην δημοτικότητά της τα τελευταία χρόνια. Αυτή η αύξηση γίνεται συγκεκριμένα πιο αισθητή την χρονιά του 2020 όπου σημειώνει το «ρεκόρ» των ερωτήσεων που αναρτήθηκαν στην σελίδα. Αυτό αποδίδεται σε διάφορους λόγους, όπως είναι για παράδειγμα η «ανοιχτού-κώδικα» σχεδίαση της R, αφού εκτός από το ήδη πλούσιο οικοσύστημά που παρέχει σε βιβλιοθήκες και εργαλεία, μπορεί και η ίδια η κοινότητά να συνεισφέρει στην επέκταση και εξέλιξή της, κάτι που κάνει τους προγραμματιστές να την προτιμήσουν. Αυτό το διαπιστώνουμε και από τα αποτελέσματα της συγκεκριμένης έρευνας [5], αφού οι συγγραφείς της έδειξαν πως ένα πολύ μεγάλο ποσοστό των προγραμματιστών, την διάλεξαν για τα projects τους για την ανάλυση διάφορων δεδομένων του Covid-19, κάτι που δικαιολογεί και την αύξησή της δημοτικότητας την συγκεκριμένη χρονιά.

Ένα από τα βασικά χαρακτηριστικά της γλώσσας είναι η ενεργή και ποιοτική της κοινότητα. Το μεγάλο ποσοστό του 70% των απαντημένων ερωτήσεων μαζί με το ποσοστό του 56% των ερωτήσεων που έχουν επιβεβαιωμένη απάντηση είναι η απόδειξη του υποστηρικτικού και εξυπηρετικού κοινού της. Αυτός ο ισχυρισμός υποδηλώνει την συνεργατική φύση της κοινότητας και την αφοσίωσή της στην ανταλλαγή γνώσεων, κάτι που έχει μεγάλη σημασία αφενός για τους αρχάριους που δυσκολεύονται στα αρχικά εμπόδια της εκμάθησης της γλώσσας και αφετέρου για τους πιο έμπειρους χρήστες της που αναζητούν λύση σε πιο προχωρημένα θέματα.

Σύμφωνα με τα πιο δημοφιλή tags, τις πιο εμφανιζόμενες οντότητες και τα θέματα που εξήχθησαν μπορούμε εύκολα να αντιληφθούμε πως η συγκεκριμένη γλώσσα χρησιμοποιείται κυρίως στην επιστήμη των δεδομένων. Πιο συγκεκριμένα, τα

κυρίαρχα θέματα που αναδύονται από τις ερωτήσεις του Stackoverflow, όπως προαναφέρθηκε στην ανάλυσή μας, είναι τα DataFrame Operations και Data Visualization. Αυτό γίνεται αντιληπτό ακόμη και αν παρατηρήσουμε διάφορες λέξεις που εμφανίζονται στα πιο συχνά εμφανιζόμενα tags και οντότητες. Για παράδειγμα, λέξεις όπως «dplyr», «dataframe», «datatable» είναι λέξεις που αφορούν το D.O., καθώς αντίστοιχα και λέξεις όπως το «ggplot2», «plot», «tidyverse» είναι λέξεις που αφορούν το D.V.

Το γεγονός ότι έχει στραφεί η περισσότερη προσοχή στα παραπάνω δύο πεδία, δεν υποδεικνύει μόνο πως οι βασικές χρήσης της R περικλείονται σε αυτά αλλά επίσης και πως οι δυσκολίες που αντιμετωπίζουν οι χρήστες της βρίσκονται εκεί. Αυτό, όμως, μπορεί να οφείλεται στις πλούσιες παροχές που παρέχει η R σε αυτούς τους τομείς, καθώς είναι μια γλώσσα που εξειδικεύεται σε αυτούς, απαιτώντας μια απότομη καμπύλη εκμάθησης για τους αρχάριους αλλά και μεσαίους χρήστες της. Προβάλλοντας αυτά τα θέματα μπορεί να βοηθήσει τους εκπαιδευτές της γλώσσας να επικεντρώσουν το ενδιαφέρον τους σε αυτά καθώς επίσης και τους συντηρητές και προγραμματιστές των βιβλιοθηκών της γλώσσας να μπορέσουν να απλοποιήσουν και να καλυτερέψουν την εμπειρία του χρήστη.

Βιβλιογραφία

[1] Georgiou, Konstantinos et al. “A Study of Knowledge Sharing related to Covid-19 Pandemic in Stack Overflow.” *ArXiv* abs/2004.09495 (2020): n. pag.

[2] Tahmooresi, Hamed et al. “An Analysis of Python's Topics, Trends, and Technologies Through Mining Stack Overflow Discussions.” *ArXiv* abs/2004.06280 (2020): n. Pag

[3] M. Yazdaninia, D. Lo and A. Sami, "Characterization and Prediction of Questions without Accepted Answers on Stack Overflow," *2021 IEEE/ACM 29th International Conference on Program Comprehension (ICPC)*, Madrid, Spain, 2021, pp. 59-70, doi : 10.1109/ICPC52881.2021.00015.

[4] Chakraborty, P., Shahriyar, R., Iqbal, A., & Uddin, G. (2021). How do developers discuss and support new programming languages in technical Q&A site? An empirical study of Go, Swift, and Rust in Stack Overflow. *Information and Software Technology*, *137*, 106603.

[5] P. A. M. Oliveira, P. A. Santos Neto, G. Silva, I. Ibiapina, W. L. Lira and R. M. C. Andrade, "Software Development During COVID-19 Pandemic: an Analysis of Stack Overflow and GitHub," *2021 IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH)*, Madrid, Spain, 2021, pp. 5-12, doi: 10.1109/SEH52539.2021.00009.

[6] Johri, Vishal, and Srividya Bansal. "Identifying trends in technologies and programming languages using Topic Modeling." *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. IEEE, 2018.

[7] Nischal Shrestha, Colton Botta, Titus Barik, and Chris Parnin. 2022. Here we go again: why is it difficult for developers to learn another programming language? Commun. ACM 65, 3 (March 2022), 91–99. https://doi.org/10.1145/3511062.

[8] M. Papoutsoglou, N. Mittas and L. Angelis, "Mining People Analytics from StackOverflow Job Advertisements," 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Vienna, Austria, 2017, pp. 108-115, doi: 10.1109/SEAA.2017.50.

[9] Muhammad Asaduzzaman, Ahmed Shah Mashiyat, Chanchal K. Roy, and Kevin A. Schneider. 2013. Answering questions about unanswered questions of stack overflow. In Proceedings of the 10th Working Conference on Mining Software Repositories (MSR '13). IEEE Press, 97–100.

[10] S. Wang, T. Chen and A. Hassan, "How Do Users Revise Answers on Technical Q&A Websites? A Case Study on Stack Overflow" in IEEE Transactions on Software Engineering, vol. 46, no. 09, pp. 1024-1038, 2020. doi: 10.1109/TSE.2018.2874470.

[11] H. Zhang, S. Wang, T. -H. Chen and A. E. Hassan, "Reading Answers on Stack Overflow: Not Enough!," in IEEE Transactions on Software Engineering, vol. 47, no. 11, pp. 2520-2533, 1 Nov. 2021, doi: 10.1109/TSE.2019.2954319.

[12] L. Lord, J. Sell, F. Bagirov and M. Newman, "Survival Analysis within Stack Overflow: Python and R," *2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data)*, Barcelona, Spain, 2018, pp. 51-59, doi: 10.1109/Innovate-Data.2018.00015.

[13] J. Zou, L. Xu, W. Guo, M. Yan, D. Yang and X. Zhang, "Which Non-functional Requirements Do Developers Focus On? An Empirical Study on Stack Overflow Using Topic Analysis," *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, Florence, Italy, 2015, pp. 446-449, doi: 10.1109/MSR.2015.60.

[14] Xiong, Yunxiang & Meng, Zhangyuan & Shen, Beijun & Yin, Wei. (2017). Mining Developer Behavior Across GitHub and StackOverflow. 578-583. 10.18293/SEKE2017-062.

[15] Stack Overflow. (n.d.). Retrieved from https://insights.stackoverflow.com/survey

[16] https://r-community.org/stackoverflow/

[17] https://en.wikipedia.org/wiki/Python_(programming_language)

[18] https://www.kaggle.com/code/paultimothymooney/kaggle-survey-2022-all-results

[19] LIDDY, Elizabeth D. Natural language processing. 2001.

[20] BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent Dirichlet allocation. *Journal of machine Learning research*, 2003, 3.Jan: 993-1022.

[21] https://api.stackexchange.com/docs

[22] James Curran and Stephen Clark. 2003. Language Independent NER using a Maximum Entropy Tagger. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 164–167.

[23] S. JUGRAN, A. KUMAR, B. S. TYAGI and V. ANAND, "Extractive Automatic Text Summarization using SpaCy in Python & NLP," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 582-585, doi: 10.1109/ICACITE51222.2021.9404712.

[24] CHURCH, K. (2017). Word2Vec. Natural Language Engineering, 23(1), 155-162. doi:10.1017/S1351324916000334.

[25] https://en.wikipedia.org/wiki/Stack_Overflow

[26] https://stackoverflow.blog/2017/10/10/impressive-growth-r/